International Journal of Peptide Research and Therapeutics (2021) 27:2257–2273
https://doi.org/10.1007/s10989-021-10251-7

New Computational Approach for Peptide Vaccine Design Against SARS-COV-2

Subhamoy Biswas[1] · **Smarajit Manna**[2,3] · **Ashesh Nandy**[3] · **Subhash C. Basak**[4]

Accepted: 3 July 2021 / Published online: 10 July 2021

Abstract
The design for vaccines using in silico analysis of genomic data of different viruses has taken many different paths, but lack of any precise computational approach has constrained them to alignment methods and some alignment-free techniques. In this work, a precise computational approach has been established wherein two new mathematical parameters have been suggested to identify the highly conserved and surface-exposed regions which are spread over a large region of the surface protein of the virus so that one can determine possible peptide vaccine candidates from those regions. The first parameter, *w*, is the sum of the normalized values of the measure of surface accessibility and the normalized measure of conservativeness, and the second parameter is the area of a triangle formed by a mathematical model named 2D Polygon Representation. This method has been, therefore, used to determine possible vaccine targets against SARS-CoV-2 by considering its surface-situated spike glycoprotein. The results of this model have been verified by a parallel analysis using the older approach of manually estimating the graphs describing the variation of conservativeness and surface-exposure across the protein sequence. Furthermore, the working of the method has been tested by applying it to find out peptide vaccine candidates for Zika and Hendra viruses respectively. A satisfactory consistency of the model results with pre-established results for both the test cases shows that this in silico alignment-free analysis proposed by the model is suitable not only to determine vaccine targets against SARS-CoV-2 but also ready to extend against other viruses.

Keywords Peptide vaccines · In silico drug design · Alignment-free sequence analysis · Viral epidemics · SARS-CoV-2

Introduction

Since time immemorial, epidemics and pandemics have ravaged the human race. Many of these have been identified with viruses and sometimes antidotes have been discovered by keen observation as in the case of smallpox (Riedel 2005). The recent developments in immunogenetics, information technology, immunoinformatics, artificial intelligence and other sciences, especially the success of the Human Genome Project, have emboldened scientists to seek a closer understanding of the functioning of viruses and designing of anti-viral therapeutics. Traditional vaccines that bolster a human body's immune system are generally designed to be based on live attenuated, inactivated or subunit vaccines. These have the problem of long and costly development process and sometimes the vaccines revert to the original viral form thus complicating the process (Chit et al. 2014; Li et al. 2014; Lo et al. 2013).

The advent of bioinformatics and enhanced knowledge of the immune response to an invading pathogen as in Tomar and De (2014) and Backert and Kohlbacher (2015) led to a different pathway to the development of anti-viral vaccines. B cells and T cells are responsible for facilitating adaptive immunity response in the human body by developing antibodies. These cells only recognize the similar antigenic part of the pathogen. Surface-situated, antigenic parts or regions of the pathogen are called epitopes. So, it would be prudent to determine such regions and design vaccines that can elicit anti-viral response. Such ideas led to the concept of

✉ Subhamoy Biswas
subhab365@gmail.com

[1] Department of Electrical Engineering, Jadavpur University, Kolkata 700032, India

[2] Jagadis Bose National Science Talent Search, Kolkata 700107, India

[3] Centre for Interdisciplinary Research and Education, Kolkata 700068, India

[4] Department of Chemistry and Biochemistry, University of Minnesota Duluth, Duluth, Minnesota 55812, USA

"reverse vaccinology" and "vaccinomics" as in Rappouli (2001) and Poland et al. (2011, 2016), and the development of peptide vaccine models (Purcell et al. 2007; Nandy et al. 2018; Dudek et al. 2010). These advancements led to design of remote and individual specific vaccine that hold the promise of development of community-specific vaccines by discarding the "one size fits all" paradigm prevalent in current vaccine industry. Among several approaches, peptide vaccines have taken a leading role. The Zika virus pandemic of 2015–2016 gave added impetus to different studies for peptide vaccines. For example, Shawan et al. (2014) and Badawi et al. (2016) used the immunoinformatics approach to design peptide vaccines. Mirza et al. (2016) predicted antigenic epitopes of Zika viral proteins using immunoinformatics coupled with molecular dynamic simulations. Dar et al. (2016) used an in silico approach to predict promiscuous T-cell epitopes in the Zika polyprotein. Islam et al. (2012) identified conserved high-scoring epitopes in the Chikungunya virus using alignment techniques. Chakraborty et al. (2010) used sequence alignment methods to determine conserved segments for peptide vaccine design for all four types of the dengue virus. In the context of the currently raging pandemic caused by SARS-CoV-2, there is a huge demand placed on the production of vaccines on a very urgent basis. In this regard, many researchers have adopted approaches towards peptide vaccine design as in Abdelmageed et al. (2020), Slathia and Sharma (2020), Yazdani et al. (2020), Dagur et al. (2020), Kalita et al. (2020) and Durojaye et al. (2020).

Advancement of computer-assisted technologies and availability of viral sequence databases have paved the way for the design of peptide vaccines. In this context, different mathematical models and computational algorithms are used to study the appropriate proteins for their solvent exposure, their extent of mutation and other attributes. In our previous studies as in Dey et al. (2017, 2018, 2016) and Ghosh et al. (2012), graphical plots and manual observations have been relied upon to identify the regions of the virus, which have high surface exposure and low mutation probability i.e., more conserved. But there is need for a more robust and automated mathematical or computational approach that bypasses manual interventions so that the most highly conserved and surface accessible peptides can be figured out precisely through mathematical descriptors. This would be valuable for a quick response against newly emerging epidemics and pandemics. For this reason, in this article, a new computational approach has been introduced. The method begins with the definition of a new mathematical parameter—w parameter to rank all possible 12-length peptides from the full-length sequence of the target protein based on their surface exposure (ASA) and conservativeness (PV). The following step involves grouping the higher ranked 12-length peptides into different consolidated peptide zones based on their location and defining a new mathematical model—the 2D Polygon Representation, which can assign a score to those consolidated zones by representing each of them with the help of a simple triangle constructed by three independent variables. Two of these variables are already known beforehand, that is, the surface accessibility and conservativeness, but on the other hand a third biophysical parameter has also been studied in this regard, that is, the length or span of the zones. Using this approach, the best peptide candidates for vaccine development have been determined with respect to the current pandemic caused by SARS-CoV-2, in a much more quick, precise and automatic way. Additionally, these regions have been verified not only using the older approach of manual estimation but also with the comparative analyses made by several other researchers working in this field, to see if there is any consistency in the results given by the analysis. Finally, the method has been applied to design peptide vaccine candidates for two other viruses—Hendra virus and Zika virus. The best possible peptide regions determined for either of the two cases were found to match with an established set of peptide vaccine candidates, showing that the method is consistent when applied for other viruses as well. Therefore, this effort provides a quick starting point for wet-lab experiments to begin with, cutting the lead-time and costs by a significant amount, for vaccine design against any virus.

Some comparative studies on peptide vaccine design against SARS-CoV-2 include (Durojaye et al. 2020), which has used the main proteinase of the SARS-CoV-2 virus as the target protein and (Abdelmageed et al. 2020 and Dagur et al. 2020) which have used the envelope E protein for the same. However, in the analysis presented here for SARS-CoV-2, it is the surface-situated spike glycoprotein of SARS-CoV-2 that has been exclusively studied. Similar studies on peptide vaccine design using this spike protein have been demonstrated in Slathia and Sharma (2020), Yazdani et al. (2020) and Kalita et al. (2020).

Materials and Methods

Prerequisites

The protein sequences of the surface-situated spike glycoprotein of SARS-CoV-2 have been downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/). They are now characterized numerically using the hypothetical 20D Coordinate representation of protein sequences (Nandy et al. 2009). Using the Graphical Sliding Window Method (GSWM) (Ghosh et al. 2012; Biswas et al. 2019), a measure of conservativeness of all possible 12-length peptides, i.e., a Protein Variability (PV) index, was evaluated. Here the length "12" has been considered because peptide candidates

selected for vaccine design against any virus are short in length and their number of constituent amino acids generally ranges from 10 to 15. So, the peptide lengths have been uniformly chosen as 12 as a rough average of the two numbers (10 and 15). The Average Solvent Accessibility (ASA) of each of these peptides was also obtained (on a scale of 0–100) using the SABLE server (Porollo et al. 2003; Adamczak et al. 2004; Wagner et al. 2005).

Ranking the 12-Length Peptides Based on Surface Exposure and Conservativeness

The purpose of this approach described here is to filter out those peptide stretches from the target protein sequence, which have high surface-exposure and are conserved that is, having low chances of mutation. As explained before, the value of ASA gives the level of surface-exposure of the selected peptide. On the other hand, the magnitude of PV index denotes the extent to which the region is conserved in the face of mutation. As concluded in Nandy et al. (2009), the lower the value of PV, the higher is the level of conservation, and so, the lower is the chance of mutation. Thus, in order to serve the desired purpose, the peptide stretches having high ASA and low PV such that they are highly surface exposed and less vulnerable to mutation respectively, need to be determined first. For the purpose of simplification, the value $1/PV$ has been considered instead of PV, such that both ASA and $1/PV$ can be treated in the same way, that is, to have them as high as possible, so that, the peptide becomes suitable for vaccine design.

Now, for each 12-length peptide, the ASA and $1/PV$ values are not in the same range, although they are equally important in this aspect. The range of ASA value is from 0 to 100, whereas for $1/PV$, it is from 0 to 1. Hence it is imperative to normalize both the parameters before their use for the analysis. In this regard, we define two parameters:

$$(ASA)_n = \frac{ASA - \min(ASA)}{\max(ASA) - \min(ASA)} * 10$$

$$(1/PV)_n = \frac{(1/PV) - \min(1/PV)}{\max(1/PV) - \min(1/PV)} * 10$$

Here, $(ASA)_n$ = normalized version of ASA in a scale of 0–10. $(1/PV)_n$ = normalized version of $1/PV$ in a scale of 0–10.

Having obtained both the values of $(ASA)_n$ and $(1/PV)_n$, for each 12-length peptide, a new parameter w has been defined as:

$$w = (ASA)_n + (1/PV)_n$$

It is obvious that, since it is required to have high ASA and high $1/PV$, so in this regard, w must also be high. So, all the 12-length peptides considered from the primary target protein sequence, have been first ranked based on this w parameter. Consequently, the top ranks will be occupied by those peptides which have higher w value than the rest.

2D Polygon Representation

The rank list, therefore, sorts the 12-length peptides in descending order of their w value. Now, from this rank list, the top ranks have been filtered and grouped into zones based on their location in the sequence. For this analysis, we have restricted ourselves to the top 100 ranks for sequence length greater than 1000, top 75 ranks for sequence length in between 500 and 1000 and top 50 ranks for length below 500.

The grouping of the 12-length peptides is an important step in our approach. To explain this, we represent a 12-length peptide as $(f, f + 11)$ where, f gives the starting amino acid position of the 12-length peptide. Obviously, $f + 11$ becomes the ending position. We consider a similar 12-length peptide $(f_0, f_0 + 11)$. It is to be noted that the aforementioned representation of the 12-length peptides does not indicate any coordinate point but simply an understanding of a 12-length peptide by its starting and ending positions. Since the representations for both the 12-length peptides depend on the starting position only, so here we can represent the position of the two 12-length peptides to be on a single straight line. As a result, we can define a quantity ε to be the difference in their positions. It is easy to see that $\varepsilon = |f - f_0|$. ε will definitely be of the form of a natural number (ε cannot be 0 as two 12-length peptides cannot occur at the same position). Now if, ε is a low value, then that means the 12-length peptides are lying very close to one another. Therefore as ε increases, the two 12-length peptides become more and more distant from each other. Now for these two peptides, the process of grouping refers to the process of designing a representation that can fully contain the representations of both the 12-length peptides. This representation denotes nothing but a new peptide of length r, given as $\big(\min(f, f_0), \max(f + 11, f_0 + 11)\big) = (\min(f, f_0), \max(f, f_0))$ and

$$r = \max(f, f_0) - \min(f, f_0) + 1$$

So, to group n number of 12-length peptides represented as $(f_1, f_1 + 11), (f_2, f_2 + 11), \ldots\ldots\ldots, (f_n, f_n + 11)$, the grouped interval will be given as

$$(\min(f_1, f_1, \ldots\ldots, f_1), \max(f_1, f_1, \ldots\ldots, f_1))$$

Now the criteria for grouping lies in how close these n number of 12-length peptides are, that is, how small value of ε is mutually between each of these n 12-length peptides. If the regions are far off from one another and then

on top of that, if they are grouped together, then not only it overlooks the variation in surface accessibility and conservation in between them, but it also generates a smaller number of options to choose from when we finally design our shortlist of vaccine candidates. So, the above method of grouping is only done when the 12-length peptides are close to one another. As a boundary condition, we have considered any two 12-length peptides to fall in the same group if $\varepsilon = |f - f_0| < 12$, that is, the distance between their starting positions can be atmost 11, that is, the starting position of one cannot be greater than the ending position of the other.

After the grouping has been done, we define the *ASA* value of each of these grouped zones as the average of the *ASA* values of all the 12-length peptides which constitute that zone. It is to be noted that the *ASA* values of the 12-length peptides that we have considered over here are the original *ASA* values that were obtained from the SABLE server. Similarly, we define the *PV* value for each grouped peptide zone as the average of the *PV* values of its constituent 12-length peptides. The modified peptide stretches thus enlisted, must also have high *ASA* and low *PV*. To mathematically describe these definitions, we consider a case where we end up with 'k' number of peptide zones after grouping. The *ASA* values of the zones can be represented of the form, $(ASA)_{g,i}$, where $i = 1, 2, \ldots \ldots, k$. The suffix '$g$' denotes that here the representation is being made for the grouped zones. In a similar way, the PV values of the 'k' peptide zones can be shown as $(PV)_{g,1}, (PV)_{g,2}, \ldots, (PV)_{g,k}$. However, just like before, in order to have both the parameters to be treated in the same way, the value $1/(PV)_{g,i}$ has been considered in place of $(PV)_{g,i}$ for the ith peptide zone. Just like before, the parameters $(ASA)_{g,i}$ and $1/(PV)_{g,i}$ currently are not normalized and they belong to different range of values. So, on a scale of 0–10, both of their normalized versions can be defined as:

$$(ASA)_{g,n,i} = \frac{(ASA)_{g,i} - \min((ASA)_{g,i})}{\max\left((ASA)_{g,i}\right) - \min((ASA)_{g,i})} * 10$$

$$(1/(PV)_{g,i})_n = \frac{1/(PV)_{g,i} - \min(1/(PV)_{g,i})}{\max\left(1/(PV)_{g,i}\right) - \min(1/(PV)_{g,i})} * 10$$

Tentatively, therefore we have two mathematical values to describe the 'k' number of grouped peptide zones—$(ASA)_{g,n,i}$ and $(1/(PV)_{g,i})_n$ where i represents the ith peptide zone.

But there is one more characteristic which needs to be studied in particular for this case—the length of the grouped peptide zone, or in other words, the span across which it is spread out. Thus, if a peptide zone has a greater length, that means, it is much more spread out on the target protein. Now, this characteristic is of great importance for peptide vaccine design, primarily for two reasons. Firstly, such a peptide zone with a larger length means that it is highly conserved and surface-exposed over a wide region. So, if in case, a part of that zone gets affected by any mutation in future, then the remaining portion of that zone, if it remains conserved, can continue to be used for vaccine design. Secondly, if a portion of such a peptide zone turns to be perfectly matching with a human sequence, then designing a peptide vaccine based on the entire zone may cause complications related to auto-immunity. In that situation, if the zone has a large length, then even after removing the matching portion, we will be still left with a peptide region that has a length optimal enough to generate a sufficient immune response.

Hence, it is safe to comment that although initially, the rank list of all possible 12-length peptides was prepared based on only $(ASA)_n$ and $(1/PV)_n$ values because all of them were of the same length (that is, 12) but now for each of the 'k' number of grouped peptide zones, we are taking into account their length, defined as h_i, along with the other two values $(ASA)_{g,n,i}$ and $(1/(PV)_{g,i})_n$ where, $i = 1, 2 \ldots \ldots, k$. Here, again, the length value h_i has been normalized on the same scale of 0–10, just like the other two parameters. We define the normalized version of it as $(h_i)_n$.

In this regard, for each ith peptide zone, we define a triangle ABC as shown in Fig. 1. A point O is located inside the triangle such that $\angle AOB = \angle BOC = \angle COA = 120°$ and

$$length\ of\ OA = |OA| = (ASA)_{g,n,i},$$

$$length\ of\ OB = |OB| = (1/(PV)_{g,i})_n,$$

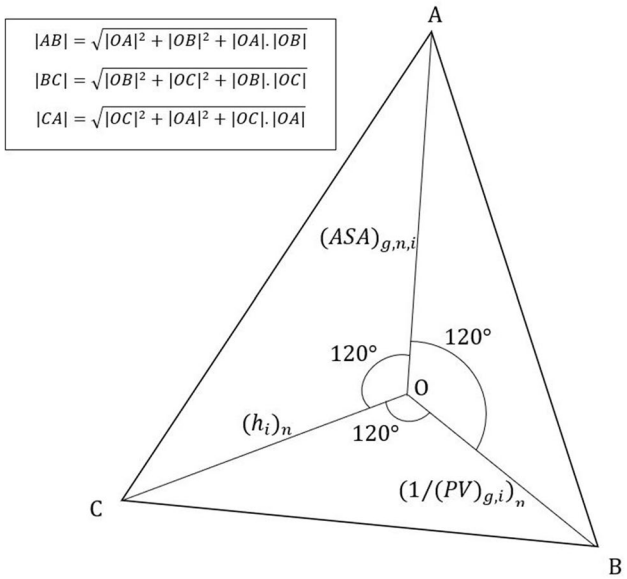

Fig. 1 Pictorial representation of the 2D Polygon model. Here, $OA = (ASA)_{g,n,i}$, $OB = (1/(PV)_{g,i})_n$, $OC = (h_i)_n$. Also, $\angle AOB = \angle BOC = \angle COA = 120°$

$$length\ of\ OC = |OC| = (h_i)_n = \frac{10 * (h_i - \min(h_i))}{\max(h_i) - \min(h_i)}$$

Therefore, it is clear that $|OA|$, $|OB|$ and $|OC|$ all must be high for the peptide to be suitable for selection. Now, if $|OA|$, $|OB|$ and $|OC|$ need to be high, then area of ΔABC for that peptide zone must also be high. Thus, if a particular peptide zone has to be suitable for selection compared to other zones, then its area of ΔABC must be larger in comparison with that of the other ones.

Therefore, the area of the triangle can be determined as:

$$area(\Delta ABC) = \sqrt{s(s-a)(s-b)(s-c)}$$

where,

$$a = \sqrt{|OB|^2 + |OC|^2 + |OB|.|OC|}$$

$$b = \sqrt{|OC|^2 + |OA|^2 + |OC|.|OA|}$$

$$c = \sqrt{|OA|^2 + |OB|^2 + |OA|.|OB|}$$

$$s = \frac{a+b+c}{2}$$

Using this mathematical approach, defined as the "2D Polygon Representation", the area of the triangle for each grouped peptide zone has been calculated. For a safe assumption, peptide zones in the top 50th percentile have been considered for the further analyses and the remaining ones have been discarded.

Further Analysis

The peptide zones in the top 50th percentile as per the 2D Polygon model score have been tested further for their epitope potential. For this purpose, the Immune Epitope Database Analysis Resource (IEDB-AR) (Vita et al. 2018) has been used. The list of possible epitopes from the full-length sequence of the spike protein were obtained from the database and matched with the peptide zones to screen those which had sufficient epitope potential.

The ones which showed good binding capacity have been considered for the final step of the analysis—to check their capacity to generate any autoimmune threats. This has been accomplished using the Basic Local Alignment Search Tool (BLAST) as in Altschul et al. (1990), by aligning them with the corresponding matches found for *Homo sapiens*. Significant matching means that the peptide region has the possibility of autoimmune threat and has to be discarded. In this way, a final shortlist for the conserved surface protein is obtained. For spike glycoprotein of SARS-CoV-2, a preliminary analysis was previously done using this computational model based on a smaller number of sequences (72) as in Biswas et al. (2020). But as many as 2812 amino acid sequences (available in the database as on 8 May, 2020) have been considered here and a detailed analysis pertaining to the selection of the most suitable peptide regions have been performed, resulting in a more comprehensive analysis.

Verification of the Entire Approach

The entire approach described so far, for determining the best peptide vaccine candidates, should not be limited to the case of only SARS-CoV-2. Rather, it should be such that its application can be extended to the purpose of any other virus as well. Hence, to verify this, the same approach as described before has been used to determine peptide vaccine candidates for Zika and Hendra viruses as well. For the Zika virus, the candidates were selected based on its surface-situated E protein. The final shortlist of candidates obtained have been matched with an already established set of results given for the Zika virus E protein in Dey et al. (2017). In the same way, for Hendra virus, the surface situated G glycoprotein was chosen and thereby, suitable peptide vaccine stretches obtained using our analysis were compared with similar established results for the G glycoprotein given in Dey et al. (2018). Consequently, a majority of the stretches matched with the published results for both the cases, showing that this method, overall, can be utilized for peptide vaccine design for any virus in general.

Results

Accordingly, 2812 full-length protein sequences of spike glycoprotein of SARS-CoV-2 were obtained from the NCBI database, which have been extensively used by this new method for determining the suitable peptide vaccine targets for SAR-CoV-2. The spike glycoprotein is a 1273 amino acid long sequence. Therefore, the top 100 ranks have been selected from the rank list given by the w parameter for spike glycoprotein, out of which the top 10 peptides of length 12 have been shown in Table 1. All of these 100 peptides of 12-length have been grouped based on their location in the sequence. In this way, 16 highly surface exposed and conserved peptide zones in the spike glycoprotein of SARS-CoV-2 have been figured out, which have been shown in Table 2. Figure 2 shows the superimposed ASA and PV profiles for the spike protein of SARS-CoV-2. For showing the constituent amino acids within the regions indicated by the study, here, the sequence BCF79924.1 (Japan, 2020) (https://www.ncbi.nlm.nih.gov/protein/1842103922) has been used as reference. In order to verify whether the results

Table 1 Top 10 peptides of length 12 for spike glycoprotein of SARS-CoV-2 given by *w* parameter

Rank	Starting position of the peptides	Score	PV	Peptide (Length = 12)
1	1027	19.27289	1	TKMSECVLGQSK
2	1026	18.86894	1	ATKMSECVLGQS
3	1030	18.82406	1	SECVLGQSKRVD
4	1028	18.62657	1	KMSECVLGQSKR
5	334	18.39318	1	NLCPFGEVFNAT
6	982	18.35727	1	SRLDKVEAEVQI
7	1031	18.17774	1	ECVLGQSKRVDF
8	1029	18.03411	1	MSECVLGQSKRV
9	985	18.00718	1	DKVEAEVQIDRL
10	1023	17.98025	1	NLAATKMSECVL

are satisfactory, eye-estimation method of the ASA-PV profiles of the spike protein based on the same 2812 sequences has been employed in the current analysis as well. This eye-estimation gave 14 peptide regions (listed in Table 2), out of which, 12 were found to be intersecting or lying in the close neighbourhood of the 16 regions obtained by grouping the top ranks.

Out of these 16 peptide regions obtained using the improvised method discussed here, 8 zones fell in the top 50th percentile based on the score generated by the 2D Polygon model. Therefore, as per the definition of the 2D Polygon representation, these zones were not only highly conserved and surface accessible but also large enough in length. Now these 8 zones have been tested for their epitope potential using the IEDB-AR server. The analysis was made for both the T-cell and B-cell immune responses. For the T-cell epitope analysis, we have used only the MHC II binding predictions from the IEDB-AR web server because here the focus is on antibody production against specific antigenic sites on the outer surface of the spike protein. The HLA DP/DQ and HLA DRB alleles chosen for this purpose cover a major part of the world population (Paul et al. 2015). Two separate analyses were made, one for HLA DP/DQ and the other for HLA DRB, and accordingly, for each of the two cases, two different lists of all possible epitopes of length 15, from the full-length sequence, were obtained from IEDB-AR. Now for a particular peptide zone enlisted as per the 2D Polygon model, each of the two lists of epitopes was matched with it. Thus, from each of the two lists, one epitope of length 15 was suggested for that zone such that the starting and ending positions of both the 15-length peptides more or less matched with the original zone, keeping the percentile rank as minimum as possible, with 10 as a safe threshold for the percentile rank. In this way, for every peptide zone determined by the 2D model, we determined two best matches, one as per the HLA DP/DQ analysis and another as per the HLA DRB analysis. However, those peptide zones for which neither of the two best possible 15-length epitopes from the two lists had percentile rank below 10, were discarded. We found that the zones 329–345 (FPNITNLCPFGEVFNAT), 414–440 (QTGKIADYNYKLPDDFTGCVIAWNSNN) and 386–400 (KLNDLCFTNVYADSF) had none of the 2 percentile values below 10, so they could not be considered for the next step of the analysis. On the contrary, the remaining 5 zones had at least one of the 2 percentile ranks to be below 10. So, for those 5zones, the two best-matching 15-length epitopes were compared with one another and the one having a lesser percentile rank, was compared once again with the parent peptide zone. The part which was found to be intersecting after comparison was figured out. This exercise has been done for all the 5 zones, in the end giving us 5 corresponding intersection regions which will be considered for the BLAST analysis.

Additionally, out of these 5 intersection zones, two of them namely—SNLKPFERDIST and PKKSTNLVKNKCVNF and were found to be present among the linear and discontinuous B-cell epitopes for the spike protein obtained from the predictions made by the IEDB-AR Ellipro server (Ponomarenko et al. 2008). The lists of all discontinuous and linear B-cell epitopes have been shown in Tables 3 and 4 respectively. In Table 5, the IEDB-AR results have been listed and summarized for all the 5 grouped peptide zones, wherein, the best possible percentile ranks along with the corresponding HLA alleles for each zone have also been mentioned. The starred marks for each case in the table indicates the best out of the two 15-length peptides chosen for determining the intersecting zone for the next and final step, that is, BLAST.

Finally, each of the 5 intersecting regions was aligned with all possible human sequences in the Non-redundant Protein Database of the BLAST server. Any significant match between an intersecting zone and a human sequence meant that the former can cause autoimmunity in the human host. The significance of the match was studied based on the E values obtained for each alignment for a particular intersecting zone. A low E value indicated that the strength of the match was high (Kerfeld et al. 2011). Therefore, the strongest match for every intersecting zone was the one that gave the least E value. For a safe assumption, we chose 1 as a threshold for the E value. If for any intersecting zone, the corresponding best match had an E value below 1, then those regions have been discarded, as they have a greater chance of causing an autoimmune disease in the human host. As such, the region LLTDEMIAQYTSALL was eliminated because it gave 0.37 as its best possible E value. Table 6 gives a summary of the 4 remaining intersecting zones and their corresponding best matches as per the BLAST analysis.

In the end, the remaining 4 intersecting zones that satisfied the BLAST analysis, have been considered as potential

Table 2 Peptide zones for spike glycoprotein of SARS-CoV-2 predicted using the 2D model and comparison with the eye-estimated regions for the same, using all the 2812 full-length sequences of the protein used in the current analysis

As per 2D Polygon Representation model, using the currently available 2812 sequences (as of 8 May, 2020)			As per eye-estimation of the ASA and PV profiles, using the same 2812 sequences	
Start–End position of the region based on 2D Polygon Representation	Peptide Stretch (based on the 2D model)	2D Polygon Score	Start–End position of the region based on eye-estimation	Peptide Stretch (based on eye-estimation)
1021–1046**	SANLAATKMSECVLGQSKRVDFCGKG	66.20194	1019–1049**	RASANLAATKMSECVLGQSKRVDFCGKGYHL
329–345**	FPNITNLCPFGEVFNAT	30.595	323–347**	TESIVRFPNITNLCPFGEVFNATRF
975–1004**	SVLNDILSRLDKVEAEVQIDRLITGRLQSL	64.50292	953–1007**	NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTY
458–470**	KSNLKPFERDIST	17.15506	456–475**	FRKSNLKPFERDISTEIYQA
414–440**	QTGKIADYNYKLPDDFTGCVIAWNSNN	65.6183	412–448**	PGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGN
523–546**	TVCGPKKSTNLVKNKCVNFNFNGL	39.63615	523–551**	TVCGPKKSTNLVKNKCVNFNFNGLTGTGV
386–400**	KLNDLCFTNVYADSF	22.54376	349–399**	SVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADS
373–386	SFSTFKCYGVSPTK	11.57532	349–399**	SVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADS
860–878**	VLPPLLTDEMIAQYTSALL	36.08575	858–877	LTVLPPLLTDEMIAQYTSAL
289–300	VDCALDPLSETK	14.40109	–	–
1088–1108	HFPREGVFVSNGTHWFVTQRN	17.0471	–	–
660–671	YECDIPIGAGIC	13.04917	655–671	HVNNSYECDIPIGAGIC
901–917	QMAYRFNGIGVTQNVLY	11.67947	907–923	NGIGVTQNVLYENQKLI
1148–1161	FKEELDKYFKNHTS	16.8394	1145–1161	LDSFKEELDKYFKNHTS
316–335	SNFRVQPTESIVRFPNITNL	13.16672	–	–
782–793	FAQVKQIYKTPP	7.009527	779–795	QEVFAQVKQIYKTPPIK
–	–	–	1168–1185	DISGINASVVNIQKEIDR
–	–	–	195–209	KNIDGYFKIYSKHTP

Double asterisks refer to those peptide regions which fell in the top 50th percentile as per the 2D Polygon score for SARS-CoV-2, Hendra and Zika viruses respectively

candidates for vaccine design against SARS-CoV-2. These 4 regions have been obtained after IEDB-AR and BLAST analyses of the peptide zones derived from the 2D Polygon model. Now, that we have obtained our shortlist for SARS-CoV-2 using the new method, we have also performed the same analyses on the 14 peptide zones obtained from eye-estimation to check whether the results are consistent or not. The 2D Polygon representation filtered out 7 peptide zones out of all the 14 ones (marked with star in Table 2), based on their span, conservativeness and surface exposure, just like before. Analysis using the MHC II DRB and MHC II DP/DQ alleles from the IEDB-AR showed that all 7 zones had really good binding capacity. The 7 zones were matched individually and the intersecting zone for each of them was figured out, in the same way as it has been done in the previous case. These 7 intersecting zones being tested using the BLAST server showed only one region—YSVLYNSASFSTFKC that could cause any autoimmune disease. Therefore, the remaining 6 intersecting zones formed the final shortlist based on the eye-estimation method. These were now compared with the previously obtained 4 intersecting regions, which showed three cases where a good match was obtained. The comparison between the two shortlists obtained using

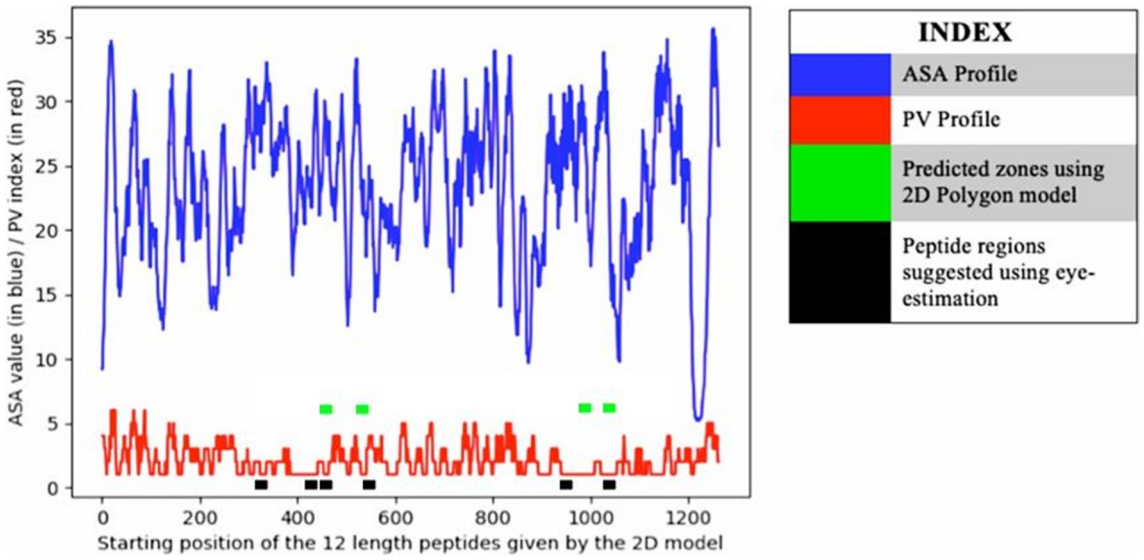

Fig. 2 ASA and PV profiles of the full-length sequence of spike glycoprotein of SARS-CoV-2: The components shown in the figure are: *ASA* profile (blue), *PV* profile (red), predicted peptide stretches using the new approach (green) and peptide stretches determined using eye-estimation (black). (Color figure online)

the older and the newer methods have been shown in Table 7 (the 3 matching regions have been highlighted).

The 4 shortlisted candidates, predicted using the current approach, have been marked by green horizontal lines whereas the zones suggested by eye-estimation have been marked by black lines in Fig. 2. The comparative study therefore shows that this new approach is well consistent when applied to determine peptide candidates for SARS-CoV-2. But it also needs to be seen whether it is able to maintain the consistency for other viruses as well.

Comparison of the results obtained here for spike glycoprotein of SARS-CoV-2 with the other equivalent analyses made by several other researchers on the same protein has also shown consistency to a good extent. In Slathia and Sharma (2020), several MHC II based epitopes for peptide vaccine design have been listed. On comparison, the shortlisted zone 987–1001 suggested in our analysis was found among the list. Similarly, a zone 505–534 was suggested in Yazdani et al. (2020) that intersects with the finally predicted peptide zone 527–541 mentioned in our analysis.

An interesting result was observed in Di Paola et al. (2020) where few residues on the surface-situated spike protein of SARS-CoV-2 were mentioned which, on a general basis, could function as specific sites called as allosteric modulation regions which in turn could allow allosteric drugs to function on the areas where the incoming SARS-CoV-2 virus interact with the ACE2 receptor of the human host. Therefore, it will be of great importance if the finally shortlisted regions predicted using our new method contain those residues. In Di Paola et al. (2020), the amino acid numbering convention that was used was to indicate the first amino acid position as the 0th index. Here, we have used the convention of denoting the first aa position as the 1st index. As per the convention in Di Paola et al. (2020), there were three classes of allosteric modulation regions (AMR); the first class of moderate dynamical effect on other allosteric regions included 320Q, 321P, 322 T, 323E, 539 N, 548 T and 578P; the second class of large intermolecular allosteric effect in the receptor binding domain (RBD) included 531 N and 532L and the third class of moderate to high dynamical effect on the AMRs, RBD and ACE2 include only 580 T. Therefore, as per our convention of naming the amino acid positions, all the aforementioned residues will be 321Q, 322P, 323T, 324E, 540N, 549T, 579P, 532N, 533L and 581T. Out of these, the residues 532 N, 533L and 540N are well included in our finally shortlisted peptide zone PKKSTNLVKNKCVNF (527–541). This shows the region 527–541 in particular, as predicted by our new method, contains certain sites using which the binding between the SARS-CoV-2 and the ACE2 receptor can be easily controlled, using allosteric drugs.

Verification of the Approach Using the Case of G Glycoprotein of Hendra Virus

The first test case that we chose was to predict suitable vaccine candidates using this method for the Hendra virus. The surface-situated G glycoprotein has been considered in this case, which is a 604 amino acids (aa) long sequence. The finally shortlisted peptide stretches obtained based on this G glycoprotein have been matched with the established results given for the same in Dey et al. (2018). There were

Table 3 Discontinuous epitopes of spike glycoprotein of SARS-CoV-2 based on the analysis using IEDB-AR Ellipro

Residues	Number of residues	Score
A:Y707, A:S708, A:N709, A:N710, A:S711, A:I712, A:A713, A:I714, A:P715, A:T716, A:N717, A:F718, A:A783, A:Q784, A:V785, A:K786, A:Q787, A:I788, A:Y789, A:K790, A:T791, A:P792, A:P793, A:I794, A:K795, A:D796, A:F797, A:G798, A:G799, A:F800, A:P863, A:L864, A:L865, A:E868, A:M869, A:Q872, A:Y873, A:S875, A:A876, A:A879, A:G880, A:I882, A:T883, A:S884, A:G885, A:W886, A:T887, A:F888, A:G889, A:A890, A:G891, A:A892, A:A893, A:L894, A:Q895, A:I896, A:P897, A:F898, A:A899, A:M900, A:Q901, A:M902, A:A903, A:Y904, A:F906, A:N907, A:G908, A:I909, A:G910, A:V911, A:T912, A:Q913, A:N914, A:V915, A:L916, A:Y917, A:E918, A:N919, A:Q920, A:K921, A:L922, A:I923, A:A924, A:N925, A:L1034, A:G1035, A:Q1036, A:Q1071, A:E1072, A:K1073, A:N1074, A:F1075, A:T1076, A:T1077, A:A1078, A:P1079, A:A1080, A:I1081, A:C1082, A:H1083, A:D1084, A:G1085, A:K1086, A:A1087, A:H1088, A:F1089, A:P1090, A:R1091, A:E1092, A:G1093, A:V1094, A:F1095, A:V1096, A:S1097, A:N1098, A:G1099, A:T1100, A:H1101, A:W1102, A:F1103, A:V1104, A:T1105, A:Q1106, A:R1107, A:N1108, A:F1109, A:Y1110, A:E1111, A:P1112, A:Q1113, A:I1114, A:I1115, A:T1116, A:T1117, A:D1118, A:N1119, A:T1120, A:F1121, A:V1122, A:S1123, A:G1124, A:N1125, A:C1126, A:D1127, A:V1128, A:V1129, A:I1130, A:G1131, A:I1132, A:V1133, A:N1134, A:N1135, A:T1136, A:V1137, A:Y1138, A:D1139, A:P1140, A:L1141, A:Q1142, A:P1143, A:E1144, A:L1145, A:D1146, A:S1147	164	0.751
A:R328, A:F329, A:P330, A:N331, A:I332, A:T333, A:N334, A:L335, A:C336, A:P337, A:F338, A:G339, A:E340, A:V341, A:F342, A:N343, A:A344, A:T345, A:R346, A:F347, A:A348, A:S349, A:V350, A:Y351, A:A352, A:W353, A:N354, A:R355, A:K356, A:R357, A:I358, A:S359, A:N360, A:C361, A:V362, A:A363, A:D364, A:V367, A:L368, A:S371, A:A372, A:S373, A:F374, A:S375, A:T376, A:Y380, A:T393, A:N394, A:V395, A:Y396, A:A397, A:D398, A:S399, A:F400, A:V401, A:I402, A:R403, A:G404, A:D405, A:E406, A:V407, A:R408, A:Q409, A:I410, A:A411, A:P412, A:G413, A:Q414, A:T415, A:G416, A:K417, A:I418, A:A419, A:D420, A:Y421, A:N422, A:Y423, A:K424, A:L425, A:P426, A:D427, A:D428, A:F429, A:V433, A:I434, A:A435, A:W436, A:N437, A:S438, A:N439, A:N440, A:L441, A:D442, A:N448, A:Y449, A:N450, A:Y451, A:L452, A:Y453, A:R454, A:L455, A:F456, A:R457, A:K458, A:S459, A:N460, A:L461, A:K462, A:P463, A:F464, A:E465, A:R466, A:D467, A:I468, A:S469, A:T470, A:F490, A:P491, A:L492, A:Q493, A:S494, A:Y495, A:G496, A:F497, A:Q498, A:P499, A:T500, A:N501, A:V503, A:G504, A:Y505, A:Q506, A:P507, A:Y508, A:R509, A:V510, A:V511, A:V512, A:L513, A:S514, A:E516, A:L517, A:L518, A:H519, A:A520, A:P521, A:A522, A:T523, A:V524, A:C525, A:G526, A:P527, A:K528, A:K529, A:S530, A:T531, A:N532, A:L533, A:V534, A:K535, A:N536, A:K537, A:N544, A:T553, A:E554, A:S555, A:N556, A:K557, A:F559, A:L560, A:P561, A:F562, A:Q563, A:V576, A:D578, A:P579, A:Q580, A:T581, A:L582, A:E583, A:I584, A:L585	182	0.735
A:A27, A:Y28, A:T29, A:N30, A:S31, A:F32, A:F59, A:S60, A:N61, A:V62, A:T63, A:W64, A:F65, A:H66, A:A67, A:I68, A:H69, A:P82, A:V83, A:L84, A:P85, A:N87, A:F92, A:A93, A:S94, A:T95, A:E96, A:K97, A:S98, A:N99, A:I100, A:I101, A:R102, A:G103, A:W104, A:I105, A:F106, A:G107, A:T108, A:T109, A:L110, A:D111, A:S112, A:K113, A:S116, A:L117, A:L118, A:I119, A:V120, A:N121, A:N122, A:A123, A:T124, A:N125, A:V126, A:V127, A:I128, A:K129, A:V130, A:C131, A:E132, A:F133, A:Q134, A:F135, A:C136, A:N137, A:D138, A:P139, A:F140, A:L141, A:G142, A:V143, A:C166, A:T167, A:F168, A:E169, A:Y170, A:V171, A:S172, A:F186, A:K187, A:N188, A:L189, A:R190, A:E191, A:F192, A:G199, A:I203, A:S205, A:K206, A:H207, A:T208, A:P209, A:I210, A:N211, A:L212, A:V213, A:R214, A:D215, A:L216, A:P217, A:Q218, A:G219, A:L223, A:L226, A:V227, A:L229, A:P230, A:I231, A:G232, A:I233, A:N234, A:I235, A:T236, A:R237, A:F238, A:Q239, A:T240, A:L241, A:L242, A:A263, A:A264, A:Y265, A:Y266, A:V267	125	0.732
A:E702, A:N703, A:S704, A:V705, A:A706	5	0.613
A:N801, A:F802, A:S803, A:Q804, A:I805, A:L806, A:P807, A:D808, A:P809, A:S810, A:K811, A:S813, A:K814, A:R815	14	0.554
A:D985, A:P986, A:P987	3	0.548
A:T747, A:E748, A:S750, A:N751, A:L754, A:Q755, A:G757, A:S758	8	0.546

15 full-length sequences of G glycoprotein that were analysed in Dey et al. (2018) for obtaining the suiable peptides regions. Therefore, in this analysis, we have used the exact same 15 sequences to determine similar peptide candidates before they are being compared. Figure 3 shows the *ASA-PV* profile for the G glycoprotein. Here the sequence AEB21198.1 (https://www.ncbi.nlm.nih.gov/nuccore/AEB21198.1) from the NCBI database has been utilized as reference to represent the sequences in the tables and figures.

In a similar way as before, in the first step, a ranking list of all possible 12-length peptides retrieved from the full-length G protein sequence has been obtained using the *w* parameter. Since the sequence is only 604 aa long, so we preferred to consider only the top 75 ranks from the ranking list. The top 75 ranks have been then grouped into peptide zones based on their location in the sequence. This gave 8 consolidated peptide zones which were now analysed using the 2D Polygon model Table 8 shows the 8 consolidated regions obtained after grouping of the top 75 ranks.

Based on the 2D Polygon model score obtained, the regions YGTMDIKKINDGLLDSKILGA (29–49),

Table 4 Linear epitopes of spike glycoprotein of SARS-CoV-2 based on the analysis using IEDB-AR Ellipro

Start–End position of the linear epitopes predicted by Ellipro	Peptide stretch	Score
1071–1147	QEKNFTTAPAICHDG……PELDS	0.882
92–192	FASTEKSNIIRGWIFG……NLREF	0.811
433–537	VIAWNSNNLD……KSTNLVKNK	0.767
328–364	RFPNITNLCPFGEVF……NCVAD	0.754
236–267	TRFQTLLALHRSYL……AAYYV	0.728
553–564	TESNKKFLPFQQ	0.728
393–428	TNVYADSFVIRGDE……KLPDD	0.718
60–86	SNVTWFHAIHVSGT……PVLPF	0.699
203–219	IYSKHTPINLVRDLPQG	0.686
702–718	ENSVAYSNNSIAIPTNF	0.678
576–585	VRDPQTLEIL	0.674
879–925	AGTITSGWTFGAGAA……KLIAN	0.629
783–815	AQVKQIYKTPPIKDFGG……SKR	0.622
371–376	SASFST	0.575
226–234	LVDLPIGIN	0.507

Table 5 Summary of the IEDB-AR analysis of the 5 grouped peptide zones obtained for spike protein of SARS-CoV-2, which had good binding capacity

Start–End	Grouped peptide zone	MHC II DP/DQ			MHC II DRB		
		Score	Adjusted peptide	Allele	Score	Adjusted peptide	Allele
1021–1046	SANLAATKMSECV-LGQSKRVDF-CGKG	5*	SANLAATKM-SECVLG	HLA-DQA1*01:02/ DQB1*06:02	14	MSECVL-GQSKRVDFC	HLA-DRB1*03:01
975–1004	SVLNDILSRLD-KVEAEVQIDR-LITGRLQSL	3.8	SRLDKVEAEV-QIDRL	HLA-DQA1*03:01/ DQB1*03:02	1.1*	VEAEVQIDRLIT-GRL	HLA-DRB1*03:01
458–470	KSNLKPFERDIST	14	SNLKPFERDISTEIY	HLA-DQA1*03:01/ DQB1*03:02	1.8*	SNLKPFERDISTEIY	HLA-DRB3*01:01
523–546	TVCGPKKSTN-LVKNKCVNFN-FNGL	16	NLVKNKCVNFN-FNGL	HLA-DPA1*02:01/ DPB1*05:01	5.1*	PKKSTN-LVKNKCVNF	HLA-DRB1*13:02
860–878	VLPPLLTDEMI-AQYTSALL	12	LLTDEMIAQYT-SALL	HLA-DPA1*02:01/ DPB1*01:01	1.6*	LLTDEMIAQYT-SALL	HLA-DRB1*15:01

The star mark indicates 15-length adjusted peptide that has been chosen to form the intersecting zone corresponding to the particular peptide zone

The single asterisk marks represent the 15-length epitope chosen out of the two choices for every grouped peptide zone, such that the chosen ones form intersecting zones with the grouped peptide regions for SARS-CoV-2, Hendra and Zika viruses respectively

Table 6 Best possible BLAST matches and their corresponding E values for the 4 intersecting peptide zones finally shortlisted as peptide vaccine candidates for SARS-CoV-2

Intersecting peptide zone	Best BLAST match	E value of the match	Accession/PDB ID
SANLAATKMSECVLG	Crystal structure of influenza A NS1A protein in complex with F2F3 fragment of human cellular factor CPSF30, Northeast Structural Genomics Targets OR8C and HR6309A	22	2RHK_C
VEAEVQIDRLITGRL	Chromosome 14 open reading frame 103	3.9	EAW81633.1
SNLKPFERDIST	Rho-associated, coiled-coil containing protein kinase 1 variant	9.1	AAI13115.1
PKKSTNLVKNKCVNF	Chromosome 17, hCG 2,045,508	15	EAW89537.1

Table 7 Comparison of the final shortlists obtained from the newer and older approaches showed that there were 3 instances where a good level of matching was observed, thus, indicating that our new approach is indeed consistent in terms of determining peptide candidates for SARS-CoV-2

Eligible vaccine candidates selected as per the new approach using the 2D Polygon model		Eligible vaccine candidates selected as per the older method of eye-estimation	
Start–End position	Peptide stretch	Start–End position	Peptide stretch
1021–1035	SANLAATKMSECVLG	1019–1033	RASANLAATKMSECV
–	–	959–973	LNTLVKQLSSNFGAI
527–541	PKKSTNLVKNKCVNF	537–551	KCVNFNFNGLTGTGV
459–470	SNLKPFERDIST	460–474	NLKPFERDISTEIYQ
–	–	431–445	GCVIAWNSNNLDSKV
–	–	323–337	TESIVRFPNITNLCP
987–1001	VEAEVQIDRLITGRL	–	–

The matching peptide zones have been placed side-by-side for easy comparison

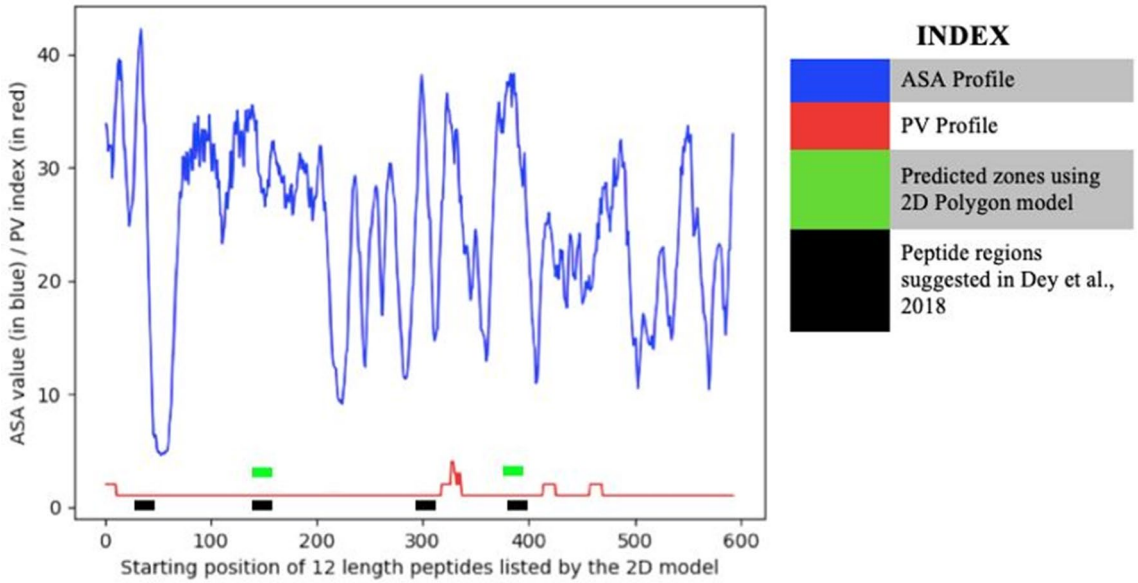

Fig. 3 ASA and PV profiles of the full-length sequence of G glycoprotein of Hendra virus: The components shown in the figure are: *ASA* profile (blue), *PV* profile (red), predicted peptide stretches using the new approach (green) and peptide stretches determined in Dey et al. (2018) (black). (Color figure online)

Table 8 The 8 grouped peptide zones obtained after grouping of the top 75 ranks for G glycoprotein

Starting position	Ending position	Consolidated Peptide stretch
29	49	YGTMDIKKINDGLLDSKILGA**
11	28	NNNLSGKIKDQGKVIKNY**
370	400	LPRTEFQYNDSNCPIIHCKYSKAENCRLSMG**
297	313	VSHVGDPILNSTSWTES
122	154	ANIGLLGSKISQSTSSINENVNDKCKFTLPPLK**
84	110	KESLQSVQQQIKALTDKIGTEIGPKVS
548	564	QVPLAEDDTNAQKTITD
593	604	FAVKIPAQCSES

The zones marked with star indicate the ones which fell into the top 50th percentile as per the 2D Polygon analysis

Double asterisks refer to those peptide regions which fell in the top 50th percentile as per the 2D Polygon score for SARS-CoV-2, Hendra and Zika viruses respectively

Table 9 Summary of the IEDB-AR study for G glycoprotein where the best possible 15-length epitopes for both the MHC II DRB and DP/DQ analyses have been listed

Start–End	Grouped peptide zone	MHC II DP/DQ			MHC II DRB		
		Score	Adjusted peptide	Allele	Score	Adjusted peptide	Allele
29–49	YGTMDIKKINDG-LLDSKILGA	13	KINDGLLDSKIL-GAF	HLA-DQA1*01:02/ DQB1*06:02	14	KINDGLLDSKIL-GAF	HLA-DRB1*03:01
11–28	NNNLSGKIKDQGK-VIKNY	37	LSGKIKDQGK-VIKNY	HLA-DPA1*02:01/ DPB1*05:01	12	LSGKIKDQGK-VIKNY	HLA-DRB3*01:01
370–400	LPRTEFQYND-SNCPIIHCKYS-KAENCRLSMG	28	PRTEFQYNDSNCPII	HLA-DQA1*01:01/ DQB1*05:01	6.1*	EFQYNDSNCPII-HCK	HLA-DRB3*02:02
122–154	ANIGLLGSKISQST-SSINENVNDKCK-FTLPPLK	4.8*	VNDKCKFTLP-PLKIH	HLA-DPA1*02:01/ DPB1*14:01	5.8	ANIGLLGSKISQSTS	HLA-DRB1*15:01

The starred regions were chosen to form the intersecting zone corresponding to the particular peptide zone

The single asterisk marks represent the 15-length epitope chosen out of the two choices for every grouped peptide zone, such that the chosen ones form intersecting zones with the grouped peptide regions for SARS-CoV-2, Hendra and Zika viruses respectively

NNNLSGKIKDQGKVIKNY (11–28), LPRTEFQYND-SNCPIIHCKYSKAENCRLSMG (370–400) and ANIGLLGSKISQSTSSINENVNDKCKFTLPPLK (122–154) fell into the top 50th percentile, meaning that they not only had high surface exposure and were conserved but also, they had a large length, quite favourable for vaccine design. These 4 regions have been marked in star in Table 8.

Using the 4 grouped peptide zones obtained, we now performed the IEDB-AR study with the help of the MHC II DRB and DP/DQ alleles based on the three affected countries India, Bangladesh and Malaysia, as per Dey et al. (2018). The study confirmed that there were two regions that did not show sufficient epitope potential—11–28 and 29–49. For both these cases, neither of the DRB or the DP/DQ analyses gave any 15-length epitope with percentile rank within 10. Hence both these regions have been discarded. The summary of the IEDB-AR study of the 4 grouped peptide zones has been shown in Table 9.

So now, we are left with only two grouped peptide regions with have been matched with the IEDB-AR table to obtain their corresponding intersection zones. These intersection zones have been tested using the BLAST server to check whether there is any case of autoimmunity or not. Table 10 shows the summary of the BLAST analysis for these two intersecting zones. Fortunately, none of the regions gave E value below 1, showing that they have negligible chances of causing autoimmune threats.

Therefore, the two peptide stretches EFQYNDSNCPII-HCK and VNDKCKFTLPPLKIH form the final shortlist of peptide vaccine candidates for Hendra virus. We have compared these regions with the same given in Dey et al. (2018) for each of the three affected countries—India, Bangladesh and Malaysia. Table 11 gives this comparison. From the comparative analysis, we can easily infer that the two regions we predicted were found among the published results showing that our model was indeed consistent in this test case.

Additionally, the two regions predicted by the current method have been depicted in green horizontal lines in Fig. 3 whereas, the shortlisted regions listed in Dey et al. (2018) have been marked in black horizontal lines, for better understanding.

Verification of the Approach Using the Case of E Protein of Zika Virus

The second test case that we chose was to predict suitable vaccine candidates using this method for the Zika virus. The surface-situated E protein has been considered in this case, which is naturally a 504 amino acids (aa) long sequence. Here for verification, we have taken the reference of Dey

Table 10 Best possible BLAST matches and their corresponding E values for the 2 intersecting peptide zones finally shortlisted as peptide vaccine candidates for G protein of Hendra virus

Intersecting zone	Best BLAST match	E value of the match	Accession/PDB ID
EFQYNDSNCPIIHCK	Immunoglobulin light chain junction region in Homo sapiens	9.1	MCC90139.1
VNDKCKFTLPPLK	Coxsackievirus and adenovirus receptor isoform 4 precursor in Homo sapiens	45	NP_001193994.1

Table 11 Comparison of the final shortlist for peptide candidates for G protein of Hendra virus obtained from the new approach with that given in Dey et al. (2018), after the IEDB-AR and BLAST analyses

Final shortlist for G protein of Hendra virus as per the new method	Final shortlists obtained for G protein of Hendra virus for each of the three affected countries as per Dey et al. (2018)		
	India	Bangladesh	Malaysia
EFQYNDSNCPIIHCK (374–388)	PIAECQYSKPENCRL (383–397)	FKYNDSNCPIAECQY (375–389)	PITKCQYSKPENCRL (383–397)
VNDKCKFTLPPLK (142–154)	VNEKCKFTLPPLKIH (142–156)	VNEKCKFTLPPLKIH (142–156)	NVNEKCKFTLPPLKI (141–155)
–	KKINEGLLDSKILSA (35–49)	KINEGLLDSKILSAF (36–50)	YGTMDIKKINEGLLD (29–42)
–	AVSVVGDPILNSTYW (296–310)	VVGDPVLNSTYWSNS (299–313)	PILNSTYWSGSLMMT (303–317)

The comparison shows that all the regions we predicted were present among the published results in Dey et al. (2018)

et al. (2017). In Dey et al. (2017), the analysis considered 25 full-length E protein sequences of 504 aa length and 35 fragments of E protein of 251 aa length. The portion of the full-length sequence represented by the 251 aa fragments was retrieved from each of the 25 full-length sequences. This gave 60 251 aa fragments in total, based on which the vaccine candidates were shortlisted in Dey et al. (2017). We considered the same 60 fragment proteins of the E protein and derived the relevant peptide candidates before comparison with the shortlist given in Dey et al. (2017). Figure 4 shows the *ASA-PV* profile for the E protein. Here the protein sequence AHL43501.1 (https://www.ncbi.nlm.nih.gov/protein/AHL43501.1) from the NCBI database has been used to represent the regions in the corresponding tables and figures for our Zika virus analysis.

At first, a ranking list of all possible 12-length peptides retrieved from the 251 aa fragment protein sequence has been obtained using the *w* parameter. Since the sequence is only 251 aa long, so we preferred to consider only the top 50 ranks from the ranking list. The top 50 ranks have been then grouped into peptide zones based on their location in the sequence. This gave 9 consolidated peptide zones which were now analysed using the 2D Polygon model Table 12 shows the 9 consolidated regions obtained after grouping of the top 50 ranks.

Using the 2D Polygon model score, there were 5 peptide zones that were screened from the 9 regions. This means, these 5 zones had high solvent accessibility, high conservation from mutation and a broad span, thus making them suitable for vaccine design. These 5 regions have

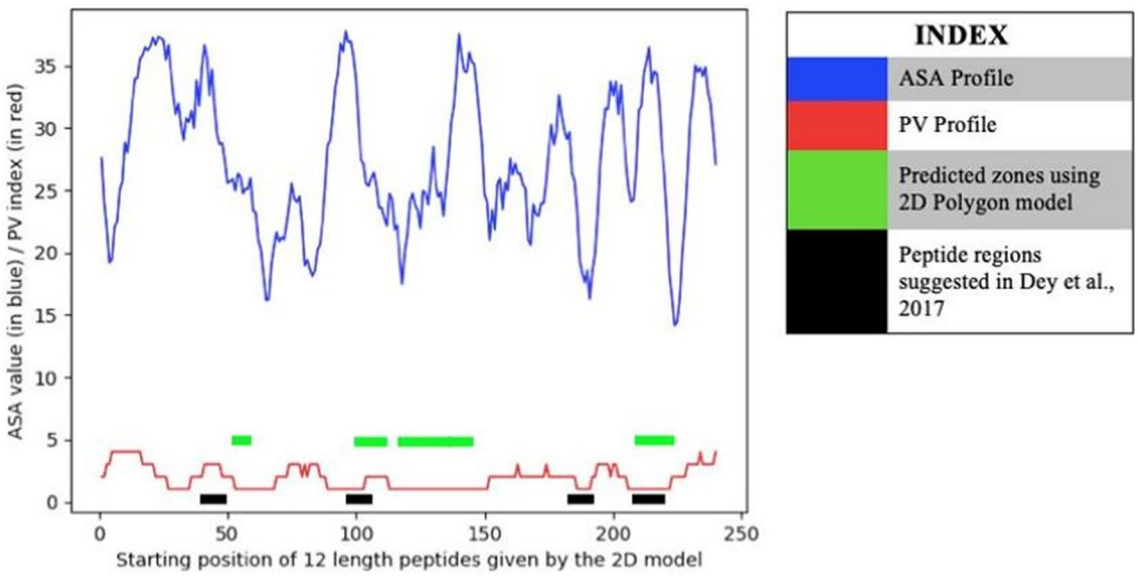

Fig. 4 ASA and PV profiles of the 251 aa fragment sequence of E protein of Zika virus: The components shown in the figure are: *ASA* profile (blue), *PV* profile (red), predicted peptide stretches using the new approach (green) and peptide stretches determined in Dey et al. (2017) (black). (Color figure online)

Table 12 The 9 grouped peptide zones obtained after grouping of the top 50 ranks for E protein of Zika virus

Starting position	Ending position	Consolidated Peptide stretch
212	233	PAQMAVDMQTLTPVGRLITANP**
120	145	AKRQTVVVLGSQEGAVHTALA-GALEA**
102	114	GTPHWNNKEALVE**
53	66	FGSLGLDCEPRTGL**
113	130	VEFKDAHAKRQTVVVLGS**
139	158	LAGALEAEMDGAKGRLFSGH
68	79	FSDLYYLTMNNK
38	51	EVTPNSPRAEATLG
188	201	PAETLHGTVTVEVQ

The zones marked with star indicate the ones which fell into the top 50th percentile as per the 2D Polygon analysis

Double asterisks refer to those peptide regions which fell in the top 50th percentile as per the 2D Polygon score for SARS-CoV-2, Hendra and Zika viruses respectively

been highlighted in star in Table 12. Considering the 5 grouped peptide zones obtained, we now performed the IEDB-AR study with the help of the MHC II DRB and DP/DQ alleles as mentioned in Paul et al. (2015), which can cover a majority of world population. Table 13 gives the results given by IEDB-AR analysis regarding the best possible 15-length epitopes for each corresponding peptide zone for both the cases of DRB and DP/DQ alleles. Accordingly, all the zones exhibited sufficient epitope potential as for each zone, we were able to find at least one 15-length epitope for either of the two types of alleles or both to have percentile rank below 10. As such, none of the 5 grouped peptide zones were rejected. In a similar way as before, we determined the corresponding intersecting zone for each of the 5 consolidated zones. Using the BLAST server, we once again checked whether any of these 5 intersecting zones show any kind of autoimmunity. Table 14 gives the best BLAST matches for each of the intersecting zones along with the E value of that match. All of the regions gave the minimum possible E

Table 13 Summary of the IEDB-AR study for the E protein where the best possible 15-length epitopes for both the MHC II DRB and DP/DQ analyses have been listed

Start–End	Grouped peptide zone	MHC II DP/DQ			MHC II DRB		
		Score	Adjusted peptide	Allele	Score	Adjusted peptide	Allele
212–233	PAQMAVDMQTLT-PVGRLITANP	20	VDMQTLTPVGR-LITA	HLA-DPA1*02:01/ DPB1*14:01	8.3*	PAQMAVDMQTLT-PVG	HLA-DRB4*01:01
120–145	AKRQTVVVLG-SQEGAVHTALA-GALEA	2*	SQEGAVHTALA-GALE	HLA-DQA1*05:01/ DQB1*03:01	7.7	SQEGAVHTALA-GALE	HLA-DRB1*09:01
102–114	GTPHWNNKEALVE	7.4*	TGTPH-WNNKEALVEF	HLA-DQA1*03:01/ DQB1*03:02	25	TGTPH-WNNKEALVEF	HLA-DRB1*13:02
53–66	FGSLGLDCEPRTGL	44	FGSLGLDCEPRT-GLD	HLA-DQA1*03:01/ DQB1*03:02	5.1*	ATLGGFGSLGLD-CEP	HLA-DRB1*15:01
113–130	VEFKDAHAKRQT-VVVLGS	7.9*	DAHAKRQTV-VVLGSQ	HLA-DPA1*02:01/ DPB1*14:01	21	VEFKDAHAKRQT-VVV	HLA-DRB5*01:01

The starred regions were chosen to form the intersecting zone corresponding to the particular peptide zone

The single asterisk marks represent the 15-length epitope chosen out of the two choices for every grouped peptide zone, such that the chosen ones form intersecting zones with the grouped peptide regions for SARS-CoV-2, Hendra and Zika viruses respectively

Table 14 Best possible BLAST matches and their corresponding E values for the 5 intersecting peptide zones to be finally shortlisted as peptide vaccine candidates for E protein of Zika virus

Intersecting zone	Best BLAST match	E value of the match	Accession/PDB ID
PAQMAVDMQTLTPVG	MON1 homolog A (yeast), isoform CRA_a in Homo sapiens	2.7	EAW65037.1
SQEGAVHTALAGALE	Protein PEAK3 in Homo sapiens	22	NP_940934.1
GTPHWNNKEALVE	Tumor protein p63-regulated gene 1-like protein in Homo sapiens	11	NP_877429.2
FGSLGLDCEP	Immunoglobulin heavy chain junction region in Homo sapiens	5.3	MBN4382203.1
DAHAKRQTVVVLGS	Immunoglobulin heavy chain junction region in Homo sapiens	2.8	MOP52445.1

Table 15 Comparison of the final shortlist for peptide candidates for E protein of Zika virus obtained from the new approach with that given in Dey et al. (2017), after the IEDB-AR and BLAST analyses

Finally shortlisted regions for E protein of Zika virus using the new approach		Finally shortlisted peptide regions for E protein of Zika virus listed in Dey et al. (2017)	
Start–End position	Peptide Stretch	Start–End position	Peptide stretch
212–226	PAQMAVDMQTLTPVG	215–223	MAVDMQTLT
130–144	SQEGAVHTALAGALE	–	–
102–114	GTPHWNNKEALVE	97–106	AGADTGTPHW
53–62	FGSLGLDCEP	–	–
117–130	DAHAKRQTVVVLGS	–	–
–	–	43–50	SPRAEATL
–	–	180–190	AAFTFSKVPAE

The comparison shows that two of the five regions we predicted were present among the published results in Dey et al. (2017)

value greater than 1, showing that they have negligible chances of causing autoimmune threats. So, the 5 intersecting zones, all of which passed both the IEDB-AR and BLAST tests, finally form the shortlist of possible vaccine candidates against Zika virus based on its E protein. On comparison of these regions with the ones given in Dey et al. (2017), it was observed that two of them were found among the established regions, showing there was a decent level of consistency between the two approaches for this test case. The above comparison has been shown in Table 15.

Moreover, the five regions predicted by the current method have been depicted in green horizontal lines in Fig. 4 whereas, the shortlisted regions listed in Dey et al. (2017) have been marked in black horizontal lines for better representation of the comparison. Also, in the figure, the regions 117–130 and 130–144 have been depicted as a single uniform horizontal stretch.

Conclusion

The main objective of the research reported in this paper is the development and applications of a new quantitative approach for the formulation of peptide vaccine libraries for the surging SARS-CoV-2 virus. So far, regarding our earlier approach towards rational peptide vaccine design against different viruses, eye-estimation of the superimposed graphical profiles for ASA and PV of the viral sequence was employed to find surface-exposed and conserved peptide regions of the viral sequence for further analysis to prescribe the most suitable peptide regions for vaccine design. But this manual method often incorporates bias and a slight deviation from precision, which is not desirable. In this perspective, the two-stepped approach presented in this article turns out to be useful in replacing the tedious manual intervention for scanning through the graphical profiles, with this new precise computational model by automating the search procedure for suitable peptide candidates. Along with this, the robust and simple nature of the approach makes it easier to execute it through various programming platforms. In the analysis presented here, the most conserved and surface-exposed peptide regions in the spike glycoprotein of SARS-CoV-2 have been identified for in silico peptide vaccine design against the virus, with the help of this newly devised approach. Finally, 4 peptide regions have been shortlisted for SARS-CoV-2 using this model. A similar study for determining possible vaccine candidates for SARS-CoV-2 has also been done using the older method of eye-estimation in the current analysis using the same sequences as used for the 2D model. A good level of matching among both the results shows that the new approach is indeed consistent in predicting the peptide vaccine candidates for SARS-CoV-2. Further comparison was also made with results given by other researchers using the same spike protein of SARS-CoV-2. As such, some of our zones matched with these analyses as well.

However, the design of the approach doesn't end here. We still need to see whether this method is applicable for other viruses as well. In that case, the method will be able to predict vaccine candidates for any other virus in general. For this verification with other viruses, we chose two test cases—the first one was to predict vaccine candidates for Hendra virus using its surface-situated G glycoprotein, and the second test case was to perform the same for Zika virus using its surface-situated E protein. For the analysis on Hendra virus, the new method yielded two peptide regions as the finally shortlisted candidates, both of which matched with a pre-established set of vaccine candidates for Hendra virus. Again, for Zika virus E protein, we obtained 5 finally shortlisted peptide stretches, out of which two regions matched with a similar pre-established set of peptide vaccine candidates for Zika virus. This shows that the method was also able to predict peptide vaccine candidates with a decent level of consistency for other viruses as well.

Thus, it can be said that this new computational approach has paved way for a much improvised in silico

and alignment-free technique to design peptide vaccines not only with respect to the current pandemic caused by SARS-CoV-2, but for any other new virus in future as well. Regarding the current situation where the pandemic is taking a toll on many lives, by automating the selection of the best peptide candidates, this approach will help to generate a much more rapid solution to the issue of vaccine design.

The software application that we have prepared for the execution of the entire analysis described in this article can be obtained by downloading the relevant installation setup given in the following GitHub repository link: https://github.com/SubhamoyBiswas/Installation-Setup-for-Peptide-Vaccine-Analysis-Tool-PVAT.

Acknowledgements We would like to acknowledge Dr. Sumanta Dey from Centre for Interdisciplinary Research and Education (CIRE), Kolkata, India for providing the accession IDs of the sequences to be analyzed for Zika and Hendra viruses, required for verifying the performance of the model before applying to the case of SARS-CoV-2. We would also like to acknowledge Mr. Shreyans Chatterjee from Department of Microbiology, St. Xavier's College, Kolkata, India for helping us in better understanding of the workflow of T-cells and Mr. Tathagata Dutta, also from CIRE, Kolkata, for his ideas in the initial period of the exercise.

Author Contributions SB: Conceptualization, Methodology, Software, Validation, Data Curation and Writing—Original Draft. SM: Validation, Writing—Original Draft, Writing—Review & Editing and Supervision. AN: Resources, Writing—Review & Editing and Supervision. SCB: Resources, Writing—Review & Editing and Supervision.

Funding No funding source has been involved in this work.

Data Availability As supplementary data, we have attached the predictions given by IEDB-AR using the HLA alleles MHC II DRB and DP/DQ for the spike protein of SARS-CoV-2 in the following link: https://drive.google.com/drive/folders/1O4pPlK6PaLWXJciADQF52qpMhuM25t2h.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdelmageed MI, Abdelmoneim AH, Mustafa MI, Elfadol NM, Murshed NS, Shantier SW, Makhawi AM (2020) Design of a multiepitope-based peptide vaccine against the E protein of human COVID-19: an immunoinformatics approach. Biomed Res Int 2683286:2314–6133. https://doi.org/10.1155/2020/2683286

Adamczak R, Porollo A, Meller J (2004) "accurate prediction of solvent accessibility using neural networks based regression", proteins: structure. Funct Bioinform 56:753–767. https://doi.org/10.1002/prot.20176

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Backert L, Kohlbacher O (2015) Immunoinformatics and epitope prediction in the age of genomic medicine. Genom Med 7:119. https://doi.org/10.1186/s13073-015-0245-0

Badawi MM, Osman MM, Alla AAEF, Ahmedani AM, Abdalla MH, Gasemelseed MM, Elsayed AA, Salih MA (2016) Highly conserved epitopes of ZIKA envelope glycoprotein may act as a novel peptide vaccine with high coverage: immunoinformatics approach. Am J Biomed Res 4:46–60. https://doi.org/10.12691/ajbr-4-3-1

Biswas S, Dey T, Chatterjee S, Manna S, Nandy A, Das S, Nandy P, Basak SC (2019) A novel approach to Peptide Vaccine Design for Ebola virus. In: 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. https://doi.org/10.3390/mol2net-05-06712

Biswas S, Chatterjee S, Dey T, Dey S, Manna S, Nandy A, Basak SC (2020) In Silico Approach for Peptide Vaccine Design for CoVID 19. In: MDPI AG in MOL2NET 2020, International conference on multidisciplinary sciences, 6th edition session USINEWS-04: US-IN-EU Worldwide Science Workshop Series, UMN, Duluth, USA, 2020. https://doi.org/10.3390/mol2net-06-06787

Chakraborty S, Chakravorty R, Ahmed M, Rahman A, Waise TMZ, Hassan F, Rahman M, Shamsuzzaman S (2010) A computational approach for identification of epitopes in dengue virus envelope protein: a step towards designing a universal dengue vaccine targeting endemic regions. Silico Biol 10:235–246. https://doi.org/10.3233/ISB-2010-0435

Chit A, Parker J, Halperin SA, Papadimitropoulos M, Krahn M, Grootendorst P (2014) Toward more specific and transparent research and development costs: the case of seasonal influenza vaccines. Vaccine 32(26):3336–3340. https://doi.org/10.1016/j.vaccine.2013.06.055

Dagur HS, Dhakar SS, Gupta A (2020) Epitope-based vaccine design against novel coronavirus SARS-CoV-2 envelope protein. EJMO 4(3):201–208. https://doi.org/10.14744/ejmo.2020.01978

Dar H, Zaheer T, Rehman MT, Ali A, Javed A, Khan GA, Babar MM, Waheed Y (2016) Prediction of promiscuous T-cell epitopes in the Zika virus polyprotein: an in silico approach. Asian Pac J Tropical Med 9:844–850. https://doi.org/10.1016/j.apjtm.2016.07.004

Dey S, De A, Nandy A (2016) Rational design of peptide vaccines against multiple types of human papillomavirus. Cancer Inform. https://doi.org/10.4137/CIN.S39071

Dey S, Nandy A, Basak SC, Nandy P, Das S (2017) A Bioinformatics approach to designing a Zika virus vaccine. Comput Biol Chem 68:143–152. https://doi.org/10.1016/j.compbiolchem.2017.03.002

Dey S, Roy P, Dutta T, Nandy A, Basak SC (2018) Rational design of peptide vaccines for highly lethal nipah and hendra viruses. BioRxiv. https://doi.org/10.1101/425819

Di Paola L, Hadi-Alijanvand H, Song X, Hu G, Giuliani A (2020) The discovery of a putative allosteric site in the SARS-CoV-2 spike protein using an integrated structural/dynamic approach. J Proteome Res 19(11):4576–4586. https://doi.org/10.1021/acs.jproteome.0c00273

Dudek NL, Perlmutter P, Aguilar M, Croft NP, Purcell AW (2010) Epitope discovery and their use in peptide based vaccines. Curr Pharm Des 16:3149. https://doi.org/10.2174/138161210793292447

Durojaye OA, Mushiana T, Cosmas S, Ibiang GO, Ibiang MO (2020) An in silico epitope-based peptide vaccine design against the 2019-nCoV. Egypt J Med Hum Genet 21:35. https://doi.org/10.1186/s43042-020-00071-7

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. https://doi.org/10.1371/journal.pone.0040749
Islam R, Sakib MS, Zaman A (2012) A computational assay to design an epitope-based peptide vaccine against Chikungunya virus. Future Virol 7:1029–1042. https://doi.org/10.2217/fvl.12.95
Kalita P, Padhi AK, Zhang KYJ, Tripathi T (2020) Design of a peptide-based subunit vaccine against novel coronavirus SARS-CoV-2. Microbial Pathog 145:104236. https://doi.org/10.1016/j.micpath.2020.104236
Kerfeld CA, Scott KM (2011) Using BLAST to teach "E-value-tionary" concepts. PLoS Biol 9(2):e1001014. https://doi.org/10.1371/journal.pbio.1001014
Li W, Joshi MD, Singhania S, Ramsey KH, Murthy AK (2014) Peptide vaccine: progress and challenges. Vaccines 2:515–536. https://doi.org/10.3390/vaccines2030515
Lo YT, Pai TW, Wu WK, Chang HT (2013) Prediction of conformational epitopes with the use of a knowledge-based energy function and geometrically related neighboring residue characteristics. BMC Bioinform 14:S3. https://doi.org/10.1186/1471-2105-14-S4-S3
Mirza MU, Rafique S, Ali A, Munir M, Ikram N, Manan A, Salo-Ahen OMH, Idrees M (2016) Towards peptide vaccines against Zika virus: immunoinformatics combined with molecular dynamics simulations to predict antigenic epitopes of Zika viral proteins. Sci Rep 6:37313. https://doi.org/10.1038/srep37313
Nandy A, Ghosh A, Nandy P (2009) Numerical characterization of protein sequences and application to voltage-gated sodium channel α subunit phylogeny. In Silico Biol 9:77–87. https://doi.org/10.3233/ISB-2009-0389
Nandy A, Dey S, Roy P, Basak SC (2018) Epidemics and peptide vaccine response: a brief review. Curr Top Med Chem 18(26):2202–2208. https://doi.org/10.2174/1568026618666181112144745
Paul S, Lindestam Arlehamn CS, Scriba TJ, Dillon MB, Oseroff C, Hinz D, McKinney DM, Carrasco Pro S, Sidney J, Peters B, Sette A (2015) Development and validation of a broad scheme for prediction of HLA class II restricted T cell epitopes. J Immunol Methods 422:28–34. https://doi.org/10.1016/j.jim.2015.03.022
Poland GA, Kennedy RB, Ovsyannikova IG (2011) Vaccinomics and personalized vaccinology: is science leading us toward a new path of directed vaccine development and discovery? PLoS Pathog 7:e1002344. https://doi.org/10.1371/journal.ppat.1002344
Poland GA, Whitaker JA, Poland CM, Ovsyannikova IG, Kennedy RB (2016) Vaccinology in the third millennium: scientific and social challenges. Curr Opin Virol 17:116–125. https://doi.org/10.1016/j.coviro.2016.03.003
Ponomarenko JV, Bui H, Li W, Fusseder N, Bourne PE, Sette A, Peters B (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinform 9:514. https://doi.org/10.1186/1471-2105-9-514
Porollo A, Adamczak R, Wagner M, Meller J (2003) "Maximum feasibility approach for consensus classifiers: applications to protein structure prediction", CIRAS (conference proceedings)
Purcell AW, McCluskey J, Rossjohn J (2007) More than one reason to rethink the use of peptides in vaccine design. Nat Rev 6:404–414. https://doi.org/10.1038/nrd2224
Rappuoli R (2001) Reverse vaccinology, a genome-based approach to vaccine development. Vaccine 19:2688–2691. https://doi.org/10.1016/s0264-410x(00)00554-5
Riedel S (2005) Edward Jenner and the history of smallpox and vaccination. Proc (baylor Univ Med Cent) 18(1):21–25. https://doi.org/10.1080/08998280.2005.11928028
Shawan MMAK, Mahmud HA, Hasan MM, Parvin A, Rahman MN (2014) In Silico Modeling and Immunoinformatics probing disclose the epitope based peptide vaccine against Zika virus envelope glycoprotein. Ind J Pharma Biol Res 2:44–57. https://doi.org/10.30750/ijpbr.2.4.10
Slathia P, Sharma P (2020) Prediction of T and B cell epitopes in the proteome of SARS-CoV-2 for potential use in diagnostics and vaccine design. ChemRxiv. https://doi.org/10.26434/chemrxiv.12116943.v1
Tomar N, De RK (2014) Immunoinformatics: a brief review. In: De R, Tomar N (eds) Immunoinformatics. Methods in molecular biology (methods and protocols), vol 1184. Humana Press, New York. https://doi.org/10.1007/978-1-4939-1115-8_3
Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B (2018) The immune epitope database (IEDB): 2018 update. Nucleic Acids Res. https://doi.org/10.1093/nar/gky1006
Wagner M, Adamczak R, Porollo A, Meller J (2005) Linear regression models for solvent accessibility prediction in proteins. J Comput Biol 12:355–369. https://doi.org/10.1089/cmb.2005.12.355
Yazdani Z, Rafiei A, Yazdani M, Valadan R (2020) Design an efficient multi-epitope peptide vaccine candidate against SARS-CoV-2: an in silico analysis. BioRxiv. https://doi.org/10.1101/2020.04.20.051557