Journal of Biotechnology **14**(4): 605-618, 2016

COMPUTATIONAL APPROACH FOR SELECTION OF EPITOPE-BASED DENGUE VACCINE TARGETS

Phuc Nguyen[1], Ly Le[1, 2, ✉]

[1]*International University, Vietnam National University Ho Chi Minh City*
[2]*Ho Chi Minh Institute for Computational Science and Technology*

✉ To whom correspondence should be addressed. E-mail: ly.le@hcmiu.edu.vn

Received: 11.9.2016
Accepted: 28.12.2016

SUMMARY

High antigenic variability in the envelope (E) protein of different virus strains has been a major obstacle in designing effective vaccines for Dengue virus (DENV). To maintain their biological function, some parts of viral proteins remain stable during evolution thus one possible approach to solve this problem is to recognize specific regions within different protein sequences of E that have the tendency to stay constant through evolution. These regions may possess some special attributes to become a vaccine candidate against dengue virus. In this study, a computational approach was utilized to identify and analyze highly conserved amino acid sequences of the DENV E protein. Sequences of 9 amino acids or more were specifically focused due to their immune-relevant as T-cell determinants. Different bioinformatics tools were responsible for revealing conserved regions in the DENV E protein and constructing the phylogenetic tree from the sequence database. The tools also predicted immunogenicity of the identified vaccine targets. Ultimately, two peptide regions of at least 9 amino acids were chosen due to their high conserved attribute in more than 95% of all collected DENV sequences. Moreover, both of them was found to be immune-relevant by their correspondence to known or putative HLA-restricted T cell determinants. The conserved attribute of these sequences through the entire analysis of this study supports their potential as candidates for further *in vitro* experiments for rational design a universal vaccine which has longer and broader impact.

Keywords: *Bioinformatics; conserved regions; Dengue; envelope protein; phylogenetic tree; T-cell epitopes; HLA*

INTRODUCTION

Dengue is currently listed as one of the neglected tropical diseases even though it has plagued human populations for at least 250 years (Hotez *et al.,* 2014). Millions of people are at risk of infections that may lead to fatal symptoms such as Dengue Fever (DF) and Dengue Hemorrhagic Fever (DHF). An estimated 50% of the global human population is exposed to dengue fever transmitted by the *Aedes aegypti* or *Ae. albopictus* mosquito (Rigau-Pérez *et al.,* 1998). Up to now, dengue virus (DENV) has spread to more than 100 countries, mainly tropical and subtropical areas. Statisticians estimated that there are millions of infections and thousands of deaths caused by DENV each year (Gubler *et al.,* 2014).

The dengue is caused by dengue viruses (DENV), which form the dengue complex in the genus *Flavivirus*, family *Flaviviridae*. There are four serologically distinct types of dengue virus: DENV-1, DENV-2, DENV-3, and DENV-4. There are three main structural proteins of DENV: capsid, pre-membrane (prM), and envelope (E) protein. One potential way to develop vaccines against DENV entry focused on E protein (Murrell *et al.,* 2011). It is because E protein assumes different conformations in the life cycle of the virus (Kuhn *et al.,* 2002; Rey, 2003; Perera & Kuhn, 2008). Therefore, it is thought to be responsible for many important roles such as mediating viral entry by interacting with host cell surface receptors (Rey, 2003). E protein is also the primary target of adaptive immune response (Murphy & Whitehead, 2011).

Many approaches have been directed at the DENV E protein such as viral receptors, hydrophobic pockets, monoclonal antibodies (Murrell *et al.,* 2011). However, none of these

studies demonstrated an effective way to combat DENV due to various reasons (Murrell *et al.,* 2011). The most common obstacles are the lack of knowledge of DENV E protein involvement in the infection process and the high variability of DENV (Murrell *et al.,* 2011). Currently, there is no licensed vaccine, antiviral drugs or specific treatment against dengue (Gubler *et al.,* 2014). Preventive measures are the best strategy, which consists mainly of environmental management, spraying insecticides, and personal protective measures (Murrell *et al.,* 2011; Mustafa *et al.,* 2015).

In recent years, the fields of bioinformatics and immune-informatics are emerging rapidly suggesting alternative solutions to the DENV problems. Many complex problems such as understanding immune responses and vaccine/drug design have been solved with the combinations of experimental and in silico methods (Korber *et al.,* 2006). Bioinformatics tools have the ability to systematically scan for candidate epitopes from larger sets of protein antigens such as those encoded by complete viral genomes. Considerable time and costs could be saved by this approach. The strategy of computer-aided vaccine design seems to be quite effective in dealing diseases such as multiple sclerosis (Bourdette *et al.,* 2005), malaria (López *et al.,* 2001), and tumors (Knutson *et al.,* 2001).

With the advent of computer-aided vaccine design method, a throughout investigation of epitopes on DENV E protein using bioinformatics tools is necessary for understanding viral function and for the design of unique polyvalent vaccines capable of inducing a neutralizing antibody response against each DENV serotype. In this study, the E protein from all four types of dengue virus was analyzed using bioinformatics and immune-informatics tools for identification of an immunogenic hot spot that may be used as a peptide vaccine candidate.

MATERIALS AND METHODS

Data collection

To obtain maximum geographical and chronological diversity, the envelope protein sequences from many variants belong to all 4 type of dengue virus (DENV-1, DENV-2, DENV-3 and DENV-4) were retrieved from Uniport protein database (www.uniprot.org) (Consortium, 2014). Different variants of E protein (isolated from different geographical locations in different time points) were manually selected from the above proteomics database (Table 1). The following criteria were used when compiling the database (Twiddy *et al.,* 2003; Costa *et al.,* 2012): (1) all sequences had a known date of collection and geographic location; (2) sequences collected in the same year from the same geographical location were excluded. This resulted in a most updated database of 688 DENV E protein sequences from 78 countries covering the temporal range of 51 years (1961 – 2012).

Table 1. Characteristics of DENV database used in this study.

Serotype	Number of sequences	Time intervals (years)
DENV-1	131	1972-2012
DENV-2	218	1969-2012
DENV-3	184	1974-2012
DENV-4	155	1961-2012

Multiple sequence alignment

The alignments were generated on the basis of different algorithms of multiple sequence alignment calculation for higher accuracy and consistency. Sequences were first aligned in MUSCLE (Edgar, 2004), and then the alignment was verified and corrected if necessary in 3 other different algorithms which have been widely used and previously proved accurate in MSA: CLUSTALW (Thompson *et al.,* 2002), MAFFT (Katoh *et al.,* 2002) and Kalign (Lassmann & Sonnhammer, 2005). All of these algorithms were already integrated into Unipro UGENE package upon installation for the ease of use (Okonechnikov *et al.,* 2012).

Construction of phylogenetic tree

The multiple sequence alignment (MSA) was generated from the retrieved protein sequences. This MSA formed the basis for construction of phylogenetic tree, which was used to elucidate the evolutionary distance among the envelope proteins as well as to identify the conserved regions in the envelope proteins. Unipro UGENE package was employed to construct and visualize the un-rooted phylogenetic tree (Okonechnikov *et al.,* 2012). MUSCLE was utilized for generating MSA and PHYLIP neighbor-joining method was used for the construction of phylogenetic tree (Edgar, 2004). Evolutionary distances that were used to infer the

phylogenetic tree were computed using the Kimura distance matrix model with 1000 bootstrap replicates (Lockhart *et al.,* 1994).

Identification of conserved regions, termed as pan-DENV sequences

The DENV protein sequences were examined by a consensus sequence based approach (Novitsky *et al.,* 2002) to identify sequence fragments that were common across the 4 types. These conserved sequence fragments are also known as pan-DENV sequences. The consensus sequences for the envelope protein of each type (intra-type consensus) were first derived by multiple sequence alignments to select the predominant residue at each amino acid position. The 4 intra-type consensus sequences for envelope protein (one from each type) were then aligned to reveal sequence fragments identical across each of the types that were at least 9 amino acids long. Only sequence fragments that were identical in at least 80% of the sequences of each of the 4 types were retained for further analyses.

Information entropy analysis of pan-DENV sequences

Shannon information entropy (Strait & Dewey, 1996; Khan *et al.,* 2006) was used to study the diversity of DENV protein sequences within each type (intra-type diversity) and across all DENVs (pan-DENV diversity) and to assess the predicted evolutionary stability of the identified pan-DENV sequences. Entropy, H, representing the variability of nonamer peptides (peptides that are composed of 9 amino acids) centered at any given alignment site, is computed from the probability, pa of each nonamer peptide occurring at that site:

$$H = -\sum_{a} p_a \log_2(p_a)$$

All entropy analyses were carried out by using BioEdit (Hall, 1999).

Functional and structural correlates analyses of pan-DENV sequences

The known and putative structural and functional properties of pan-DENV sequences were elucidated by used of the Prosite (Hulo *et al.,* 2006), via ScanProsite (Gattiker *et al.,* 2002), and Pfam (Punta *et al.,* 2011) databases which comprise information on protein families, domains and functional site. Then, conserved sequences were mapped on the three-dimensional (3-D) structures of DENV in the PDB database (Sussman *et al.,* 1998) by use of ConSurf (Armon *et al.,* 2001). The variability pattern plotted on the DENV envelope protein structure was visualized by UCSF Chimera package (Pettersen *et al.,* 2004). The PDB ID of 3-D crystal structures of DENV-1, DENV-2, DENV-3 and DENV-4 envelope protein used in this study respectively were 3G7T, 1OAN, 1UZG, and 3UAJ (Sussman *et al.,* 1998).

Identification of pan-DENV sequences common to other viruses and organisms

Pan-DENV sequences that overlapped at least 9 consecutive amino acid sequences of other viruses and organisms were identified by performing BLAST (Altschul *et al.,*1997) search against viral protein sequences reported at NCBI, excluding DENV sequences, and against protein sequences of all organisms excluding viruses (Boratyn *et al.,* 2012). This process was completed by choosing "Exclude" option in "Organism" field for Dengue virus (taxid: 10052 and 12637); Dengue virus 1 (taxid: 11053); Dengue virus 2 (taxid: 11060), Dengue virus 3 (taxid: 11069) and Dengue virus 4 (taxid: 11070).

Identification of known and predicted pan-DENV HLA super-type binding sequences

The Immune Epitope Database (IEDB: www.iedb.org) (Vita *et al.,* 2010) were utilized to conduct series of search for reported immunogenic, human T-cell determinants (both class I and II) of DENV that either fully or partially overlapped with the pan-DENV sequences. Moreover, dedicated algorithms based on several prediction models were used to identify candidate putative HLA-binding sequences to multiple HLA class I and II supertype alleles within the pan-DENV sequences. Putative HLA supertypes class I-restricted peptides were identified by NetCTL (Larsen *et al.,* 2005); Artificial neural network (ANN) (Nielsen *et al.,* 2003); Stabilized matrix method (SMM) (Peters, Sette, 2005), and class II-restricted peptides by NN-align (Nielsen, Lund, 2009) and SMM-align (Nielsen *et al.,* 2007).

RESULTS AND DISSCUSSION

Multiple sequence alignment

The DENV E protein sequences generally exhibit low intra-type but high inter-type variability. More than 3/4 amino acid positions are highly conserved (≥ 80% identity) within each serotype.

However, when the whole 688 sequences are aligned, this level of conservation is significantly reduced (≈ 60% identity).

Evolution divergence of the DENV E protein

E protein sequences from 691 variants belonging to 4 types of dengue virus were retrieved from the database. MSA generated from these sequences were exploited to construct a phylogenetic tree which showed the evolutionary divergence among the envelope protein sequences. A significant evolutionary divergence was observed in the phylogram obtained for the E proteins. Most of the four serotypes of DENV were clearly arranged in well-define clades with high statistical support (Figure 1). In this result, DENV-1 is closest to the hypothetical ancestor, thus, it may be least evolved member in the group. However, when comparing DENV-1 with DENV-4 and DENV-3 clusters, the distances among them are not significant, possibly meaning that these 3 stereotypes show little differences in their sequence composition (Figure 1). Overall, DENV-4 is the most evolved and divergent among the 4 serotypes. This high variety in DENV-4 sequences may create great obstacles in development of effective polyvalent vaccines against 4 serotypes.

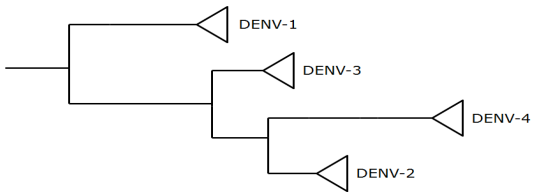


Figure 1. Summarized DENV phylogeny, picturing the evolution of the serotypes.

Conserved pan-DENV sequences

From the outputs of the MSA, there were 2 pan-DENV sequences of at least 9 amino acids that were present in ≥80% of all sequences of each DENV type (Figure 2; Table 2). The size of the pan-DENV sequences ranged from 10 to 15, with a combined size of 25 residues, corresponding approximately to 5% of the complete DENV E protein (≈495 amino acid).

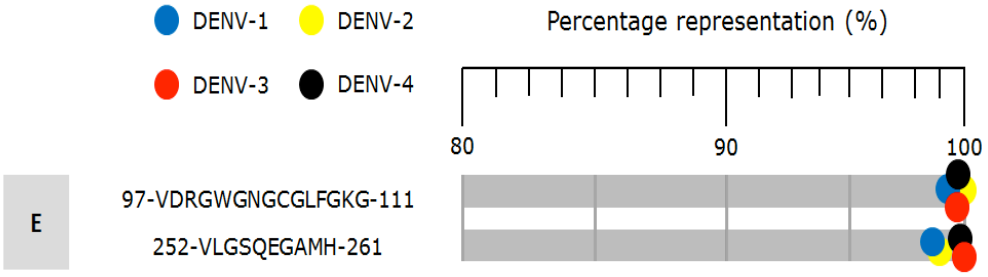


Figure 2. Pan-DENV sequences and their representations in the 4 DENV types. The 2 pan-DENV sequences of at least 9 amino acids that were found present ≥80% of the collected sequences of each DENV type are shown. Amino acid positions were numbered according to the sequence alignments of the 4 DENV types.

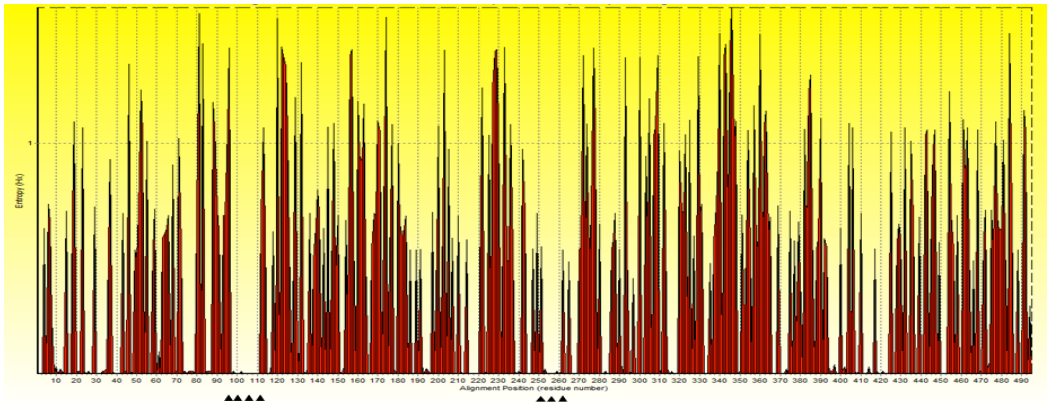
Table 2. The intra-type percentage representation of pan-DENV sequences. [a] Amino acid positions numbered according to the sequence alignments of the 4 DENV types; [b] Rounded to 1 decimal place.

DENV protein	Pan-DENV sequence [a]	% intra-type representation [b]			
		DENV-1	DENV-2	DENV-3	DENV-4
E 97-111	$_{97}$VDRGWGNGCGLFGKG$_{111}$	100	99.5	99.5	99.4
E 252-261	$_{252}$VLGSQEGAMH$_{261}$	99.2	99.1	100	100

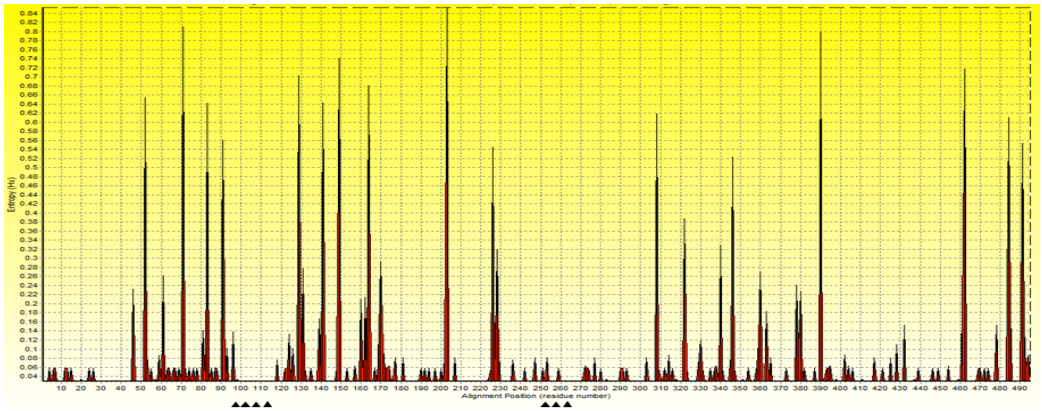
Evolutionary diversity of DENV protein nonamer peptide sequences

By applying Shannon information entropy, the evolutionary diversity of each DENV type, and all of 4 serotypes, was investigated (Strait, Dewey, 1996). The result showed numerous long regions of low entropy (≤1), illustrating the relatively high degree of intra-type sequence conservation. In general, the average intra-type nonamer entropy value of E protein of DENV-1, -2, -3 and 4 approximately ranged from 0.05 to 1

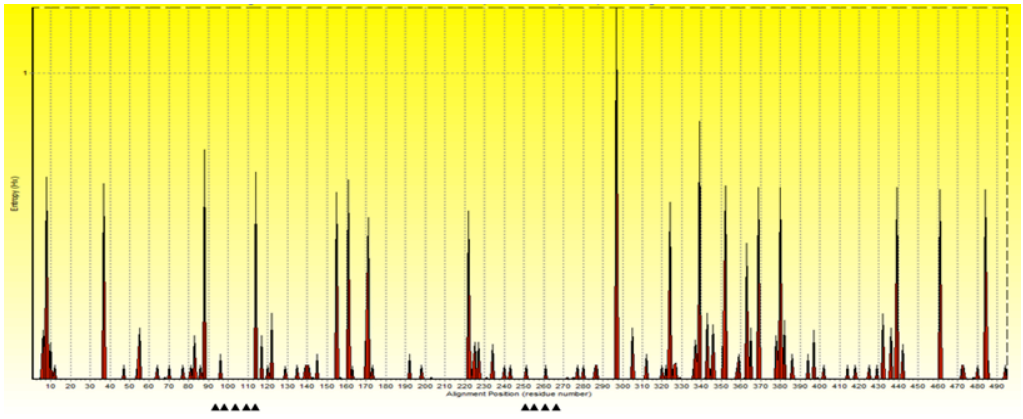
(Figure 3; Table 3). Both of the 2 pan-DENV sequences identified earlier had entropy value ≤0.3 corresponding to the intra-type representation ≥90% (Khan *et al.,* 2006). Therefore, by applying appropriate consensus- and entropy-based analyses, highly conserved and evolutionary stable pan-DENV sequences in E proteins are illuminated, in spite of the distinct viral diversity defining multiple DENV variants (Holmes, Burch, 2000).



A



B



C

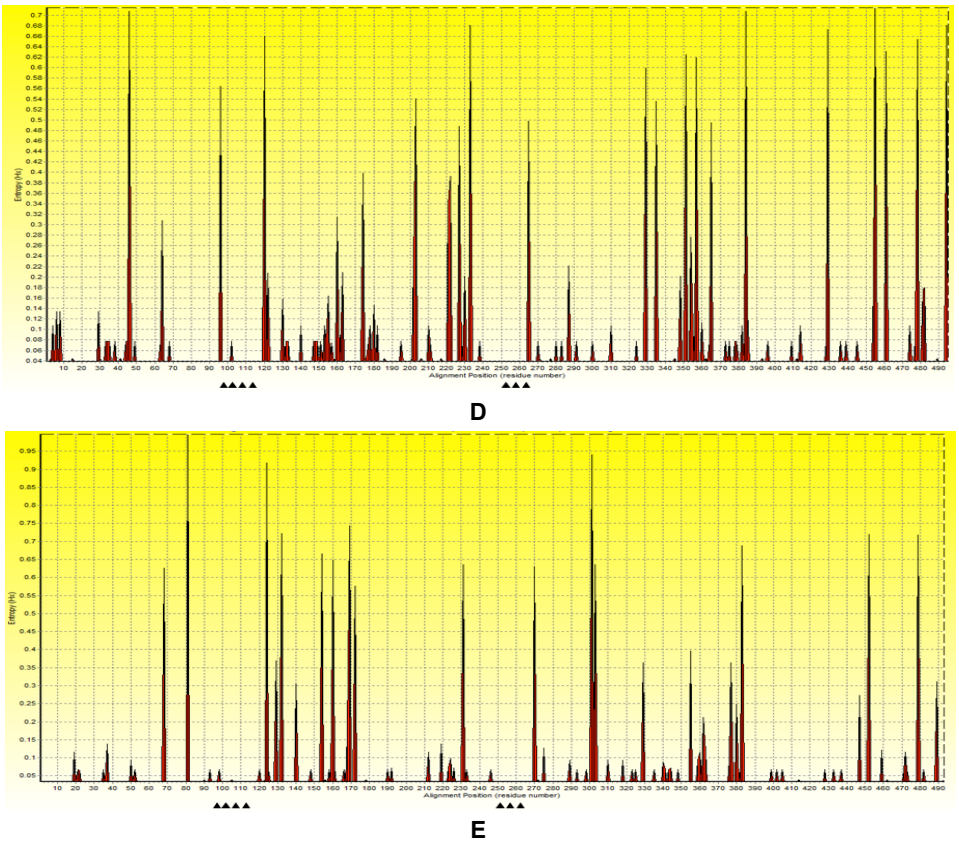


Figure 3. Shannon entropy of nonamer peptides within and across DENV type sequences. The entropy values were computed from the alignments of DENV sequences using the BioEdit software. Values were plotted for all 4 DENV types (A), DENV-1 (B), DENV-2 (C), DENV-3 (D), and DENV-4 (E) sequences. The triangles below indicate the locations of the pan-DENV sequences in the corresponding proteins.

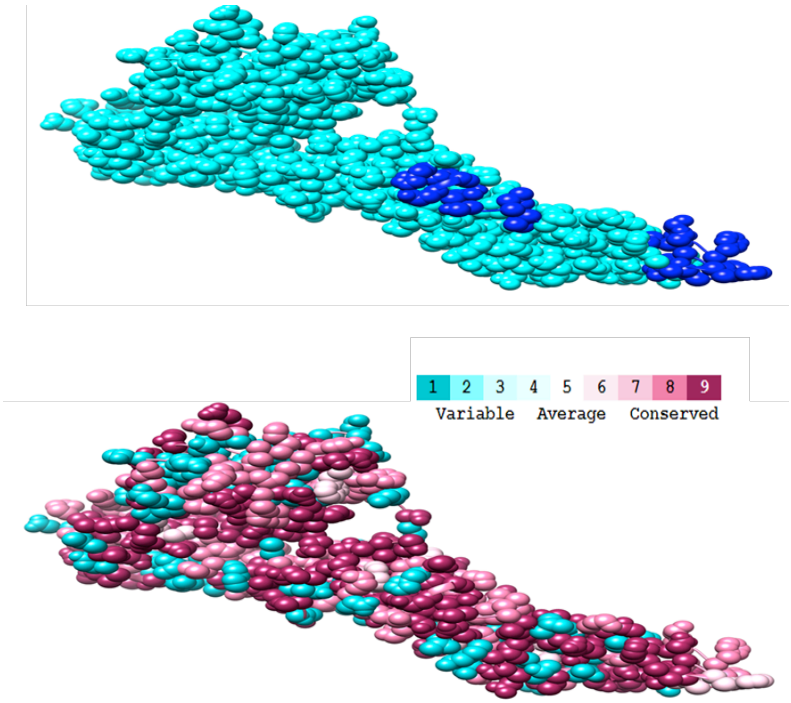
Table 3. Pan-DENV sequences, entropy and representation of variants. [a] Amino acid positions numbered according to the sequence alignments of the 4 DENV types; [b] Maximum nonamer peptide entropy across all DENV sequences (rounded to 1 decimal place); [c] Minimum and maximum percentage representation of sequences that different from pan-DENV sequences in the whole database (rounded to whole number).

DENV protein	Pan-DENV sequence [a]	Peptide entropy [b]	% variant representation [c]
E 97-111	$_{97}$VDRGWGNGCGLFGKG$_{111}$	0.2	1
E 252-261	$_{252}$VLGSQEGAMH$_{261}$	0.2	1

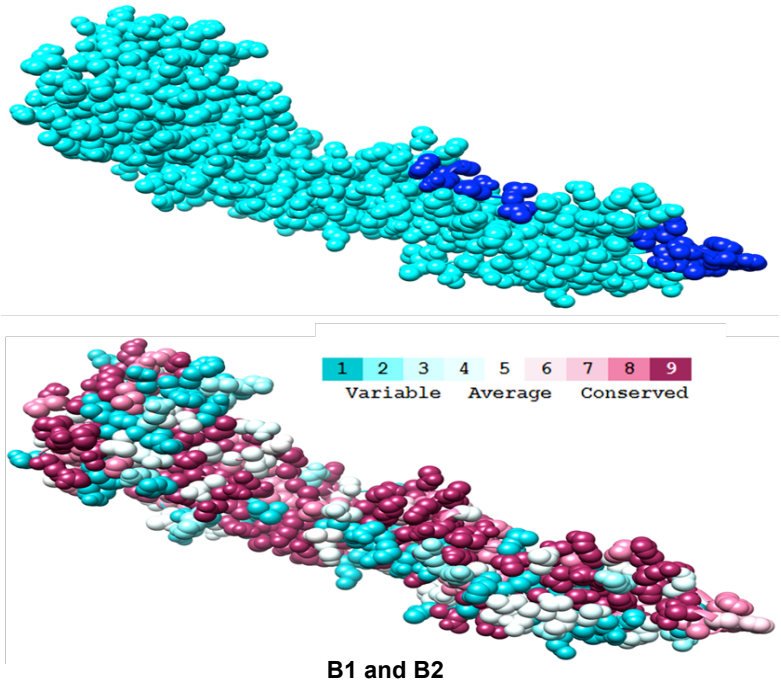
Functional and structural correlates of the pan-DENV sequences

Critical sites and domains are likely to be represented by highly conserved protein sequences (Valdar, 2002). A search of the literature and the Prosite and Pfam databases (Hulo *et al.,* 2006; Punta *et al.,* 2011) revealed that both of the identified pan-DENV sequences were associated with biological activities (Table 4). They are corresponded to the fusion peptide (position 98 to 110) and dimerization domain (Allison *et al.,* 2001; Modis *et al.,* 2004).

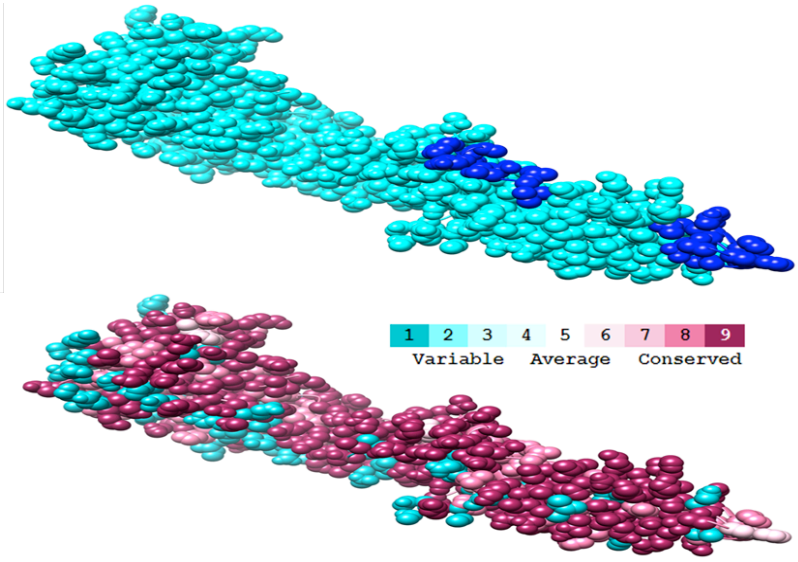
It is generally recognized that amino acids buried inside proteins are subject to greater interactions and packing constraints than those exposed on the outer surface (Haydon & Woolhouse, 1998). The mutational variability was investigated and visualized with the aid of ConSurf (Glaser *et al.,* 2003). Based on the Consurf results, both of the pan-DENV sequences were exposed at the surface of the corresponding structures but still fairly conserved (Figure 4). In addition, there were little differences among the 3-D structure of each DENV type.



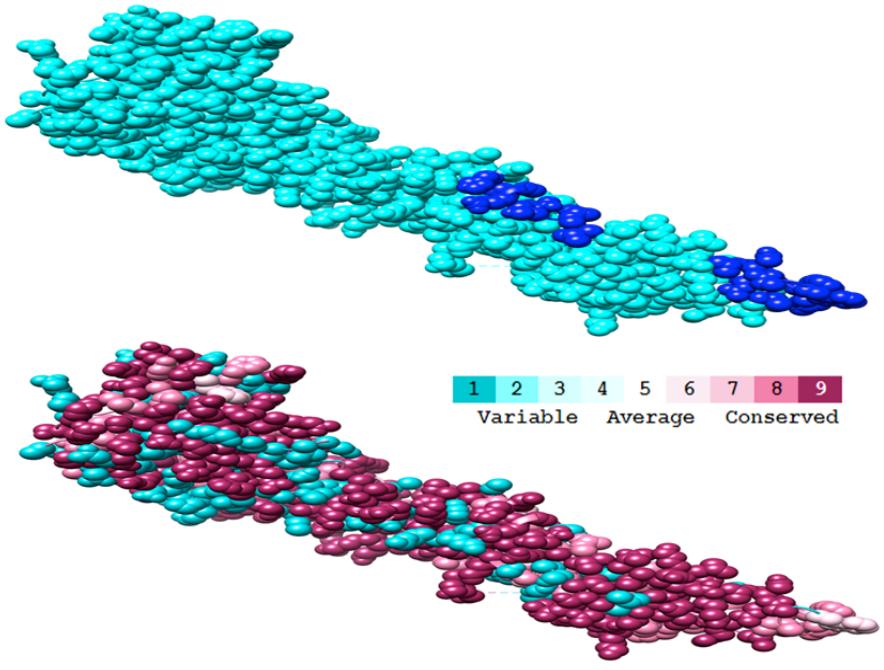
A1 and A2



B1 and B2



C1 and C2



D1 and D2

Figure 4. The mutational variability color scale plotted on the chain A of protein structure of each DENV type with the aid of Consurf and visualized by Chimera. There are 2 sub-figures for each DENV type; one is for mapping the location of pan-DENV sequences on the DENV 3-D structure (*colored in blue*) and the other to illustrate the conserved level of pan-DENV sequences calculated by Consurf (*results based on the colors*): *DENV-1* (A1, A2); *DENV-2* (B1, B2); *DENV-3* (C1, C2) and *DENV-4* (D1, D2).

Table 4. Functional and structural properties of pan-DENV sequences. [a] Amino acid positions numbered according to the sequence alignments of the 4 DENV types; [b] Described in the literature and/or identified using the Prosite (Hulo *et al,*., 2006) and Pfam (Bateman *et al.*,, 2004) databases. Prosite (PS) and Pfam (PF) accession numbers: PS00008, N-Myristoylation; PF00869, Dimerisation Domain.

DENV protein	Pan-DENV sequence [a]	Functional domains and motifs [b]	Putative post-transcriptional modifications [b]
E 97-111	$_{97}$VDRGWGNGCGLFGKG$_{111}$	Dimerization Domain, Fusion Peptide	N-Myristoylation
E 252-261	$_{252}$VLGSQEGAMH$_{261}$	Dimerization Domain	-

Distribution of pan-DENV sequences in nature

The 2 identified pan-DENV sequences overlapped at least 9 amino acid sequences of as many as 25 other viruses of the family *Flaviviridae*, genus *Flavivirus* including *Apoi, Banzi, Bouboui, Duck flavivirus, Edge Hill, Entebbe bat virus, Kedougou, Kokobera, Kunjin, Louping ill, Montana myotis leukoencephalitis, New Mapoon, Rio Bravo, Sepik, Spondweni, St. Louis encephalitis, Stratford Yellow fever, Tick-borne encephalitis, Uganda, West Nile, Wesselbron, Yokose, Zika virus*. Most of the sequences from these species were also belong to the E protein of *Flavivirus (*DENV also belongs to this genus*),* thus, associated dimerization/fusion domains. The number of the overlapped *E 97-111* sequences was high among known sequences of several of the highly studied Flaviviruses: *Kunjin, St. Louis encephalitis, Tick-borne encephalitis, West Nile, Yellow fever and Zika viruses* (Table 5).

Table 5. Distribution of pan-DENV sequences in nature. [a] Amino acid positions numbered according to the sequence alignments of the 4 DENV types; [b] Species (#) column indicates the number of viral species that shared at least 9 consecutive amino acids of the pan-DENV sequence; [c] Percentage representation is only shown for viral species with ≥ 10 total sequences reported at the NCBI Entrez protein database. These viral species included: LEV, *St. Louis encephalitis virus*; WNV, *West Nile virus*; TBEV, *Tick-born encephalitis virus*; *Yellow fever virus*; KJ, *Kunjin virus*; ZV, *Zika virus*.

Pan-DENV sequence [a]	Species (#) [b]	Percentage representation (%)/ number of sequence analyzed [c]					
		LEV	*WNV*	*TBEV*	*YFV*	*KJ*	*ZV*
E 97-111	22	100/27	100/149	87/229	93/261	100/36	_
E 252-261	3	_	89/56	_	_	_	100/53

Known and predicted HLA supertype-restricted, pan-DENV T-cell determinants

Literature survey and database search revealed that 1 of the pan-DENV sequences (*E 252-261*) overlapped at least 9 amino acids of 1 previously reported DENV T-cell determinants immunogenic in human, KKQDVVVLGSQEGAM, the amino acids present in *E 252-261* are underlined (Simmons *et al.,* 2005). Further evaluation of the immune-relevance of the pan-DENV sequences included a search for candidate putative promiscuous HLA supertype-restricted T-cell determinants within *E 97-111* and *E 252-261* by use of several computational algorithms described in the *Methods*. In general, both pan-DENV sequences, identified in E protein were predicted to contain 6 HLA class I supertype-restricted nonamers only, no class II-restricted peptides were identified (Table 6). Most of the predicted promiscuous HLA-binding nonamer were present in ≥95% of the sequences of each DENV type (Table 7). Clusters (hotspots) of two or more overlapping HLA-binder nonamer core peptide were present in both of the putative supertype pan-DENV sequences, mostly in *E 97-111*. These clusters contained 3 or more nonamer binders overlapping by 7 or 8 amino acids, covering most corresponding conserved region.

Table 6. Candidate putative HLA supertype-restricted binding nonamer peptides in pan-DENV sequences, predicted by immune-informatic algorithms. [a] Amino acid positions of the pan-DENV sequences and the predicted nonamers are numbered according to the sequence alignments of the 4 DENV types; [b] HLA supertype-restrictions that were predicted by at least two prediction models are highlighted in bold; [c] Peptides specific to HLA class I supertypes were predicted by use of NetCTL (A24, B7), ANN (A03, A01, A24) and SMM (A01, A03, A24, B27, B62); [d] Sequences identified as specific to class II were predicted by use of NN-align and SMM-align as described in methods.

Pan-DENV protein sequence and the predicted HLA supertype-restricted binding nonamer(s) [a]	**HLA supertype-restriction of predicted nonamer peptide** [b]				
	Class I [c]			**Class II** [d]	
	NetCTL	*ANN*	*SMM*	*NN-align*	*SMM-align*
$_{97}$VDRGWGNGCGLFGKG$_{111}$					
$_{99}$RGWGNGCGL$_{107}$	B7	_	B27	_	_
$_{99}$RGWGNGCGLF$_{108}$	_		A24	_	_
$_{99}$RGWGNGCGLFGK$_{110}$	_	A03	A01	_	_
$_{100}$GWGNGCGL$_{107}$	_	_	A24	_	_
$_{100}$GWGNGCGLF$_{108}$	A24	A24, A01	B62, A24	_	_
$_{100}$GWGNGCGLFGK$_{110}$	_	_	A03	_	_
$_{252}$VLGSQEGAMH$_{261}$					
$_{252}$VLGSQEGA$_{259}$	_	_	A24	_	_
$_{252}$VLGSQEGAM$_{260}$	_	_	B62		

Table 7. Intra-type representation of candidate putative HLA supertype-restricted nonamer peptides predicted by immune-informatics algorithms. [a] Amino acid positions of the pan-DENV sequences and the predicted nonamers are numbered according to the sequence alignments of the 4 DENV types; [b] Rounded to whole number.

Pan-DENV protein sequence and the predicted HLA supertype-restricted binding nonamer(s) [a]	**Intra-type representation (%)** [b] **and total sequences analyzed (#)**			
	DENV-1	*DENV-2*	*DENV-3*	*DENV-4*
$_{97}$VDRGWGNGCGLFGKG$_{111}$				
$_{99}$RGWGNGCGL$_{107}$	100/131	100/218	99.5/164	99.4/155
$_{99}$RGWGNGCGLF$_{108}$	100/131	100/218	99.5/164	99.4/155
$_{99}$RGWGNGCGLFGK$_{110}$	100/131	100/218	99.5/164	99.4/155
$_{100}$GWGNGCGL$_{107}$	100/131	100/218	99.5/164	99.4/155
$_{100}$GWGNGCGLF$_{108}$	100/131	100/218	99.5/164	99.4/155
$_{100}$GWGNGCGLFGK$_{110}$	100/131	100/218	99.5/164	99.4/155
$_{252}$VLGSQEGAMH$_{261}$				
$_{252}$VLGSQEGA$_{259}$	100/131	99.1/218	100/164	100/155
$_{252}$VLGSQEGAM$_{260}$	100/131	99.1/218	100/164	100/155

DISCUSSION

In this study, highly conserved pan-DENV sequences in all collected DENV database were identified and characterized. A large number of sequences analyzed (688), and their wide distribution in term of geography and time (1961 – 2012), provided decent information for a board survey of DENV protein diversity in nature. The 2 pan-DENV protein sequences of at least 9 amino acids, covering 25 amino acids or 5% of the complete DENV E protein, were conserved in ≥95% of the collected DENV sequences. Both of them are structural proteins and belong to the domain II of DENV

envelope, responsible for fusion/dimerization process (Allison *et al.,* 2001). These conserved sequences have shown remarkable stability over the entire database, as illustrated by their low entropy values and variant frequencies. Moreover, both of the pan-DENV sequences were conserved in 25 other *Flavivirus*, this provided more evidences of prolonged evolutionary stability within this genus (Henchal &Putnak 1990; Kuno *et al.,* 1998; Billoir *et al.,* 2000). It is possible that these pan-DENV sequences withstand selection pressure to maintain important biological and/or structural properties; in this case, it is fusion peptide and dimerization domain (Allison *et al.,* 2001; Modis *et al.,* 2004). Hence, these conserved sequences are unlikely to significantly diverge in newly emerging DENV isolates in the future, and represent attractive targets for the vaccine development.

Furthermore, there is also evidence that many of the conserved sequences are immunologically relevant. Putative T-cell epitopes for 6 HLA class I supertypes with board application to the immune responses of human population worldwide, were predicted by computational algorithms. Many the putative T-cell determinants were predicted to be promiscuous to multiple HLA supertypes, in addition to multiple alleles of a given HLA supertype, for instance, $_{100}$GWGNGCGLF$_{108}$; $_{99}$RGWGNGCGLFGK$_{110}$ and $_{99}$RGWGNGCGL$_{107}$ (Table 6).

The dramatic sequence variations between the proteins of the 4 DENV types represent a fundamental issue for the development of a polyvalent DENV vaccine that provides effective protection against each DENV type. Based on the MSA result, the sequences generally exhibit low intra-type but high inter-type variability. This observation is consistent with the other results concerning the viral diversity of envelope protein which can obstruct developing a polyvalent dengue virus vaccine (Livingston, *et al.,* 1995; Rothman, 2004) due to alternative determinants, or altered peptide ligands (Sloan-Lancaster and Allen, 1996). Furthermore, the phylogenetic tree revealed that DENV-4 is the most evolved and divergent among the 4 serotypes. In this particular study, this phenomenon caused many obstacles in the selection of conserved regions, significantly reducing many potential candidates. Some of them were already conserved in 3 DENV serotype but only different in 1 amino acid of DENV-4 sequence, thus failed to comply consensus-based approach (Novitsky *et al.,* 2002).

Ultimately, while the immune correlates of DENV protection are still poorly understood, some researchers suggested that both neutralizing antibody and specific T-cell responses are required (Rothman, 2004; Whitehead *et al.,* 2007). In order to improve vaccine efficiency, the introduction of well-defined HLA-restricted epitopes within DENV vaccine candidates could be a good strategy, thus, providing significant boost T-cell help positively to elicit a strong, long-lived immunity. For polyvalent formulations, it may be relevant to focus primarily on sequences that are conserved in all 4 DENV types and to avoid the regions of T-cell immunity that are highly variable, unless they are strictly type-specific, like in the case of DENV-4. Still, the two pan-DENV E sequences (E 97–111 and E 252– 261) are good candidate sequences for neutralizing antibody responses. However, in the future, population coverage calculation of these pan-DENV sequences are needed since an additional criterion for the selection of T-cell targets is the need for determinants with broad HLA representation, as it has been emphasized in the recognition of HLA supertypes (Sette & Sidney 1999; Sette *et al.,* 2001; Khan *et al.,* 2006). In addition, further wet lab investigations are needed to validate the immunogenicity of the candidate T-cell determinants in human subjects, and to identify sequences associated with deleterious T-cell responses.

CONCLUSION

This study aimed to identify and analyze conserved regions within the DENV E protein of 4 different serotypes to serve as potential epitope-based vaccine targets. 2 peptide regions of at least 9 amino acids were revealed to be highly conserved and identical in more than 95% of all collected DENV sequences. Both of them were found to be immune-relevant by their correspondence to known or putative HLA-restricted T cell determinants. Their level of conservation of these sequences was analyzed by many bioinformatics algorithms. The final results support their potential as candidates; however, wet-lab experiments are also necessary to validate the immunogenicity of the candidate T-cell epitopes. Still, this study is not complete yet. There are many things can be done to improve this approach such as 3-D structures of DENV E protein, identification of optimum binding HLA supertypes and population coverage. Furthermore, relationships with other species should also be considered in case these two regions show positive results in wet-lab

experiments. Another recommendation is that researches in the future should focus on DENV-4 and DENV-2. Since DENV-2 is the most evolved and divergent among 4 serotypes, and DEN-2 is the most popular and widespread. Finally, with some advance tweaks, this computational approach can also be applied to identify vaccine targets for other types of virus.

Acknowledgement: *The work was funded by the Department of Science and Technology at Ho Chi Minh City under Grant number 1160/QD-SKHCN. Computing resources provided by the Institute for Computational Science and Technology, Ho Chi Minh City is gracefully acknowledged.*

REFERENCES

Allison SL, Schalich J, Stiasny K, Mandl CW, Heinz FX (2001). Mutational evidence for an internal fusion peptide in flavivirus envelope protein E. *J Virol* 75: 4268-4275.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25: 3389-3402.

Armon A, Graur D, Ben-Tal N (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol bio* 307: 447-463.

Billoir F, de Chesse R, Tolou H, de Micco P, Gould EA, de Lamballerie X (2000). Phylogeny of the genus flavivirus using complete coding sequences of arthropod-borne viruses and viruses with no known vector. *J Gen Virol* 81: 781-790.

Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL (2012). Domain enhanced lookup time accelerated BLAST. *Biology direct* 7: 1.

Bourdette D, Edmonds E, Smith C, Bowen J, Guttmann CR, Nagy Z, Simon J, Whitham R, Lovera J, Yadav V (2005). A highly immunogenic trivalent T cell receptor peptide vaccine for multiple sclerosis. Multiple Sclerosis 11: 552-561.

Consortium U (2014). UniProt: a hub for protein information. Nucleic acids research gku989

Costa RL, Voloch CM, Schrago CG (2012). Comparative evolutionary epidemiology of dengue virus serotypes. Infection, Genetics and Evolution 12: 309-314.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 32: 1792-1797.

Gattiker A, Gasteiger E, Bairoch AM (2002). ScanProsite: a reference implementation of a PROSITE scanning tool. *Applied Bioinformatics* 1: 107-108.

Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 19: 163-164.

Gubler DJ, Ooi EE, Vasudevan S, Farrar J (2014) Dengue and dengue hemorrhagic fever. CABI.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In: Nucleic acids symposium series, pp: 95-98.

Haydon DT, Woolhouse ME (1998). Immune avoidance strategies in RNA viruses: fitness continuums arising from trade-offs between immunogenicity and antigenic variability. *J Theor Bio* 193: 601-612.

Henchal EA, Putnak JR (1990). The dengue viruses. *Clinical Microbiology Reviews* 3: 376-396.

Holmes EC, Burch SS (2000). The causes and consequences of genetic variation in dengue virus. Trends in microbiology 8: 74-77.

Hotez PJ, Alvarado M, Basáñez M-G, Bolliger I, Bourne R, Boussinesq M, Brooker SJ, Brown AS, Buckle G, Budke CM (2014). The global burden of disease study 2010: interpretation and implications for the neglected tropical diseases. PLoS Negl Trop Dis 8: e2865.

Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ (2006). The PROSITE database. Nucleic acids research 34: D227-D230.

Katoh K, Misawa K, Kuma Ki, Miyata T (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research 30: 3059-3066.

Khan AM, Miotto O, Heiny A, Salmon J, Srinivasan K, Nascimento EJ, Marques ET, Brusic V, Tan TW, August JT (2006). A systematic bioinformatics approach for selection of epitope-based vaccine targets. Cellular immunology 244: 141-147.

Knutson KL, Schiffman K, Disis ML (2001). Immunization with a HER-2/neu helper peptide vaccine generates HER-2/neu CD8 T-cell immunity in cancer patients. *J Clin Invest* 107: 477-484.

Korber B, LaBute M, Yusim K (2006). Immunoinformatics comes of age. *PLoS Comput Biol* 2: e71.

Kuhn RJ, Zhang W, Rossmann MG, Pletnev SV, Corver J, Lenches E, Jones CT, Mukhopadhyay S, Chipman PR, Strauss EG (2002). Structure of dengue virus: implications

for flavivirus organization, maturation, and fusion. Cell 108: 717-725.

Kuno G, Chang G-JJ, Tsuchiya KR, Karabatsos N, Cropp CB (1998). Phylogeny of the genus Flavivirus. *J Virol* 72: 73-83.

Larsen MV, Lundegaard C, Lamberth K, Buus S, Brunak S, Lund O, Nielsen M (2005). An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur J Immunol* 35: 2295-2303.

Lassmann T, Sonnhammer EL (2005). Kalign–an accurate and fast multiple sequence alignment algorithm. BMC bioinformatics 6: 1.

Livingston PG, Kurane I, Dai L-C, Okamoto Y, Lai C-J, Men R, Karaki S, Takiguchi M, Ennis FA (1995). Dengue virus-specific, HLA-B35-restricted, human CD8+ cytotoxic T lymphocyte (CTL) clones. Recognition of NS3 amino acids 500 to 508 by CTL clones of two different serotype specificities. *J Immunol* 154: 1287-1295.

Lockhart PJ, Steel MA, Hendy MD, Penny D (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol* 11: 605-612.

López JA, Weilenman C, Audran R, Roggero MA, Bonelo A, Tiercy JM, Spertini F, Corradin G (2001). A synthetic malaria vaccine elicits a potent CD8+ and CD4+ T lymphocyte immune response in humans. Implications for vaccination strategies. *Eur J Immunol* 31: 1989-1998.

Modis Y, Ogata S, Clements D, Harrison SC (2004). Structure of the dengue virus envelope protein after membrane fusion. Nature 427: 313-319.

Murphy BR, Whitehead SS (2011). Immune Response to Dengue Virus and Prospects for a Vaccine*. *Annu Rev Immunol* 29: 587-619.

Murrell S, Wu S-C, Butler M (2011). Review of dengue virus and the development of a vaccine. *Biotechnol Adv* 29: 239-247.

Mustafa M, Rasotgi V, Jain S, Gupta V (2015). Discovery of fifth serotype of dengue virus (DENV-5): A new public health dilemma in dengue control. *Med J Armed Forces India* 71: 67-70.

Nielsen M, Lund O (2009). NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics* 10: 1.

Nielsen M, Lundegaard C, Lund O (2007). Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics* 8: 1.

Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, Brunak S, Lund O (2003). Reliable prediction of T cell epitopes using neural networks with novel sequence representations. Protein Science 12: 1007-1017.

Novitsky V, Smith U, Gilbert P, McLane M, Chigwedere P, Williamson C, Ndung'u T, Klein I, Chang S, Peter T (2002). Human immunodeficiency virus type 1 subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design? *J Virol* 76: 5435-5451.

Okonechnikov K, Golosova O, Fursov M (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28: 1166-1167.

Perera R, Kuhn RJ (2008). Structural proteomics of dengue virus. Current opinion in microbiology 11: 369-377.

Peters B, Sette A (2005). Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. BMC Bioinformatics 6: 1.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *J Computat Chem* 25: 1605-1612.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J (2011). The Pfam protein families database. Nucleic acids research gkr 1065.

Rey FA (2003). Dengue virus envelope glycoprotein structure: new insight into its interactions during viral entry. Proceedings of the National Academy of Sciences 100: 6899-6901.

Rigau-Pérez JG, Clark GG, Gubler DJ, Reiter P, Sanders EJ, Vorndam AV (1998). Dengue and dengue haemorrhagic fever. *Lancet* 352: 971-977.

Rothman AL (2004). Dengue: defining protective versus pathologic immunity. *J Clin Invest* 113: 946-951.

Sette A, Livingston B, McKinney D, Appella E, Fikes J, Sidney J, Newman M, Chesnut R (2001). The development of multi-epitope vaccines: epitope identification, vaccine design and clinical evaluation. Biologicals 29: 271-276.

Sette A, Sidney J (1999). Nine major HLA class I supertypes account for the vast preponderance of HLA-A and-B polymorphism. *Immunogenetics* 50: 201-212.

Simmons CP, Dong T, Chau NV, Dung NTP, Chau TNB, Dung NT, Hien TT, Rowland-Jones S, Farrar J (2005). Early T-cell responses to dengue virus epitopes in Vietnamese adults with secondary dengue virus infections. *J Virol* 79: 5665-5675.

Sloan-Lancaster J, Allen PM (1996). Altered peptide ligand-

induced partial T cell activation: molecular mechanisms and role in T cell biology. *Annu Rev Immunol* 14: 1-27.

Strait BJ, Dewey TG (1996). The Shannon information entropy of protein sequences. *Biophys J* 71: 148.

Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, Ritter O, Abola E (1998). Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallographica Section D Biological Crystallography 54:1078-1084.

Thompson JD, Gibson T, Higgins DG (2002). Multiple sequence alignment using ClustalW and ClustalX. Current protocols in bioinformatics 2.3. 1-2.3. 22.

Twiddy SS, Holmes EC, Rambaut A (2003). Inferring the rate and time-scale of dengue virus evolution. *Mol Biol Evol* 20: 122-129.

Valdar WS (2002). Scoring residue conservation. Proteins: Structure, Function, and Bioinformatics 48: 227-241.

Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B (2010). The immune epitope database 2.0. Nucleic acids research 38: D854-D862.

Whitehead SS, Blaney JE, Durbin AP, Murphy BR (2007). Prospects for a dengue virus vaccine. *Nat Rev Microbiol* 5: 518-528.