Virus Genes (2007) 35:175–186
DOI 10.1007/s11262-007-0103-2

Computational analysis and identification of amino acid sites in dengue E proteins relevant to development of diagnostics and vaccines

Raja Mazumder · Zhang-Zhi Hu · C. R. Vinayaka · Jose-Luis Sagripanti · Simon D. W. Frost · Sergei L. Kosakovsky Pond · Cathy H. Wu

Received: 25 January 2007 / Accepted: 11 April 2007 / Published online: 17 May 2007

Abstract We have identified 72 completely conserved amino acid residues in the E protein of major groups of the Flavivirus genus by computational analyses. In the dengue species we have identified 12 highly conserved sequence regions, 186 negatively selected sites, and many dengue serotype-specific negatively selected sites. The flavivirus-conserved sites included residues involved in forming six disulfide bonds crucial for the structural integrity of the protein, the fusion motif involved in viral infectivity, and the interface residues of the oligomers. The structural analysis of the E protein showed 19 surface-exposed non-conserved residues, 128 dimer or trimer interface residues, and regions, which undergo major conformational change during trimerization. Eleven consensus T_h-cell epitopes common to all four dengue serotypes were predicted. Most of these corresponded to dengue-conserved regions or negatively selected sites. Of special interest are six singular sites (N_{37}, Q_{211}, D_{215}, P_{217}, H_{244}, K_{246}) in dengue E protein that are conserved, are part of the predicted consensus T_h-cell epitopes and are exposed in the dimer or trimer. We propose these sites and corresponding epitopic regions as potential candidates for prioritization by experimental biologists for development of diagnostics and vaccines that may be difficult to circumvent by natural or man-made alteration of dengue virus.

Keywords Dengue · Flavivirus · Hemorrhagic virus · Selection pressure · T-cell epitope · Envelope protein

Introduction

Positive-sense single-stranded RNA (+ssRNA) viruses include highly virulent human pathogens. Within the Flaviviridae family, the genus *Flavivirus* comprises more than 70 +ssRNA viruses [1], including dengue virus (DENV). Dengue, an acute viral disease transmitted by mosquito, is one of the most widespread vector-borne viral diseases in humans. Dengue is caused by any of four antigenically distinct serotypes: dengue-1 (DENV1), dengue-2 (DENV2), dengue-3 (DENV3), and dengue-4 (DENV4). There are estimated 50–100 million cases of dengue fever annually worldwide, half a million of which result in severe forms of the disease, dengue hemorrhagic fever and dengue shock syndrome [2]. Generally, infection with one serotype confers future protective immunity against that particular serotype, but not against the others. In fact, dengue hemorrhagic fever may occur from sequential infection by different virus serotypes in a process called antibody-mediated disease enhancement, where antibodies raised against the first serotype enhance infection with the second serotype [3].

Electronic supplementary material The online version of this article (doi:10.1007/s11262-007-0103-2) contains supplementary material, which is available to authorized users.

R. Mazumder · Z.-Z. Hu · C. R. Vinayaka · C. H. Wu
Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, DC 20007, USA

J.-L. Sagripanti (✉)
Edgewood Chemical Biological Center, US Army, AMSRD-ECB-RT, Aberdeen Proving Ground, Aberdeen, MD 21010, USA
e-mail: joseluis.sagripanti@us.army.mil

S. D. W. Frost · S. L. Kosakovsky Pond
Department of Pathology, University of California at San Diego, La Jolla, CA 92093, USA

The major envelope glycoprotein (referred to as E protein hereafter) of DENV and of other flaviviruses is responsible for important phenotypic and immunogenic properties of the virion and is believed to lead the virus entry into cells [4, 5]. The E protein mediates virus assembly and virus–cell membrane fusion, and initiates infection through binding to cell surfaces. This protein is the principal component of the external surface of the DENV virion and represents the dominant virus antigen, evoking protective immune responses. Dengue serotypes can be distinguished by virus-neutralizing antibodies, but non-neutralizing antibodies against the E protein are cross-reactive. These non-neutralizing antibodies may help bring the virion into close proximity to the normal virus receptor, thus enhancing virus binding and increasing the number of infected cells, with concomitant exacerbation of the disease [6].

While it is well established that the E protein is one of the major proteins responsible for the pathogenicity and immunogenic properties of flaviviruses, the exact residues/regions responsible for these traits remain to be identified. Single-residue substitutions mapped to different parts of the E protein were reported to cause flavivirus attenuation [7], implying that several residues within the E protein are responsible for phenotypic and pathogenic properties.

The crystal structure of the soluble ectodomain of DENV2 E protein reveals a hydrophobic pocket lined by residues that influence the pH threshold for membrane fusion [8–10]. The protein has three structural domains (DI, DII, DIII) that map closely to the three antigenic regions (C, A, and B, respectively) [4]. DENV enters the host cell when the E protein binds to a yet undefined cell receptor and responds to a reduced pH of the endosome by a conformational change [11]. This conformational change induces fusion of the virus and the host cell membrane. The crystal structure of DENV3 E protein indicates that the serotype-specific mutations that allow viral evasion from immune surveillance (neutralization escape mutations) are all located on the surface of domain III, which has been implicated in receptor binding [12, 13]. The apparent involvement of the host immune system in disease pathogenesis, the so-called antibody-dependent enhancement (ADE), has hampered development of a vaccine against dengue.

The goal of this study is to identify specific regions of the E protein with high potential success as targets for future development of robust diagnostics and vaccines against dengue virus. To achieve this goal we have employed several complementary bioinformatics methods to analyze sequence conservation, selection pressure, immunogenic properties and structural features of the DENV E proteins.

Materials and methods

Residue and site numbers mentioned in the manuscript are based on the DENV2 E protein sequence unless otherwise noted.

Analysis of sequence conservation in the flavivirus genus and dengue species

Sequence conservation of E proteins was analyzed based on structure-guided multiple sequence alignment, neighbor-joining phylogenetic tree, and Shannon entropy. Twelve representative sequences were chosen from the major flavivirus groups where the genome polyproteins are available (Table 1). The taxonomic groups (http://www.ncbi.nlm.nih.gov/ICTVdb/Ictv/fs_flavi.htm) were defined based on the ICTVdB nomenclature of the International Committee on Taxonomy of Viruses [1]. The E protein sequences were extracted from polyproteins in the UniProt Knowledgebase (UniProtKB) [14]. The multiple sequence alignment was manually edited using the Cn3D sequence-structure viewer [15], guided by a reference structure alignment in MMDB [6] using three E protein structures (1OKE, 1UZG, 1SVB), which were originally deposited in PDB [16]. The manual editing involves superimposing and aligning the structures in the structure viewer and then adding individual sequences manually and aligning them with the sequences from the structure. For the C-terminal regions outside the structural alignment (~100 aa), the sequences were aligned using ClustalW [17] and then manually verified. A phylogenetic tree was generated using the neighbor-joining program in ClustalW with 1,000 bootstrap replicates. The bootstrap values indicate the confidence in the estimated tree branches.

The sequence conservation of dengue E proteins was further analyzed based on the multiple sequence alignment of 740 full-length proteins from all four serotypes. The conservation in each amino acid position was quantified based on the Shannon entropy, as estimated by the nine-component Dirichlet mixture algorithm [18]. The entropy calculation took into account conservative substitutions of amino acids with similar physicochemical properties.

Evolutionary selection analysis of dengue serotypes

Selection analysis was studied using maximum likelihood methods to calculate non-synonymous/synonymous (dN/dS) nucleotide substitution ratio based on codon alignment and phylogenetic tree. The maximum likelihood methods evaluate the probability that the chosen model has produced the observed data. DNA sequences were extracted

Table 1 E proteins and their source viruses and taxonomic groups in the genus *Flavivirus*

Virus group and name	Virus abbreviation	UniProtKB ID: residue range	PDB ID
1. Tick-borne viruses			
Mammalian tick-borne virus group			
Omsk hemorrhagic fever virus	OHFV	Q7T6D2_9FLAV:281–776	
Tick-borne encephalitis virus	TBEV	POLG_TBEVW:281–776	1SVB
Louping ill virus	LIV	POLG_LIV:281–776	
2. Mosquito-borne viruses			
Dengue virus group			
Dengue virus serotype 1	DENV1	POLG_DEN1S:281–774	
Dengue virus serotype 2	DENV2	POLG_DEN2P:281–775	1OAN, 1OK8, 1OKE
Dengue virus serotype 3	DENV3	POLG_DEN3:281–773	1UZG
Dengue virus serotype 4	DENV4	POLG_DEN4:280–774	
Japanese encephalitis virus group			
St. Louis encephalitis virus	SLEV	POLG_STEVM:276–814	
West Nile virus	WNV	POLG_WNV:291–787	
Yellow fever virus group			
Yellow fever virus	YFV	POLG_YEFV1:286–778	
3. Viruses with no known arthropod vector			
Modoc virus group			
Apoi virus	APOIV	Q9J9C2_9FLAV:272–756	
Rio Bravo virus group			
Rio Bravo virus	RBV	Q9JAD5_9FLAV:268–751	

from the NCBI nucleotide nt database [19], resulting in four datasets consisting of 146, 269, 121, and 204 sequences for DENV1, DENV2, DENV3, and DENV4, respectively, as well as a fifth dataset totaling 740 sequences with all four serotypes combined. The codon alignment and consensus neighbor-joining tree was derived using MEGA [20]. To quantify selective pressure at a given codon site, estimates of synonymous (dS) and non-synonymous (dN) substitution rates were compared using a statistical test as described earlier [21]. If $dS > dN$ (or, $dS < dN$), then a site was inferred to be negatively (or positively) selected. We employed two likelihood-based methods—Single Likelihood Ancestor Counting (SLAC) and Fixed Effects Likelihood (FEL) [22] and considered sites to be ''selected'' when both SLAC and FEL methods yielded a *P*-value of <0.01. SLAC reconstructs the most likely unobserved ancestral sequences, counts the number of non-synonymous and synonymous changes at every site, and tests whether the number of non-synonymous changes per non-synonymous site is significantly different from the number of synonymous changes per synonymous site. FEL derives the branch lengths and substitution rate bias (global) parameters from the entire alignment, and then directly estimates the ratio of non-synonymous to synonymous rates under a codon-substitution model for each site in a sequence alignment holding all global parameters fixed. The FEL method is in general more powerful but also more computationally demanding than the SLAC method [22]. For this study we reported only those sites, which were concordantly classified with both methods. For the large combined set with all four serotypes, a more stringent cutoff of $P < 10^{-5}$ was used [22]. Selection analyses were performed using the HyPhy program [23], which took into account nucleotide substitution biases and dS and dN rate variation across sites. To exclude the possibility of erroneous selection inference due to recombination between different DENV serotypes, a maximum likelihood test for phylogenetic incongruence [24] was run to screen for possible recombination in the E protein; no statistically significant recombination breakpoints were identified in any of the DENV datasets.

Prediction of T-Cell epitopes

Three online prediction programs, NetMHC 2.1 (http://www.cbs.dtu.dk/services/NetMHC/) [25], MHCPred 2.0

(http://www.jenner.ac.uk/MHCPred/) [26] and RANKPEP (http://www.mifoundation.org/Tools/rankpep.html) [27], were used to analyze the four serotypes of dengue E proteins. As both MHC-I and -II molecules are highly polymorphic and the specificity of the alleles is often very different, predictions were performed on multiple supertypes to cover the polymorphic loci. NetMHC was used for the prediction for 12 supertypes of MHC-I locus. A binding affinity threshold of 500 nM was used as the cutoff for MHC-I binding [26]. Both MHCPred 2.0 and RANKPEP were used for MHC-II binding predictions to obtain consensus MHC-II epitopes. For MHCPred prediction, 3 supertypes of MHC-II locus were used, with a final binding affinity cutoff of 60 nM for MHC-II binding to obtain approximately the top 5% of the binders that correspond to regions of synthetic peptides reported to induce T-cell immune responses in mice [28]. For RANKPEP prediction, 50 MHC-II locus types were used and the cutoff was set at 4% of top-scoring peptides with above default threshold scores as reported in Ref. [27]. Data from individual supertypes of MHC-I or -II loci were combined for the analysis of each peptide. To ensure better predictive accuracy, only consensus results above the cutoff of both MHCPred and RANKPEP were considered as MHC-II binders.

Analysis of protein structural features

The structural features of dengue E proteins were analyzed using known structures in PDB for DENV2, 1OAN (dimer), 1OK8 (post-fusion trimer), and 1OKE (protein in complex with *n*-octyl-β-D-glucoside) [8, 29], and the DENV3 structure, 1UZG (dimer) [12]. The extent of exposure of amino acid residues was determined by computing the relative accessible surface area (ASA) using the POLYVIEW server (http://polyview.cchmc.org/polyview_doc.html) [30]. Relative ASA of a residue is the ratio of the ASA of that residue in the protein to ASA of the same residue in the fully extended tripeptide alanine-residue-alanine. Based on the value of the relative ASA, the residues were grouped as buried (0.0–0.60) or exposed (0.61–1.0). To determine interactions across the dimer and trimer interface, the occluded surface (OS) area was computed using the method of Pattabiraman et al. [31]. Residues with OS area >0.5 $Å^2$ were considered as interacting. The conformational rearrangements occurring during the dimer to trimer transition were measured by the changes in the backbone torsion angles φ and ψ between the DENV2 E protein dimer (1OAN) and trimer (1OK8). The conformational angles were obtained using the DSSP program [32], and the difference in the backbone torsion angles, $\Delta\varphi$ and $\Delta\psi$, for each amino acid residue was calculated.

Results

Sequence conservation of E proteins in the *Flavivirus* genus

Sequence conservation of E proteins among dengue species and other members of the genus *Flavivirus* was studied based on structure-guided multiple sequence alignment using 12 representative sequences chosen from the major flavivirus groups where the genome polyproteins are available (Table 1). The multiple sequence alignment and the corresponding phylogenetic tree are shown in Fig. 1. The neighbor-joining tree reflects the relationships of the taxonomic groups, where the three major groups (1—Tick-borne viruses, 2—Mosquito-borne viruses, and 3—Viruses with no known arthropod vector) and their subgroups are clearly delineated.

The multiple sequence alignment shows 72 completely conserved amino acid residues in the 12 flaviviruses (Fig. 1). These residues include cysteines forming the six pairs of disulfide-bonds [33] that stabilize loop structures in the three structural domains of E protein. Other crucial structure-stabilizing interactions by the flavivirus-conserved amino acids include the D_{98}-K_{110} salt bridge, several intramolecular hydrogen bonds involving D_{10}, C_{30}, G_{100}, W_{101}, C_{105}, L_{216}, and L_{218}, and the hydrophobic environment provided by leucines (L_{216}, L_{218}, and L_{264}) and V_{208}. Also, completely conserved among flaviviruses are two pairs of residues involved in interactions of domains I and III, R_9 and E_{368}, which form a salt bridge, and H_{144} and H_{317} that are involved in hydrogen bonds with the main chain of the opposite domain across the interface [34].

The most highly conserved region in flavivirus E protein is $_{98}$DRGWGNGCGLFGKG$_{111}$, where 13 of the 14 residues are strictly conserved (Fig. 1). This peptide, contained in an internal loop between two β-strands on domain DII, corresponds to the known fusion motif [8] involved in DENV infectivity. The highly variable region $_{380}$IGVEPGQLKL$_{389}$, in the lateral loop on domain III, has been implicated in receptor-binding of DENV2 [35] and tick-borne viruses [34].

Sequence conservation of E proteins in the dengue species

The alignment of representative E protein sequences of the four dengue serotypes (Fig. 2A) reveals a very high degree of sequence conservation as expected due to their close evolutionary relationship. A total of 260 residues (~53% of all residues) are conserved in the multiple alignment of the four sequences. The conservation spans across the entire sequence length, including complete sequence identity of

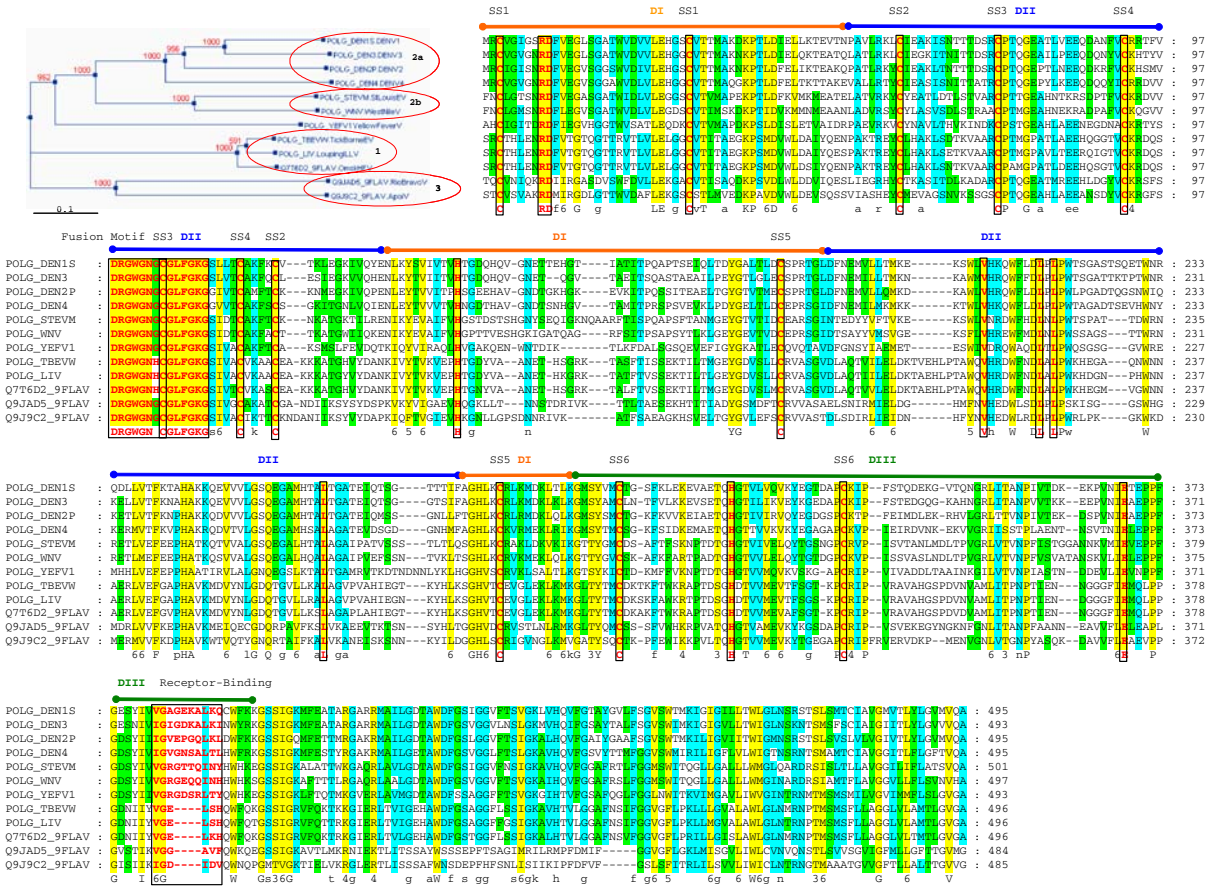

Fig. 1 Multiple sequence alignment and phylogenetic tree of 12 E proteins in the *Flavivirus* genus. The neighbor-joining phylogenetic tree is shown with a scale bar representing branch lengths, numbers above nodes representing bootstrap support (out of 1,000), and the major subgroups are circled. The consensus sequence shown below the alignment indicates residues that are completely conserved (upper-case), highly conserved (lower-case), and in conserved amino acid groups (2: polar, 3: alcohol, 4: charged, 5: aromatic, 6: hydrophobic). Additionally, the background colors for 100, 80, and 60% conserved residues are yellow, blue, and green, respectively. The labels above the alignment indicate the three structure domains (DI, DII, DIII) based on the DENV2 E protein structure (PDB code: 1OAN), the six disulfide bonds (SS1–SS6), the fusion motif, and the receptor-binding motif. The conserved residues involved in critical interactions are boldfaced in red and boxed

the fusion motif in D_{98}-G_{111} and the two known N-linked glycosylation sites at N_{67} and N_{153} [9] with the Asn-X-Thr/Ser-X potential glycosylation site motif (X can be any residue except for proline).

Figure 2B shows the Shannon entropy that quantifies the conservation of each amino acid position in the multiple sequence alignment of 740 E proteins from all four serotypes. The entropy ranges from 0.365 to 2.42 bits, with scores <0.6 for highly conserved residues, 0.6–0.9 for conservative amino acid substitutions, and >0.9 for non-conserved residues. The mean entropy of the full-length protein is 0.539 bits, and 55 and 91% of the positions have entropies of <0.6 and <0.9, respectively.

Comparisons of the entropy measure of the three structural domains DI (amino acids 1–52, 134–191, 280–295), DII (aa 53–133, 192–279) and DIII (aa 296–394), and the C-terminal region (aa 394–495) reveal regions of different sequence variability. In particular, domain DIII has the highest variability, with an entropy mean of 0.703 bits and entropies of <0.6 and <0.9 for 44 and 86% of the positions. Interestingly, serotype-specific neutralization escape mutant sites in DENV E proteins are all located on the surface of domain III [12]. Twelve sequence regions are highly conserved in DENV E proteins, containing five or more consecutive amino acid sites with entropy scores of <0.6, namely, N_8-G_{14}, V_{24}-D_{42}, R_{73}-E_{79}, V_{97}-S_{102}, D_{192}-M_{196}, V_{208}-W_{220}, V_{252}-H_{261}, G_{281}-C_{285}, E_{314}-T_{319}, E_{370}-G_{374}, K_{394}-G_{399}, and R_{411}-S_{424}, designated as sequence regions I to XII, respectively (Fig. 2B).

Amino acid sites under selection pressure in dengue E proteins

Estimates of synonymous or silent (dS) and non-synonymous or amino-acid altering (dN) nucleotide substitution rates at a given position in a codon alignment have become

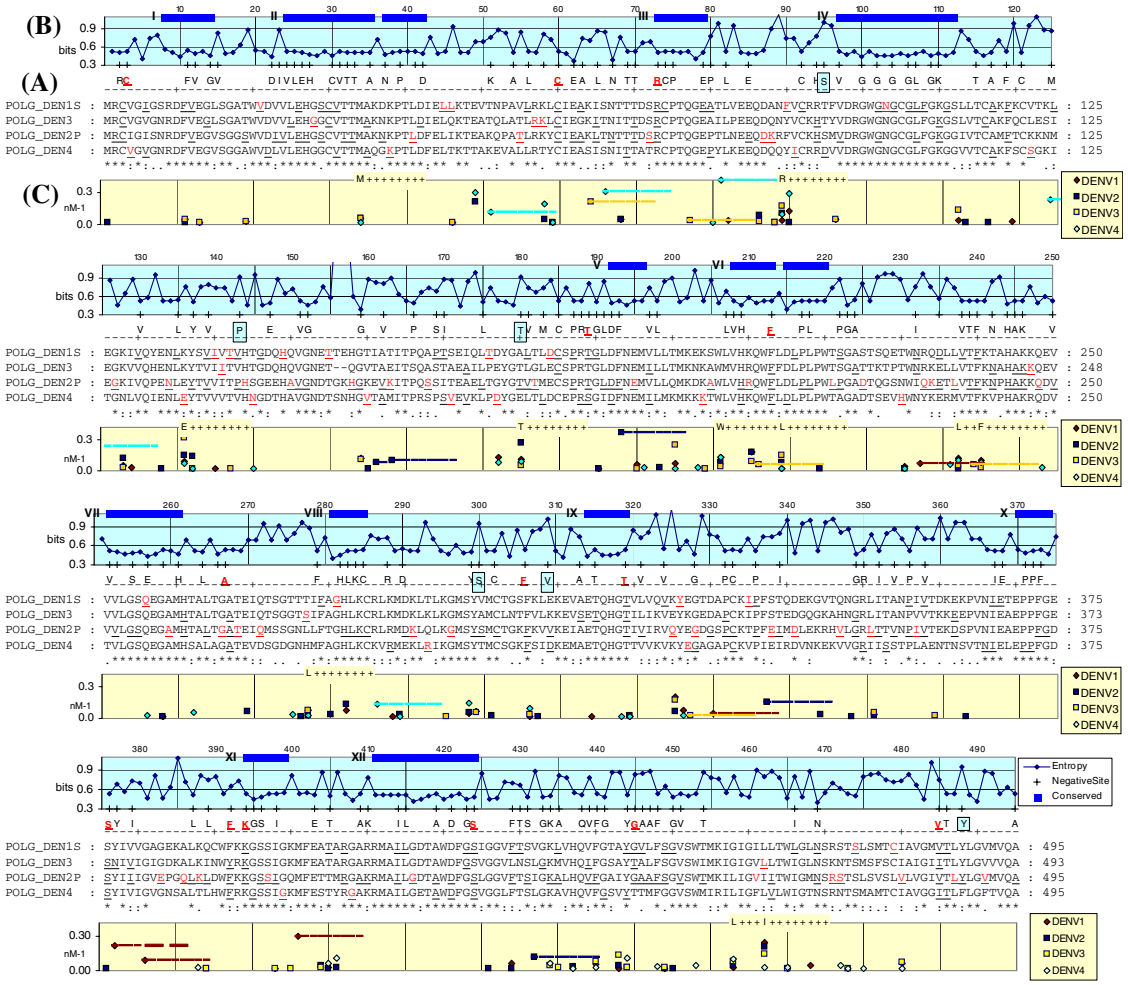

Fig. 2 (A) Multiple sequence alignment and negatively selected sites of E proteins in four dengue serotypes. The underlined residues on the alignment are negatively selected sites in each serotype and sites under negative selection pressure only in a specific serotype and not in the merged dataset are in red. The residues labeled above the alignment (based on DENV2) indicate negatively selected sites in the merged dataset with all four serotypes; underlined residues are sites that are also under negative selection pressure in at least three serotypes; and boxed residues are sites with entropy score of >0.9 bits. (B) The Shannon entropy quantifying sequence conservation of 740 E proteins from the four dengue serotypes. The entropy score (*Y*-axis) shown in the line graph ranges from 0.3 to 1.1 bits; the blue lines labeled from I to XII indicate conserved regions. (C) Predicted MHC class II binding peptides for each dengue serotype E protein. The symbols in the scatter plot mark the beginning (amino-terminal) of each 9-mer peptide. The *Y*-axis is the inverse score of the binding affinity (nM^{-1}) predicted by MHCPred. The residue labels (followed by ''+'' marks) on the top line of the graph indicate binding peptides within the region common to all four E proteins. The top 5 high-affinity MHC-II binding peptides specific to individual serotypes are indicated by dashed lines

a standard measure of selective pressure, especially in the framework of maximum likelihood phylogenetic methods [22]. If dS is estimated to be significantly more than dN at a given site, (a dN/dS value of <1), this can be taken as evidence of purifying (negative) selective pressure on that site. That is, for that particular site amino acid changes are, on average, deleterious (negative selection). When the opposite is true, i.e. dN > dS, (dN/dS > 1) there is selective pressure to generate and possibly maintain amino-acid polymorphisms, i.e. undergo adaptive change (positive selection). This unusual condition may reflect a change in the function of a gene or an immediate change in environmental conditions (such as a pathogen's response to an administered drug) that forces the organism to adapt.

To quantify selective pressure, dS and dN were estimated for each site of the alignment of individual dengue serotypes and also for all serotypes combined. The Tamura Nei (TN93) [36] model of nucleotide substitution bias (out of 203 possible models) was selected for all alignments. To each alignment we fitted one of four models of site-to-site rate variation (Constant: dS = dN = 1; Proportional: dN is proportional to dS, which varies among sites; Non-synonymous:

dS = 1, dN varies among sites; and Dual: dS and dN vary among sites independently). Strong evidence supporting the Dual model of rate variation was observed in each alignment (shown in supplementary Table S1). This finding suggests that both dS and dN vary across sites, but there is no simple correlation pattern between the two rates [37].

Ninety negatively selected sites were identified in DENV1, 161 in DENV2, 49 in DENV3, and 57 in DENV4, as shown in Fig. 2A (underlined residues on the alignment), while 186 sites were found to be under negative selection pressure in the large combined set with all four serotypes (labeled residues above the alignment). Because the signature of negative selection is the relative abundance of synonymous substitutions likely due to functional constraints, most negatively selected sites corresponded to conserved residues or substitution of an amino acid by another with similar chemical properties (conservative substitutions). Altogether 138 out of 186 (74%) negative sites in the merged set have low entropy scores of <0.6 bits. Furthermore, 14 sites (C_3, C_{60}, R_{73}, T_{189}, F_{213}, A_{267}, F_{306}, T_{319}, S_{376}, F_{392}, K_{394}, S_{424}, G_{445}, and V_{485}) are negatively selected in at least three of the four serotypes as well as in the merged dataset (underlined residues); most are highly conserved and a few have conservative substitutions. Only 6 out of 186 (3%) negative sites in the merged set have >0.9 bits entropy scores (S_{95}, P_{143}, T_{180}, S_{300}, V_{309}, and Y_{488}, boxed residues). Note that while these sites are not conserved within the dengue species, they all correspond to negatively selected sites in one or two specific serotypes, possibly reflective of selective sweeps that have become fixed in individual serotypes and are now maintained by purifying selection (negative selection). Finally, several serotype-specific negatively selected sites were also identified (19 in DENV1, 48 in DENV2, 7 in DENV3, 16 in DENV4) (Fig. 2). Most notable are the 5 (E_{383} and Q_{386}-L_{389}) DENV2-specific negative sites in the receptor-binding region.

Although there are several reports on the role of positively selected sites in pathogen–host interaction, such as evading host immunity [38–42], there are few reports of experimental validation of the functional significance of negatively selected sites. It has been shown that epitopes consisting of negatively selected sites perform better as vaccines than ones containing positively selected sites [43, 44]. The underlying assumption is that because negatively selected sites are less likely to change, due to functional constraints, vaccines or diagnostic targets directed against them may be more effective.

No positively selected sites were detected in any of the dengue serotypes in this study. Our findings agree with previous selection studies where constant dS across all sites was assumed a priori and the data were not stratified based on genotypes and passage types [45, 46]. Comparable analysis of E gene sequences from other flaviviruses, such as St. Louis encephalitis virus, West Nile virus and Yellow fever virus, also did not detect any positive selection [47]. In order to screen for possible selection on amino-acid residues prior to the divergence of serotypes, we estimated dN and dS along the four tree branches separating individual serotype clades in the joint phylogeny using a fixed effects likelihood method [22]. Five sites with evidence of ancient positive selection were suggested ($P \leq 0.001$): N_{83}, P_{132}, E_{174}, Q_{293}, and L_{458}.

MHC peptide binding in dengue E proteins

MHC-I peptide binding prediction using NetMHC [25] identified several potential binding regions for each of the four E proteins. However, the overall MHC-I binding affinity was low and the number of high binders was small. Low predicted MHC-I binding was further confirmed using MHCPred (data not shown). These results are consistent with the observation that E proteins mainly induce antibody response to DENV, while non-structural protein 3 (NS3) mainly induces T-cell immune responses to DENV [48].

Many MHC-II binding peptides (T_h-cell epitopes) were predicted by both MHCPred and RANKPEP [27] above the affinity threshold (56, 64, 50, and 53 peptides for DENV1, DENV2, DENV3, and DENV4, respectively, had affinities ranging from 1 to 60 nM) (Fig. 2C). Of these, 11 peptides are in regions common to all four serotypes (labeled by beginning residues in Fig. 2C) and represent immunogenic consensus sequence epitopes in the DENV species. The figure also indicates high-affinity MHC-II binding peptides among the top-ranking predictions (dashed lines) that are unique to one of the four serotypes. The predicted binding peptides common to all serotypes generally are present in more conserved regions with low-entropy and/or negatively selected sites, except the last two consensus binding peptides occurring in the C-terminal transmembrane region. On the other hand, most predicted serotype-specific binding peptides are in variable regions with lower sequence conservation. Interestingly, the most highly variable domain DIII contains only predicted serotype-specific binding peptides, but no consensus binding peptides common to all four serotypes.

To cross-validate the predicted results, we further mapped the 64 predicted MHC-II binders of DENV2 E protein to regions covered by synthetic peptides that were previously determined to mimic T_h-cell epitopes and to elicit antibody responses in three different mouse strains [28] (amino acid regions of peptides in Supplementary Table S2). As 15 of the predicted peptides are not completely covered within regions of the synthetic peptides, the comparison was based on the remaining 49 predicted

peptides. We noted that 39 of the 49 (80% true positive) predicted MHC-II peptides were matched with 16 synthetic peptides that experimentally tested positive for T_h-cell epitopes; while the remaining 10 binders (20% false positive) correlated to synthetic peptides that did not elicit an immune response either in vitro or in vivo [28]. Conversely, the computational methods predicted all but one of the 17 synthetic peptides shown to induce an immune response (Table S2), yielding a 94% (16/17) recall rate. The predictive accuracy observed here is consistent with the benchmarking results of epitope prediction programs [49].

Consistent with the notion that the variable surface residues are likely to be responsible for the serotype-specific immunogenic variation, we identified four specific sequence regions ($_{329}$DGS$_{331}$, $_{342}$LEKRH$_{346}$, $_{360}$EKDS$_{363}$, and $_{383}$EPG$_{385}$) that also match with predicted serotype-specific T_h-cell epitopes and/or neutralizing mAB-binding regions and experimentally determined T_h-cell epitopes [28]. These serotype-specific T_h-cell binders, coupled with the predicted consensus T_h-cell binders (Fig. 2C), reveal dengue immunogenic properties at both the species and serotype levels.

Structural features of dengue E proteins

Figure 3 shows the results of several structural computational analyses to assign functional roles to amino acid residues in dengue E proteins. Based on the relative accessible surface area (ASA), buried residues that are important to maintain the structural integrity of the protein and exposed residues that may provide clues about protein interaction and immunogenicity were identified. The dimer (pre-fusion) and trimer (post-fusion) structures of DENV2 E protein each have about 60 exposed residues (relative ASA >0.6), half of which remain exposed on both the dimer and the trimer surfaces.

The solved structures for DENV2 and DENV3 show that there are minor structural differences at the viral surface of the two serotypes. It has been suggested that the non-conserved residues exposed on the viral protein surface may be involved in differential antibody binding [12]. Among 43 non-conserved residues across the four serotypes (entropy >0.9 bits), 19 are exposed on the surface of either oligomer, including 5 in the dimer only (K_{157}, P_{243}, Q_{293}, E_{343}, and E_{360}), 1 in the trimer only (L_{342}), and 13 in both the dimer and trimer (N_{83}, K_{88}, K_{122}, E_{174}, D_{203}, Q_{227}, S_{274}, S_{300}, D_{329}, R_{345}, H_{346}, D_{362}, and G_{385}). Domain III alone has 9 exposed and non-conserved residues, including 6 exposed in both the dimer and trimer.

A total of 128 interface residues critical for dimerization and trimerization based on the occluded surface area were identified. In particular, it was noted that several residues in the fusion motif, D_{98}, L_{107}, F_{108}, and K_{110}, are involved in interactions at both the dimer and trimer interfaces of DENV2 E proteins (Fig. 3), as well as the dimer interface of DENV3 E protein (not shown). It was further noted that most of these interface residues are highly conserved within the dengue species. Approximately 73% (94 residues) of all interface residues either have entropy scores <0.6 or are negatively selected, and only 4% (5 residues) are non-conserved with entropy scores >0.9. A few [18] interface residues are among the completely conserved residues in all 12 flaviviruses. The interface residues represent potential candidates for mutation experiments that may alter oligomerization. This region may also be a potential target for inhibitors that prevent oligomerization.

There are significant conformational rearrangements in the main chain during the dimer to trimer transition. The plot of the difference in the backbone torsion angles ($\Delta\varphi$ and $\Delta\psi$) shows several regions with major conformational changes, such as 1–19, 242–246, 289–298, and 343–350 (Fig. 3). Many residues change from buried to exposed during the transition, suggesting the importance of these residues in the fusion mechanism. For example, buried residues M_1, H_{244}, K_{246}, G_{254}, G_{330}, and K_{344} in the dimer become exposed in the trimer after significant conformational changes in the main chain, while residues including S_{16}, Q_{52}, Q_{167}, S_{169}, P_{243}, D_{290}, Q_{293}, S_{331}, and E_{343} change from exposed to buried during trimerization.

Comparative analysis of computational and experimental data

To estimate the relative accuracy of our analyses, computational data on sequence conservation, negative selection, structural features, and T-cell epitopes (Figs. 3 and 4) were compared with each other and also with published, experimentally determined functional sites. Such integrated analysis allowed identification of sites that are exposed in the dimer and are also negatively selected (e.g. N_{37}, N_{67}, K_{88}, E_{195}, P_{217}, G_{266}, D_{290}, S_{300}, D_{362}, F_{373}, and E_{383}). Other sites were identified, such as $_{266}$GAT$_{268}$ and $_{445}$GAAFS$_{449}$, which (a) belong to the group of three or more consecutive sites under negative selection pressure in DENV2, (b) have a residue that is negatively selected in at least three of the serotypes (as observed in the merged dataset), and (c) are also part of a predicted epitope. Overall, our computational results are in agreement with experimental information (see supplementary Table S3). The high affinity MHC-II binding peptides predicted here correlated to 80% of the synthetic peptides shown experimentally to induce T-cell immune responses [28]. The correlation between computationally predicted data and available experimental information suggests that the

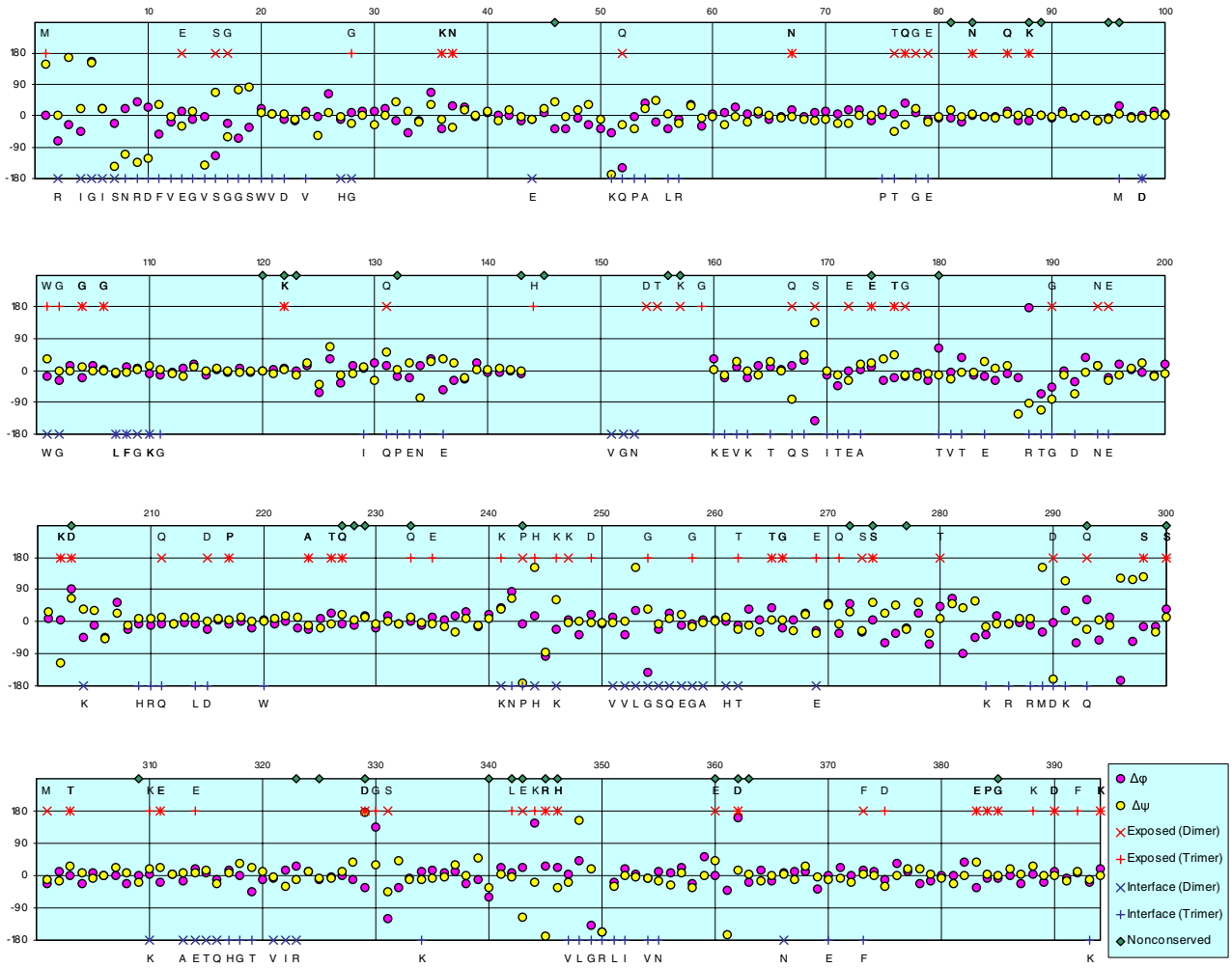

Fig. 3 Structural analysis results of DENV2 E proteins indicating exposed residues, main chain conformational changes and residues in the oligomer interface, in relationship to sequence conservation. The non-conserved residues (entropy >0.9 bits) are shown in the top line. The conformational changes are plotted based on the difference in the backbone torsion angles ($\Delta\varphi$ and $\Delta\psi$) from –180° to 180°. The region of residues 144–159 has no atomic coordinates and is not plotted. The exposed residues (relative accessible surface area >0.6) and the residues in the oligomer interface are indicated on the 180° line and –180° line, respectively, with the following notation: in DENV2 dimer (1OAN) only (×), in DENV2 trimer (1OK8) only (+), and in both dimer and trimer (*)

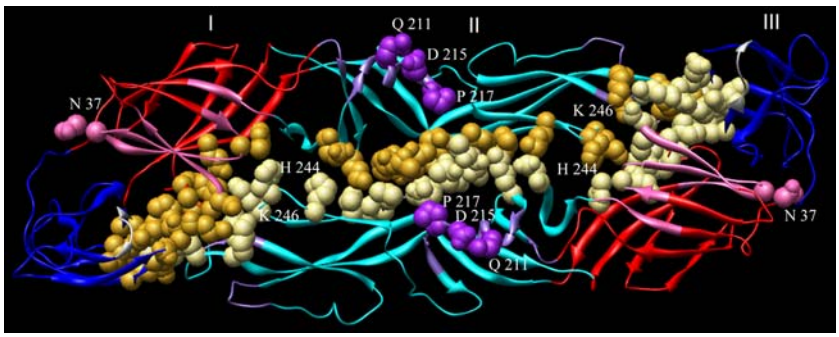

Fig. 4 Ribbon representation of the DENV2 E protein dimer structure (1OAN) showing domains I (red), II (cyan), III (blue) with 12 DENV conserved sequence regions shown in pink, purple, and grey, respectively in the three domains. The dimer interface is shown as gold spheres for chain A and light gold for chain B. In addition, residues that are conserved, part of the consensus T_h-cell epitopes and exposed in the dimer or trimer (N_{37}, Q_{211}, D_{215}, P_{217}, H_{244}, and K_{246}), are shown as spheres and their sequence positions indicated. The residues H_{244} and K_{246} are buried in the dimer (this figure) but are exposed in the trimer

computational approaches used here relate rather well to biological features.

Identification of functionally significant sites

Development of diagnostics with low rates of false negatives and of vaccines difficult to circumvent (by nature or by man) would benefit by identifying the amino acid sites that should remain unaltered in spite of natural changes or artificial modification of dengue virus. We integrated all the computational results described above in this study and searched for sites in E protein that were (a) conserved, (b) consensus T-cell epitopes, and (c) exposed. A site-by-site analysis of the sequence of E protein revealed, as expected, that different features were distributed throughout the sequence. Rather unexpectedly, however, we observed six sites that had more than one feature in a confined region of E protein. The sites having several features might be of particular importance to the viral genome and, therefore, unlikely to change without profound effects on infectivity and/or virus propagation. We considered these six singular sites (Fig. 4; N_{37}, Q_{211}, D_{215}, P_{217}, H_{244}, K_{246}) as potential candidate regions of the E protein for diagnostics and vaccine development. It was noted that 2 sites (N_{37}, P_{217}) are exposed in both dimer and trimer, 2 sites (Q_{211}, D_{215}) are exposed only in the dimer, and 2 sites (H_{244}, K_{246}) are exposed only in the trimer. Out of these six sites, H_{244} and K_{246} undergo conformational change between dimer and trimer forms.

Discussion

In this study, the sequence alignment of the flavivirus E proteins (Fig. 1) and the entropy measure and negative selection results for dengue E proteins (Fig. 2A and B) have allowed us to identify sites and regions that are conserved across the flavivirus genus or within the dengue species, as well as variable regions that reflect serotype-specific functional constraints. Such analyses of conserved sites at the different taxonomic levels allow us to differentiate residues of general importance to the infectivity of flaviviruses from residues that may be specific to the dengue viruses. For example, among the 12 dengue-conserved sequence regions (Fig. 4), regions I–IV and VI–VIII are also highly conserved in other flaviviruses, encompassing over half of the 72 completely conserved residues in flavivirus E proteins. These regions also cover 16 of the 18 DENV2 dimer or trimer interface residues that are completely conserved in all 12 flaviviruses. On the other hand, sequence region V (D_{192}-M_{196}) is dengue species-specific. Interestingly, region V overlaps with a 13-amino acid sequence region that contains 11 negatively selected sites identified to be under functional constraints. Such selection pressure analysis of codon-based DNA alignments is ideal for identifying functionally important residues in serotypes that may not be reflected in the amino acid alignments at the species and serotype level [50].

Several dengue-specific sites were identified. For example, while the N_{153} glycosylation site is conserved in several flaviviruses, the N_{67} site is unique to DENV. It appears that dengue viruses are heterogeneous in their use of the glycosylation sites [51] and the precise function of the second glycosylation site is still under investigation [52]. The significance of this site is not known, although it has been noted that the loss of the N_{67} glycosylation site may result in a higher pH threshold for conformational change [53].We have identified variable residues exposed on the surface of the E protein that are likely to be responsible for the immunogenic variation among dengue serotypes. Especially notable is the sequence region $_{342}$LEKRH_{346} in structural domain III, which consists of five surface exposed residues (1 in dimer, 2 in trimer and 2 in both dimer and trimer) in the beginning of a region (aa 343–350) that undergoes major conformational changes during the dimer to trimer transition, with E_{343} becoming buried and K_{344} becoming exposed. Four (L_{342}, E_{343}, R_{345}, H_{346}) of the five residues are not conserved among the serotypes and have entropy scores >0.9. Another variable and exposed region is within the 10-aa receptor-binding region (I_{380}-L_{389}) that contains 4 exposed residues, including the $_{383}$EPG$_{385}$ triad critical for mAB binding [12].

Recent progress in molecular-based vaccine strategies, such as recombinant subunit dengue vaccines, has provided hope for the control of the disease [6]. The phenomenon of antibody-dependent enhancement of dengue disease has spurred attempts to develop a tetravalent dengue vaccine that produces neutralizing antibodies against all four serotypes [54]. Large-scale analysis of antigenic diversity of T-cell epitopes for dengue virus [55] indicates that there are limited numbers of antigenic combinations in E protein sequence variants, and that short regions of the protein are sufficient to capture the antigenic diversity of T-cell epitopes. Taken together, the 11 predicted consensus T_h-cell epitopes that we identified, especially the 3 epitopes containing the 6 select target sites, are of special interest as potential candidate regions for inclusion in developing epitope-driven vaccines against dengue viruses. A T-cell epitope-driven vaccine design approach has been used for HIV-1 (e.g. the GAIA vaccine) [56] with promising results [57].

In addition to the six select targets that we propose there are several additional promising regions that can be identified from Figs. 1 to 3 according to specific experimental needs. Furthermore, researchers can evaluate epitopes and

diagnostic targets to see if they fall within regions that are under negative selection pressure or are exposed. For example, there are several Nucleotide- and protein-based methods that are currently available for dengue diagnostics [58] and one can choose which targets described in this study are ideal for a specific methodology. The predictive nature of this study allows prioritization of select sites and regions of the DENV E protein for laboratory experimentation with some caveats. First, the development of effective and safe vaccines should involve careful consideration in the selection of exact sequence segments and the choice of specific expression vector systems, both of which are out of the scope of this study. Second, effective diagnostics need to be validated in the laboratory and in the field.

Conclusions

This study provided a priority list of potential target sites in E protein that can be used by experimental biologists involved in dengue diagnostics and vaccine research. A battery of complementary computational tools was necessary to identify salient sites and regions having several desired features simultaneously. This form of detailed computational analysis, coupled with experimental laboratory research, could be instrumental in accelerating the development of viral diagnostics and vaccines.

Acknowledgements This work was supported by the U.S. Department of Defense Chemical and Biological Defense program administered by the Defense Threat Reduction Agency and by In-House Laboratory Independent Research (ILIR) funds from the Research and Technology Directorate, Edgewood Chemical Biological Center, Research Development and Engineering Command, US Army. SLKP and SDWF were supported in part by the National Institutes of Health (AI43638, AI47745, and AI57167, R01-GM66276), the University of California Universitywide AIDS Research Program (IS 02-SD-701). Selection analyses were performed on a computer cluster funded by a University of California, San Diego Center for AIDS Research/NIAID Developmental Award to SLKP (AI36214). We would like to thank Dr. Winona Barker for reviewing the manuscript and providing useful comments and Ms. Natalia Petrova for providing the entropy program.

References

1. M.H. van Regenmortel, M.A. Mayo, C.M. Fauquet, J. Maniloff, Arch. Virol. **145**, 2227–2232 (2000)
2. J.R. Stephenson, Bull. World Health Organ. **83**, 308–314 (2005)
3. G.N. Malavige, S. Fernando, D.J. Fernando, S.L. Seneviratne, Postgrad. Med. J. **80**, 588–601 (2004)
4. J.T. Roehrig, R.A. Bolin, R.G. Kelly, Virology **246**, 317–328 (1998)
5. K.C. Leitmeyer, D.W. Vaughn, D.M. Watts, R. Salas, I. Villalobos de Chacon, C. Ramos, R. Rico-Hesse, J Virol **73**, 4738–4747 (1999)
6. U.C. Chaturvedi, R. Shrivastava, R. Nagar, Indian J. Med. Res. **121**, 639–652 (2005)
7. E. Lee, R.A. Hall, M. Lobigs, J. Virol. **78**, 8271–8280 (2004)
8. Y. Modis, S. Ogata, D. Clements, S.C. Harrison, Nature **427**, 313–319 (2004)
9. T.P. Monath, J. Arroyo, I. Levenbook, Z.X. Zhang, J. Catalan, K. Draper, F. Guirakhoo, J. Virol. **76**, 1932–1943 (2002)
10. E. Lee, R.C. Weir, L. Dalgarno, Virology **232**, 281–290 (1997)
11. F.X. Heinz, S.L. Allison, Adv. Virus Res. **55**, 231–269 (2000)
12. Y. Modis, S. Ogata, D. Clements, S.C. Harrison, J. Virol. **79**, 1223–1231 (2005)
13. K. Hiramatsu, M. Tadano, R. Men, C.J. Lai, Virology **224**, 437–445 (1996)
14. C.H. Wu, R. Apweiler, A. Bairoch, D.A. Natale, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, B. Suzek, Nucleic Acids Res. **34**, D187–D191 (2006)
15. Y. Wang, L.Y. Geer, C. Chappey, J.A. Kans, S.H. Bryant, Trends Biochem. Sci. **25**, 300–302 (2000)
16. N. Deshpande, K.J. Addess, W.F. Bluhm, J.C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, R.K. Green, J.L. Flippen-Anderson, J. Westbrook, H.M. Berman, P.E. Bourne, Nucleic Acids Res. **33**, D233–D237 (2005)
17. J.D. Thompson, D.G. Higgins, T.J. Gibson, Nucleic Acids Res. **22**, 4673–4680 (1994)
18. K. Sjolander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I.S. Mian, D. Haussler, Comput. Appl. Biosci. **12**, 327–345 (1996)
19. D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, W. Helmberg, D.L. Kenton, O. Khovayko, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, J.U. Pontius, K.D. Pruitt, G.D. Schuler, L.M. Schriml, E. Sequeira, S.T. Sherry, K. Sirotkin, G. Starchenko, T.O. Suzek, R. Tatusov, T.A. Tatusova, L. Wagner, E. Yaschenko, Nucleic Acids Res. **33**, D39–D45 (2005)
20. S. Kumar, K. Tamura, M. Nei, Brief Bioinform. **5**, 150–163 (2004)
21. S.L. Kosakovsky Pond, S.D. Frost, S.V. Muse, Bioinformatics **21**, 676–679 (2005)
22. S.L. Kosakovsky Pond, S.D. Frost, Mol. Biol. Evol. **22**, 1208–1222 (2005)
23. S.L. Kosakovsky Pond, S.V. Muse, in HyPhy: Hypothesis Testing Using Phylogenies in Statistical Methods in Molecular Evolution ed. by R. Nielsen (Springer, NY, 2005),pp. 125–182
24. S.L. Kosakovsky Pond, D. Posada, M.B. Gravenor, C.H. Woelk, S.D. Frost, Mol. Biol. Evol. **23**, 1891–1901 (2006)
25. M. Nielsen, C. Lundegaard, P. Worning, C.S. Hvid, K. Lamberth, S. Buus, S. Brunak, O. Lund, Bioinformatics **20**, 1388–1397 (2004)
26. P. Guan, I.A. Doytchinova, C. Zygouri, D.R. Flower, Nucleic Acids Res. **31**, 3621–3624 (2003)
27. P.A. Reche, J.P. Glutting, H. Zhang, E.L. Reinherz, Immunogenetics **56**, 405–419 (2004)
28. J.T. Roehrig, P.A. Risi, J.R. Brubaker, A.R. Hunt, B.J. Beaty, D.W. Trent, J.H. Mathews, Virology **198**, 31–38 (1994)
29. Y. Modis, S. Ogata, D. Clements, S.C. Harrison, Proc. Natl. Acad. Sci. USA **100**, 6986–6991 (2003)
30. A.A. Porollo, R. Adamczak, J. Meller, Bioinformatics **20**, 2460–2462 (2004)
31. N. Pattabiraman, K.B. Ward, P.J. Fleming, J. Mol. Recognit. **8**, 334–344 (1995)
32. W. Kabsch, C. Sander, Biopolymers **22**, 2577–2637 (1983)
33. T. Nowak, G. Wengler, Virology **156**, 127–137 (1987)
34. S. Bressanelli, K. Stiasny, S.L. Allison, E.A. Stura, S. Duquerroy, J. Lescar, F.X. Heinz, F.A. Rey, EMBO J. **23**, 728–738 (2004)
35. J.J. Hung, M.T. Hsieh, M.J. Young, C.L. Kao, C.C. King, W. Chang, J. Virol. **78**, 378–388 (2004)

36. K. Tamura, M. Nei, Mol. Biol. Evol. **10**, 512–526 (1993)
37. S.L. Kosakovsky Pond, S.V. Muse, Mol. Biol. Evol. **22**(12):2375–2385 (2005)
38. G. Blanc, M. Ngwamidiba, H. Ogata, P.E. Fournier, J.M. Claverie, D. Raoult, Mol. Biol. Evol. **22**, 2073–2083 (2005)
39. M. Anisimova, Z. Yang, J. Mol. Evol. **59**, 815–826 (2004)
40. A.J. Leslie, K.J. Pfafferott, P. Chetty, R. Draenert, M.M. Addo, M. Feeney, Y. Tang, E.C. Holmes, T. Allen, J.G. Prado, M. Altfeld, C. Brander, C. Dixon, D. Ramduth, P. Jeena, S.A. Thomas, A. St John, T.A. Roach, B. Kupfer, G. Luzzi, A. Edwards, G. Taylor, H. Lyall, G. Tudor-Williams, V. Novelli, J. Martinez-Picado, P. Kiepiela, B.D. Walker, P.J. Goulder, Nat. Med. **10**, 282–289 (2004)
41. Y. Suzuki, T. Gojobori, Mol. Biol. Evol. **16**, 1315–1328 (1999)
42. Y. Suzuki, T. Gojobori, Gene **276**, 83–87 (2001)
43. Y. Suzuki, Gene **328**, 127–133 (2004)
44. Y. Suzuki, Mol. Biol. Evol. **23**, 1902–1911 (2006)
45. S.S. Twiddy, C.H. Woelk, E.C. Holmes, J. Gen. Virol. **83**, 1679–1689 (2002)
46. C. Klungthong, C. Zhang, M.P. Mammen Jr., S. Ubol, E.C. Holmes, Virology **329**, 168–179 (2004)
47. Z. Yang, J.P. Bielawski, Trends Ecol. Evol. **15**, 496–503 (2000)
48. A.L. Rothman, J. Clin. Invest. **113**, 946–951 (2004)
49. P. Guan, C.K. Hattotuwagama, I.A. Doytchinova, D.R. Flower, Appl. Bioinform. **5**, 55–61 (2006)
50. N. Goldman, Z. Yang, Mol. Biol. Evol. **11**, 725–736 (1994)
51. A.J. Johnson, F. Guirakhoo, J.T. Roehrig, Virology **203**, 241–249 (1994)
52. C.W. Davis, L.M. Mattei, H.Y. Nguyen, C. Ansarah-Sobrinho, R.W. Doms, T.C. Pierson, J. Biol. Chem. **281**, 37183–37194 (2006)
53. F. Guirakhoo, A.R. Hunt, J.G. Lewis, J.T. Roehrig, Virology **194**, 219–223 (1993)
54. D.H. Holman, D. Wang, K. Raviprakash, N.U. Raja, M. Luo, J. Zhang, K.R. Porter, J.Y. Dong, Clin. Vaccine Immunol. **14**, 182–189 (2007)
55. A.M. Khan, A. Heiny, K.X. Lee, K. Srinivasan, T.W. Tan, J.T. August, V. Brusic, BMC Bioinform. **7**(Suppl 5), S4 (2006)
56. A.S. De Groot, E.A. Bishop, B. Khan, M. Lally, L. Marcon, J. Franco, K.H. Mayer, C.C. Carpenter, W. Martin, Methods **34**, 476–487 (2004)
57. O.A. Koita, D. Dabitao, I. Mahamadou, M. Tall, S. Dao, A. Tounkara, H. Guiteye, C. Noumsi, O. Thiero, M. Kone, D. Rivera, J.A. McMurry, W. Martin, A.S. De Groot, Hum. Vaccin. **2**, 119–128 (2006)
58. C.L. Kao, C.C. King, D.Y. Chao, H.L. Wu, G.J. Chang, J. Microbiol. Immunol. Infect. **38**, 5–16 (2005)