J Vector Borne Dis 50, September 2013, pp. 188–196

Positive selection and evolution of dengue type-3 virus in the Indian subcontinent

Sukhmani K. Bedi, Apar Prasad, Krishanu Mathur & Sonika Bhatnagar

Computational and Structural Biology Laboratory, Division of Biotechnology, Netaji Subhash Institute of Technology, New Delhi, India

ABSTRACT

Background: Dengue virus infection has recently taken endemic proportions in India with dengu type-3 (DEN-3) as a predominant serotype. In this study, we carried out the selection pressure analysis of three critical immunogenic regions of DEN-3. Phylogenetic analysis was then carried out on the positively selected genomic region in the DEN-3 virus strains isolated in the Indian subcontinent over a time span of 25 yr (1984–2008). Bayesian Markov chain Monte Carlo (MCMC) calculation of the substitution rate was carried out for the DEN-3 genotype-III sequences.

Methods: Sequences corresponding to the C-prM, E-NS1 and NS1 sequence regions of DEN-3 strains were taken for the positive selection analysis. The C-prM junction sequences were then used to construct a maximum likelihood (ML) phylogenetic tree. Substitution rates were also calculated under various models of population growth.

Results: It was found that codon 86, corresponding to a conserved arginine residue in a crucial T-cell epitope of the C-protein was under significant positive selection. The K86R substitution was found to exist in almost all the Indian strains isolated after 2004. The ML tree constructed from the C-prM junction sequences indicated that strains from the 2006 dengue incidences in Delhi, namely: 04/03/del2006, 05/03/del2006, and 06/03/del2006 were the most rapidly evolving. Substitution rates of a DEN-3 genotype-III sequences from the Indian subcontinent were found to be ~3.0 times higher than those reported from other parts of the world.

Conclusion: Positive selection in the codon corresponding to R86 of the highly conserved surface C-protein is important in view of its occurrence in a T-cell epitope as well as its strict conservation in all the DEN strains. Phylogenetic analysis of the C-prM junction sequences showed that three strains of 2006 are rapidly evolving. These results were also supported by calculations of the substitution rates. Their significance in the expansion of viral epidemics requires to be investigated.

Key words C-prM; DEN-3; E-NS1; genotype-III; maximum likelihood tree; NS1; phylogenetic analysis; positive selection; predicting evolution; T-cell epitope

INTRODUCTION

Dengue is the most prevalent arbo-viral infection of humans in the tropical and subtropical countries of the world. It is estimated that 50 to 100 million infections occur yearly[1]. The dengue virus (DEN) belongs to the genus *Flavivirus* and transmitted to man through the bite of the mosquito, *Aedes aegypti*. Some other species such as *Ae. albopictus* are also potent vectors[2]. Besides being the causative agent for the benign febrile flu-like illness, DEN is also responsible for the more fatal dengue haemorrhagic fever (DHF) or dengue shock syndrome (DSS). DEN consists of four antigenically distinct serotypes (DEN-1 through DEN-4). Infection with one DEN serotype provides lifelong immunity to that virus, but there is no cross-protective immunity to the other serotypes.

Although all the four DEN serotypes have been periodically reported in the countries of the Indian subcontinent, but DEN-3 has been found responsible for most of the severe epidemics of DHF and DSS[3–7]. Sri Lanka and India experienced a long period, wherein no confirmed cases of DHF epidemics were reported. It has been shown by phylogenetic analysis of the envelope gene that the DEN-3 virus can be further classified into five distinct genotypes (I–V), each differing in its virulence and ability to infect host cells. The re-emergence (1989 in Sri Lanka and 1990 in India) of DHF has been linked to the appearance of a new genetically distinct DEN-3 genotype-III, the spread of which is aided by geographical proximity[8]. During 2001–05 outbreaks of DEN in the Indian subcontinent also had a clear predominance of DEN-3 (genotype-III)[7–11]. The role of viral evolution in determining dynamics of the disease is well-established in a number of viral diseases like influenza A, HIV-1, hepatitis C, chikungunya, and dengue[12–18]. Evidence for positive selection of amino acid residues in the envelope (E) protein of the Indian DEN-2 strains has been presented

previously[19]. However, similar studies are not available for DEN-3 sequences.

A study of the evolutionary pressures acting on antigenically important regions could help trace the evolutionary patterns of DEN-3 in the Indian subcontinent. In view of the recurrent epidemics of dengue in the region and the lack of evolutionary studies on DEN-3, we attempted to find evidences for adaptive evolution in three critical immunogenic sequence regions of DEN-3. We also estimated the rate of substitution for the C-prM junction sequences and studied the population dynamics of DEN-3 in the Indian subcontinent. From this analysis, inference about the strains of DEN-3 rapidly evolving in India could be made.

MATERIAL & METHODS

Sequence retrieval and data collection

The National Centre for Biotechnology Information (NCBI) GenBank was searched using "Dengue type-3, DEN-3 and Dengue-3" as key words and the results were filtered into different sets according to the country of the origin in the Indian subcontinent. In addition, seven other strains collected from Taiwan, Thailand, Indonesia, Brazil, China, and Philippines, used as reference strains in previous studies, were included in the dataset[7–8, 20]. These are PHI56-H87 (323468), JAK88-Den88, THA94-CO331, TAI98-TW368, BRA02-4886, TAI99-TW628, and CHI80-80-2. Due to non-availability of the DEN-3 genotype-IV C-prM sequences, these could not be included as reference sequences in the phylogenetic tree. For ease of analysis and representation, the sequences were represented in short notation in the order of (1) region, (2) year of occurrence followed by '–', and (3) name of the strain.

Selection pressure analysis

To identify the existence of positive selection pressure at individual codon sites, three likelihood procedures were used as in previous studies[19, 21–22], i.e. the single-likelihood ancestor counting (SLAC) method, fixed effects likelihood (FEL) method, and the more powerful random effects likelihood (REL) method. The strength of selection pressure is determined on the basis of the ratio of non-synonymous (dN) to synonymous (dS) substitutions per site (ratio: dN/dS). The analysis was carried out using the online Datamonkey facility (*http://www.datamonkey.org*), incorporating the General Time Reversible (GTR) model of nucleotide substitution for all the three datasets.

Positive selection analysis of epitopic sites is advantageous because the mutations in these sites are highly favourable for adaptive evolution[12, 21]. Therefore, in addition to the full-length sequences, positive selection analysis was also performed for the antigenic epitope sites[22], as obtained from the Immune Epitope Database (IEDB) (*http://www.immuneepitope.org*)[23]. The sequence positions of the epitopes were available through previously published studies[24–25]. The list of epitopes in the three regions along with their details is shown in Table 1.

Table 1. The epitope sequences in the selected regions of DEN-3

Gene	Range	Linear sequence	Reference/ IEDB ID*
C**	81 – 92	LKGFK**K**EISNML	24
E†	434 – 448	VHQIFGSAYTALFSG	1000409
	439 – 453	GSAYTALFSGVSWVM	1000409
	444 – 458	ALFSGVSWVMKIGIG	1000409
	449 – 463	VSWVMKIGIGVLLTW	1000409
	454 – 468	KIGIGVLLTWIGLNS	1000409
	474 – 488	SFSCIAIGIITLYLG	1000409
	479 – 493	AIGIITLYLGAVVQA	1000409
NS1‡	5 – 13	VINWKGKEL	1005810
	12 – 20	ELKCGSGIF	24
	266 – 274	GPWHLGKLE	24
	294 – 302	RGPSLRTTT	24

*IEDB ID for direct submission entries; **Capsid protein with the positively selected K86 shown in bold; †Envelope protein; ‡Non-structural protein-1.

Phylogenetic analysis

Multiple sequence analysis of each data set was done using TCOFFEE (*http://tcoffee.vital-it.ch*)[26]. The percentage of sequence identity was computed using BioEdit v3.6[27]. ML phylogenetic tree was constructed using PhyML (ver.3.0)[28], using bootstrap analysis with 1000 re-samplings. The tree was rooted at the oldest sequence. Two more methods, i.e. maximum parsimony (MP) and Bayesian MCMC were used to confirm the tree topology as predicted by the ML tree. The best substitution model for the datasets was determined using jModelTest[29–30], which was then used to prepare the trees. For all datasets, the best substitution model as determined by jModelTest was TIM+G (transition model with gamma distributed rate variation among sites). However, the GTR+G model was used for tree construction as the GTR model is closest to TIM, among the options available in the MrBayes 3 and PhyML.

The MP tree construction was done using the DNAPARS (deoxyribonucleic acid parsimony) module of Phylip[31], followed by bootstrap analysis with 1000 re-samplings. MrBayes 3[32] was used to build the Bayesian MCMC tree using GTR+G as the nucleotide substitution

model with 800,000 generations, sampling frequency as 100 and 25% of the generations as burn-in.

Estimation of substitution rate and comparison of various models of population dynamics

Estimation of nucleotide substitution per site, per year was done using the Bayesian MCMC protocol implemented in the BEAST package (*http://beast.bio.ed.ac.uk/*). The GTR model as determined previously to best fit the substitution model was used to estimate the phylogenetic trees under the assumption of strict or relaxed (uncorrelated lognormal) molecular clock. The demographic models compared were: Constant population size, exponential population growth, logistic and expansion growth, and a Bayesian skyline model. The analyses were run to convergence, as determined by the Tracer programme (*http://tree.bio.ed.ac.uk/software/tracer/*), and uncertainty in the estimates of all the parameters was determined as values of the 95% highest probability density (HPD) interval. The minimum effective sample size was >200 for each run. The models obtained were compared using Bayes factors.

RESULTS & DISCUSSION

Three critical immunogenic genomic regions of DEN-3 are available for sequence analysis, namely: E-NS1, C-prM and NS1

The dengue ribonucleic acid RNA genome is a single-stranded positive-sense genome of approximately 10,700 bases in length, surrounded by a nucleocapsid and covered by a lipid envelope containing the envelope and the membrane proteins. The genome contains a single open reading frame (ORF) flanked by two non-translated regions [5' and 3' untranslated region (UTR)]. The ORF encodes a precursor polyprotein that is proteolytically cleaved to form the functional protein products. The first structural protein is the C-protein followed by prM and M proteins. The last structural protein is the E protein. Following the structural region is the non-structural (NS) region comprising of five proteins: NS1, NS2, NS3, NS4 and NS5.

Though the E protein is the major target for neutralizing antibodies[33], the anti-prM monoclonal antibodies also react strongly with the M protein[34]. Experimental studies have also confirmed a cytotoxic T-cell (CTL) response against NS1[35] and the capsid protein[36]. Although the function of NS1 is yet to be fully defined, it is also known to be an important immunogen[37]. Hence, previous studies have analyzed the evolutionary pressures acting on these antigenically important regions and also used them for tracing the evolutionary pattern of dengue[19–20]. In the present study, the NCBI sequence database search identified a sufficient number of sequences at three genomic locations, namely, C-prM, E-NS1, and NS1. A total of 84 sequences were obtained for the *E-NS1* gene junction, 84 for the *C-prM* gene junction, and 70 for NS1. In all the sequences, nucleotide sequence identity ranges from 92 to 99.5% whereas the amino acid identity ranges between 94 and 100% in comparison to the oldest DEN-3 sequence, i.e. 1956 DEN-3 genotype-I isolated from Philippines. The number of sequences obtained for other regions of the DEN genome was not sufficient for our analysis. The Gi numbers, along with the strain names, year, and country of isolation, of all the sequences used in the analysis are available with the authors

A large number of mutations occur in the C-prM region of DEN-3

TCOFFEE generated multiple sequence alignment of each dataset details available with the authors. Sequence analysis showed that single nucleotide substitutions were scattered throughout the entire length of the C-prM junction. However, most of these substitutions occurred at the third codon position were not reflected in the amino acid sequence obtained. This is also evident from the higher similarity between protein sequences as compared with their nucleotide sequences. Compared to the root, several unique amino acid substitutions, affecting size, volume, polarity, hydrophobicity, and ionic properties, were found within the three genomic regions, with the largest number of mutations being in the C-prM junction region. The non-synonymous mutations resulting in amino acid substitutions in three regions are as follows:

C-prM: The multiple sequence alignment was obtained for amino acid positions 32–114 and 1–56 of C and prM, respectively. With respect to the root strain, three substitutions in the protein sequences of the C-prM junction were located in the genotype-III sequences. These include R55K, M109I, and T113A in the C-protein. Interestingly, most of the Indian strains obtained after 2004 were found to have the K86R mutation in C-protein. This is significant in the light of the strong conservation of R86 in all the sequences after 2004. Also, a search of the epitope database confirmed the presence of this substitution in an antigenically important T-cell epitope. The sequence, as well as the starting and ending positions of this T-cell epitope are shown in Table 1.

E-NS1: The multiple sequence alignment was obtained for positions 431–492 and 1–32 positions of E and NS1, respectively. Compared to the root strain, unique amino

acid substitutions were observed in the *E* gene of the strains collected from Bangladesh, including S447G, A479V, and V489T (unlike V489A for sequences from other countries). Compared to genotypes I, II, and V, genotype-III sequences possess the substitution I452V in E-protein. This is a part of a recognized epitope and change in this region may be immunologically significant (Table 1).

NS1: The multiple sequence alignment was obtained for the region 206–319 of the NS1 protein. Unique amino acid substitutions with respect to the root strain were observed in the *NS1* gene of the strains collected from Bangladesh. These include L216F and N290D of the mature NS1 protein. Specifically, the substitution Y256H, which causes a change in the ionic property of this residue, was found only in genotype-III sequences.

Position 86 of the C-protein is under positive selection in Indian DEN-3 sequences

In view of the potential amino acid substitutions, as identified above, all the three datasets were subjected to positive selection pressure analysis. The results of this analysis are as follows:

C-prM: SLAC, REL, and FEL analyses of the C-prM sequences suggested 19 potential codons to be under positive selection. When the entire dataset was used for analysis, no codon showed positive selection at a statistically acceptable significance level (i.e. p <0.1 for SLAC and FEL, Bayesian factor >100 for REL). However, positive selection analysis of the 47 Indian C-prM sequences, out of the 76 gentoype-III sequences from the Indian subcontinent, showed codon 86 and 99 to be undergoing positive selection. The values obtained for the positive selection pressure analysis by the three methods (SLAC, FEL, and REL) are listed in Table 2. As shown, only codon 86 showed positive selection at a high confidence value (p = 0.0876 and Bayesian factor = –3049). The number of synonymous substitutions (dS value) in this codon is zero, whereas the number of non-synonymous substitutions (dN value) is determined to be 3.54, by FEL analysis. For further support, the epitopic positive selection analysis was also carried out for the Indian sequences subset. It was found that the codon corresponding to the amino acid at position 86 of the epitope sequence LKGFKKEISNML (81–92) was under positive selection. This sequence constitutes a highly conserved human leukocyte antigen (HLA) class II-restricted dengue-specific peptide epitope, involved in the T-cell response (Table 1). This T-cell epitope is observed in all the serological variants of DEN, with each serotype having its distinct consensus sequence for the epitope. It interacts with $CD4^+$ cells to elicit a cytokine response from the T-helper cells[25]. Figure 1 depicts the consensus sequences for each serotype of the virus and also highlights the change in consensus observed at amino acid position 6 of the epitope for the DEN-3 strains in India post 2004. It has been demonstrated to have distinct consensus sequences in DEN serotypes I–IV. The codon is also conserved in other flavivirus such as the *Kokobera*, *Aroa*, and *Kedougou* viruses.

E-NS1: SLAC, REL, and FEL analyses of E-NS1 sequences suggested nine possible codon sites to be under positive selection. However, none of these sites was found to show positive selection at a significant level. Selection pressure analysis of the subset of sequences from different geographic locations also did not yield any positively selected codons.

NS1: None of the sites showed positive selection at a

Table 2. Parameters from selection pressure analysis using SLAC, REL and FEL methods

Codon	SLAC		FEL		REL	
	dN-dS	*p*-value	dN-dS	*p*-value	dN-dS	Bayesian factor
C-prM genotype-III (n = 76)*						
99	5.8	0.401	3.516	0.1616	–	–
110	4.2	0.459	3.040	0.200	–	–
143	5.5	0.444	–	–	–	–
86	–	–	–	–	1.424	362.269
C-prM genotype-III, India (n = 46)						
86	9.105	0.266	13.41	0.0876+	0.3455	3048.88+
99	5.1681	0.39	6.5609	0.155	–	–

*Capsid-pre-membrane protein junction; +High significance value; SLAC–Single likelihood ancestor counting; FEL–Fixed effect likelihood; REL–Random effects likelihood.

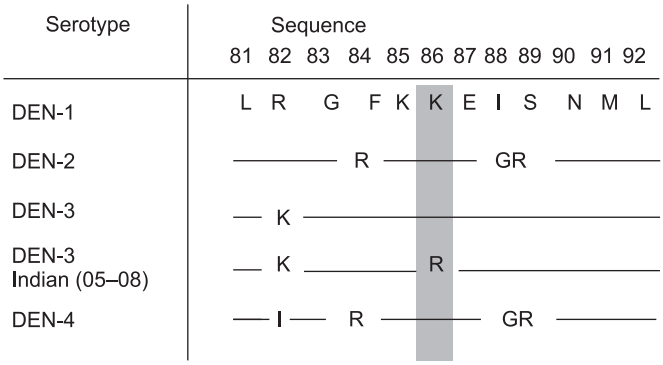

Fig. 1: Consensus sequence of T-cell epitope. Serotype-specific consensus sequences of the HLA class-II restricted T-cell epitope at amino acid position 81–92 of the capsid protein. The data about consensus sequences of DEN-1, DEN-2 and DEN-3 were taken from Ref. 25. "_" below each character represents the occurrence of the same/identical sequence as the first. Wherever there is a variation in sequence, the change in the sequence has been shown.

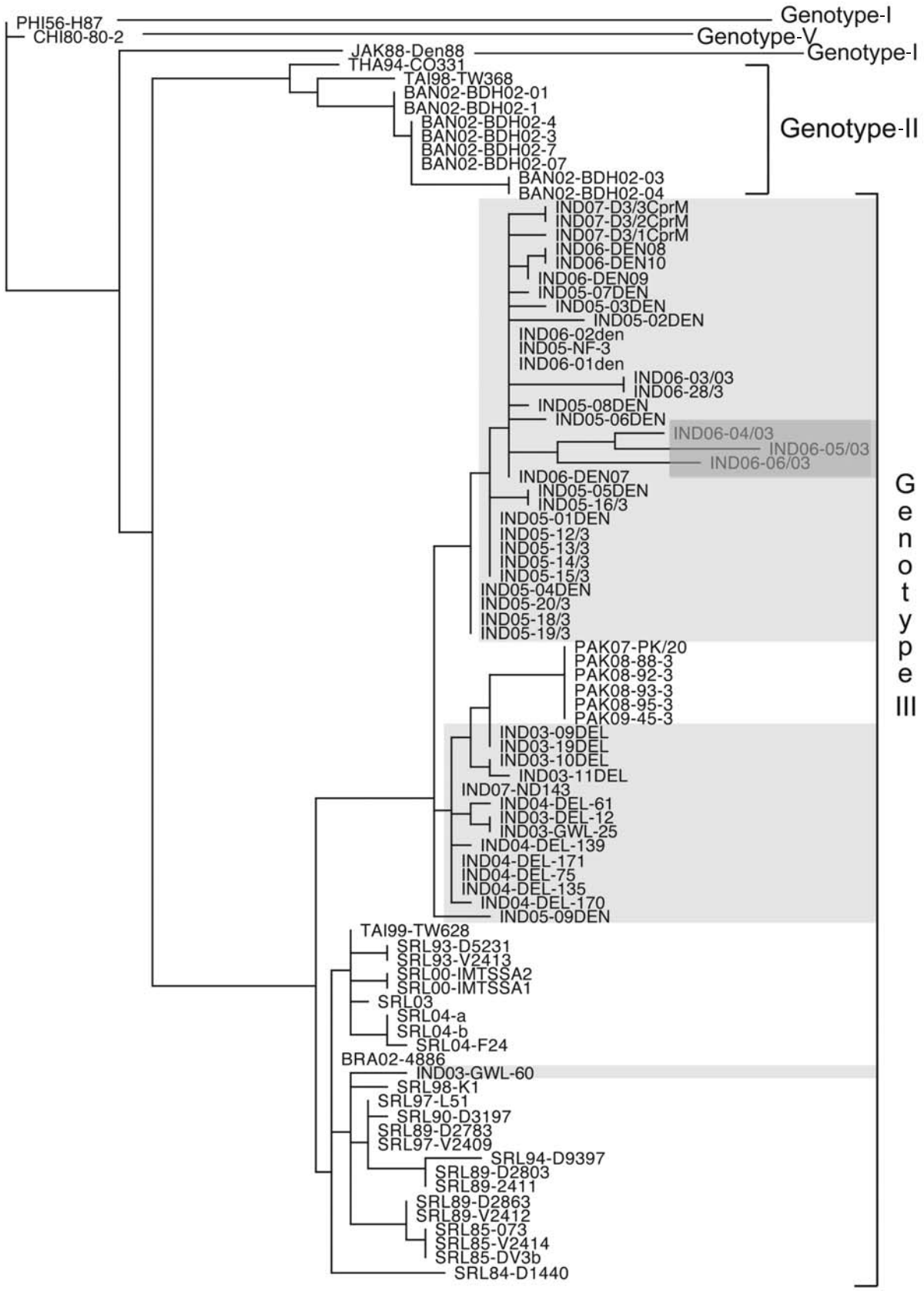

Fig. 2: The phylogenetic tree of C-prM junction sequences of DEN-3 constructed using PhyML. The tree was constructed using the GTR + G model with 1000 bootstrap re-samplings. The genotype segregation is shown using brackets, the Indian sequences are shown in light grey boxes, and the most evolving strains have been highlighted in dark grey box.

high significance level. Therefore, no site was taken-up for further analysis.

Most rapidly evolving Indian DEN-3 strains–05/03/del2006, 06/03/del2006 and 04/03/del2006

As positive selection was established in the C-prM sequences, this dataset was further used for phylogenetic analysis of the DEN-3 sequences. Additional strains from different regions, genotypes of DEN-3, as well as ancestral sequences, were added to the dataset as reference strains for tree building. The ML phylogenetic tree obtained is shown in Fig. 2. It was found that Bayesian MCMC and MP methods also gave trees that were in agreement with the ML tree. All the 47 strains of the Indian origin segregated with the genotype-III Sri Lankan strains, as well as the genotype-III sequences of Brazil and Taiwan (1999). The presence of Brazilian strains in the clade for strains from Sri Lanka is in concert with earlier studies that have established the migration of DEN-3 genotype-III from Sri Lanka to South America[8]. The strains from Bangladesh formed a distinct separate clade with the genotype-II strains of Taiwan (1998) and Thailand, in accordance with a previous study[38]. As expected, the genotype-I sequences of Philippines and Jakarta and genotype-V sequence of China have no subcontinent sequences in their vicinity. These results are in agreement with the established geographic segregation of the DEN-3 genotypes[8]. As can be observed from the evolutionary relationship between the strains, ancestors of the respective countries are responsible for the present-day strains in the same region. Unlike all the 2003 and 2004 Indian viruses, GWL-60 was found to be very closely related to the Sri Lankan strains from 1998. This occurrence has been observed previously[7]. Therefore, geographic and genotypic segregation, as well as agreement with previous studies, lends further confidence to the phylogenetic tree obtained.

The Indian clade consists of two distinct groups—one containing sequences from 2003–04 from Delhi; and the two exceptions to this are a 2007 sequence from Hyderabad (INDO7–ND143) and a 2005 sequence from Delhi (INDO5-09DEN). The other group consists of sequences from Delhi post-2004. The change observed in the C-prM sequences from India post-2004 is the K86R substitution in the C-protein. This substitution is present at amino acid position 6 of a T-cell epitope. This change can be correlated to the transition from all the four DEN serotypes circulating in 2003 and 2004 to complete predominance of DEN-3 in 2005[3]. However, further studies are required to study this transition completely.

The trunk of the tree is defined as the set of interior nodes leading from the root down to the tip, i.e. farthest removed from the root. Previous studies on human influenza posit that the trunk lineage undergoing the most amino acid replacements along the path of branches joining the root of the tree to each of the terminal nodes should identify the section of the tree from which the future trunk lineage will emerge[12, 39]. In this study, we have identified the trunk and three strains 04/03/del2006, 05/03/del2006 and 06/03/del2006 that have undergone maximum number of amino acid substitutions. These strains were also shown to be highly evolved in phylogenetic trees constructed in an earlier study[40]. In the case of the DEN-3 virus, it has been established that in spite of the temporal clustering of viral lineages, previously rare lineages often give rise to the most common genotype at any particular sampling time[41]. Therefore, the significance of the lineages identified for future pharmaco-surveillance efforts is required to be investigated further.

Substitution rates of DEN-3 genotype-III in the Indian subcontinent

As it had previously been reported that the DEN-3 genotype-III is evolving significantly faster than DEN-1 and DEN-2. Bayesian MCMC rates of nucleotide substitution were determined for the genotype-III sequences used in this study. Estimation of substitution rates was also carried out by subdividing the sequences into genotype-III sequences from India and other countries. The substitution rates obtained for the set of 76 "All" sequences were similar under all demographic models and clock rates, reflecting the robustness of this analysis (Table 3). The strict molecular clock was rejected in favour of the relaxed clock in all the population growth models tested. The values of the observed rates of substitution and highest probability density (HPD) for all the sequences under the different clocks are summarized in Table 3. The logistic growth and constant population size models were positively favoured (Bayes factor of 3.6 and 3.2, respectively). The most favoured model, the logistic population growth model has a phase of rapid growth followed by an increasingly slower growth. However, strong evidence in favour of any model was not observed.

The mean substitution rate obtained for all the DEN-3 genotype-III sequences using logistic growth model under a relaxed clock was 28.4×10^{-4} substitutions site^{-1} yr^{-1} (95% HPD = 39.2×10^{-4}–18.6×10^{-4} substitutions site^{-1} yr^{-1}). Though the rates obtained were of the same order as those reported for the Asian DEN-3 genotype-III[42–43], it was observed that the magnitude of the mean substitution rates for this set having a large number of

Table 3. Nucleotide substitution rates and population dynamics study of DEN-3 genotype-III C-prM sequences

Demographic model		Constant	Exponential	Logistic	Expansion	Bayesian skyline
No. of sequences			76			
Date range			1984–2008			
Mean substitution rate* (Strict)		19.7×10^{-4}	19.3×10^{-4}	19.4×10^{-4}	19.7×10^{-4}	17.2×10^{-4}
HPD+ substitution rate	U	29.2×10^{-4}	28.0×10^{-4}	28.0×10^{-4}	28.4×10^{-4}	25.6×10^{-4}
	L	11.5×10^{-4}	11.8×10^{-4}	11.5×10^{-4}	11.6×10^{-4}	9.8×10^{-4}
Mean substitution rate* (Relaxed)		28.1×10^{-4}	27.3×10^{-4}	28.4×10^{-4}	28.2×10^{-4}	26.6×10^{-4}
HPD** substitution rate	U	38.3×10^{-4}	38.0×10^{-4}	39.2×10^{-4}	38.9×10^{-4}	37.7×10^{-4}
	L	18.6×10^{-4}	17.1×10^{-4}	18.6×10^{-4}	18.1×10^{-4}	16.1×10^{-4}

*Substitutions site^{-1} yr^{-1}; **Highest probability density—(U) upper, and (L) lower bounds.

sequences from India and the Indian subcontinent is consistently higher (~3.0 times). Thereafter, the calculations of substitution rates were carried out by dividing this set into two different subsets, one belonging to India and another to other countries.

The rate of substitution was strikingly different in the two subsets. The rate of evolution was very close to those earlier reported before DEN-3 in previous works in other countries. Under the constant size model of population with a relaxed molecular clock, the mean substitutions rate for the 30 DEN-3 genotype-III sequences from countries other than India was 11.7×10^{-4} substitutions site^{-1} yr^{-1} (95% HPD = 5.5×10^{-4}–18.5×10^{-4} substitutions site^{-1} yr^{-1}). However, in the case of 46 Indian sequences with dates ranging from 2004–07, the mean substitution rates under the constant size model of population growth with a relaxed molecular clock was 69.2×10^{-4} substitutions site^{-1} yr^{-1} (95% HPD = $32 \times 10^{-4} - 11.2 \times 10^{-3}$ substitutions site^{-1} yr^{-1}). Both the mean substitutions rate and the HPD values are much higher for the sequences of Indian origin, showing their high evolutionary rate. This is in agreement with our observations made during positive selection and phylogenetic analysis. However, the substitution rates obtained in the two subsets should be treated with caution due to the short time span of the Indian sequences and the possible confounding effects of subdividing into two sets. Further work on collection and calculation of substitution rates in DEN-3 sequences of Indian origin from a longer time period is required to be undertaken to further verify and validate the high evolutionary rate of DEN-3 genotype-III in India.

CONCLUSION

In this study, we have analyzed the diversity of sequences among the sequences of C-prM, E-NS1 junctions as well as the NS1 protein of DEN-3 in the Indian subcontinent and established the presence of positive selection pressures in the C-prM region. In C-prM, it was found that the codon corresponding to position 86 of the C-protein is undergoing for positive selection. This codon is present in a T-cell epitope, and has been established as an important determinant for eliciting cytokine response. The K86R substitution results in the change in the basicity of the residue. It also causes an increase in surface area (by about 25Å^2), and can alter the structure as well as the binding properties in other proteins[44–45]. It is also established that vector-borne RNA viruses are highly conserved and specifically, the structural proteins are less prone to positive selection. This is attributed to their role in capsid formation[46]. The tree constructed using the C-prM junction sequences was correct and supported by previous studies. It indicated the presence of two distinct groups within the Indian clade. Using this tree, we have also identified the lineage of three strains, namely: 04/03/del2006, 05/03/del2006, and 06/03/del2006, that are evolving rapidly. The substitution rates in DEN-3 genotype-III in our dataset having a preponderance of sequences from the Indian subcontinent showed that the rates were higher than those reported previously for DEN-3. Subdividing the population of DEN-3 genotype-III showed that substitution rates for sequences from countries other than India were similar to those reported in other parts of the world[42–43] whereas those from Indian strains were much higher. The significance of these strains for immune/pharmaco-surveillance as possible virulent strains for future epidemics needs to be investigated.

To the best of our knowledge, this is the first study aimed at studying positive selection, substitution rates and predicting the evolution of DEN-3 in India. This study identifies potential sites for positive selection in the C-prM junction region of Indian strains and also sheds light on the possible causative agents for future dengue epi-

demics in India. As most of the epidemics in India have occurred in Delhi, almost all sequences in this dataset were isolated from Delhi, underlining the association of urban centres with DEN. However, some sporadic incidents of dengue have been recorded from other cities like Gwalior, Hyderabad, Bengaluru, and Thiruvanantapuram. An increase in the number of sequencing projects with a focus on the antigenically important *E* gene would be critical for understanding the evolution of the virus and will help in the identification of epitopes for the development of vaccines.

ACKNOWLEDGEMENTS

The authors thank Prof. B. Jayaram and the supercomputing facility at the Indian Institute of Technology (IIT), New Delhi for making available the computational resources required for phylogenetic analysis. They would like to acknowledge the kind help of Pooja Khurana for running the programmes and general discussion. They are also grateful to Brian O'Meara and Atul Wamble for providing prompt answers and valuable insights.

REFERENCES

1. *Fact sheet: Dengue and dengue haemorrhagic fever*. Geneva: World Health Organization 2002.
2. Gubler DJ. *The arboviruses: Epidemiology and ecology*. In: Monath TP, editor. II edn. Boca Raton, Fla: CRC Press 1988; p. 223–60.
3. Gupta E, Dar L, Kapoor G, Broor S. The changing epidemiology of dengue in Delhi, India. *Virol J* 2006; *3:* 92–6.
4. Singh UB, Maitra A, Broor S, Rai A, Pasha ST, Seth P. Partial nucleotide sequencing and molecular evolution of epidemic causing dengue-2 strains. *J Infect Dis* 1999; *180:* 959–65.
5. Dash PK, Parida MM, Saxena P, Kumar M, Rai A, Pasha ST, *et al*. Emergence and continued circulation of dengue-2 (genotype-IV) virus strains in northern India. *J Med Virol* 2004; *74:* 314–22.
6. Dash PK, Saxena P, Abhyankar A, Bhargava R, Jana AM. Emergence of dengue virus type-3 in northern India. *Southeast Asian J Trop Med Public Health* 2005; *36:* 370–7.
7. Dash P, Parida M, Saxena P, Abhyankar A, Singh CP, Tewari KN, *et al*. Reemergence of dengue virus type-3 (subtype-III) in India: Implications for increased incidence of DHF and DSS. *Virol J* 2006; *3:* 55.
8. Lanciotti RS, Lewis JG, Gubler DJ, Trent DW. Molecular evolution and epidemiology of dengue-3 viruses. *J Gen Virol* 1994; *75:* 65–75.
9. Aziz MM, Hasan KN, Hasanat MA, Siddiqui MA, Salimullah M, Chowdhury AK, *et al*. Predominance of the Den-3 genotype during the recent dengue outbreak in Bangladesh. *Southeast Asian J Trop Med Public Health* 2002; *33:* 42–8.
10. Jamil B, Hasan R, Zafar A, Bewley K, Chamberlain J, Mioulet V, *et al*. Dengue virus serotype 3, Karachi, Pakistan. *Emerg Infect Dis* 2007; *13:* 182–3.
11. Dorji T. Diversity and origin of dengue virus serotypes 1, 2, and 3, Bhutan. *Emerg Infect Dis* 2009; *15:* 1630–2.
12. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM. Predicting the evolution of human influenza A. *Science* 1999; *286:* 1921–5.
13. Zanotto PA, Kallas EG, de Souza RF, Holmes EC. Genealogical evidence for positive selection in the *nef* gene of HIV-1. *Genetics* 1999; *153:* 1077–89.
14. Manzin A, Solforosi L, Debiaggi M, Zara F, Tanzi E, Romano L, *et al*. Dominant role of host selective pressure in driving hepatitis C virus evolution in perinatal infection. *J Virol* 2000; *74:* 4327–34.
15. Hatta M, Gao P, Halfmann P, Kawaoka Y. Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. *Science* 2001; *293:* 1840–2.
16. Tong JC, Simarmata D, Lin RTP, Rénia L, Ng LFP. HLA Class-I restriction as a possible driving force for chikungunya evolution. *PloS ONE* 2010; *5:* e9291.
17. Twiddy SS, Woelk CH, Holmes EC. Phylogenetic evidence for adaptive evolution of dengue viruses in nature. *J Gen Virol* 2002; *83:* 1679–89.
18. Bennett SN, Holmes EC, Chirivella M, Rodriguez DM, Beltran M, Vorndam V, *et al*. Selection-driven evolution of emergent dengue virus. *Mol Biol Evol* 2003; *20:* 1650–8.
19. Kumar SRP, Patil JA, Cecilia D, Cherian SS, Barde PV, Walimbe AM, *et al*. Evolution, dispersal and replacement of American genotype dengue type 2 viruses in India (1956–2005): Selection pressure and molecular clock analyses. *J Gen Virol* 2009; *91:* 707–20.
20. King CC, Chao DY, Chien LJ, Chang GJJ, Lin TH, Wu YC, *et al*. Comparative analysis of full genomic sequences among different genotypes of dengue virus type 3. *Virol J* 2008; *5:* 63.
21. Wolf YI, Viboud C, Holmes EC, Koonin EV, Lipman DJ. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biol Direct* 2006; *1:* 34.
22. Pond SL, Frost SD. Datamonkey: Rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* 2005; *21:* 2531–3.
23. Vita R, Zarebski L, Greenbaum J, Emami H, Hoof I, Salimi N, *et al*. The immune epitope database 2.0. *Nucleic Acids Res* 2010; *38:* D854–62.
24. Khan AM, Miotto O, Heiny AT, Salmon J, Srinivasan KN, Nascimento E, *et al*. A systematic bioinformatics approach for selection of epitope-based vaccine targets. *Cell Immunol* 2006; *244:* 141–7.
25. Mangada MM, Rothman AL. Altered cytokine responses of dengue-specific CD4 + T-cells to heterologous serotypes. *J Immunol* 2005; *175:* 2676–83.
26. Notredame C, Higgins DG, Heringa J. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 2000; *302:* 205–17.
27. Hall TA. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Sym Ser* 1999; *41:* 95–8.
28. Guindon SP, Rodrigo AG, Dyer KA, Huelsenbeck JP. Modeling the site-specific variation of selection patterns along lineages. *Proc Natl Acad Sci USA* 2004; *101:* 12957–62.
29. Posada D. jModeltest: Phylogenetic model averaging. *Mol Biol Evol* 2008, *25:* 1253–6.
30. Guindon SP, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*

2003; *52:* 696–704.
31. Felsenstein J. PHYLIP – Phylogeny inference package (version 3.2). *Cladistics* 1989; *5:* 164–6.
32. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 2003; *19:* 1572–4.
33. Roehrig JT. Antigenic structure of flavivirus proteins. *Adv Vir Res* 2003, *59:* 141–75.
34. Falconar AK. Identification of an epitope on the dengue virus membrane (M) protein defined by cross-protective monoclonal antibodies: Design of an improved epitope sequence-based on common determinants present in both envelope (E and M) proteins. *Arch Virol* 1999; *144:* 2313–30.
35. Mathew A, Kurane I, Rothman AL, Zeng LL, Brinton MA, Ennis FA. Dominant recognition by human CD8+ cytotoxic T-lymphocytes of dengue virus nonstructural proteins NS3 and NS1.2a. *J Clin Invest* 1996; *98:* 1684–91.
36. Gagnon SJ, Zeng W, Kurane I, Ennis FA. Identification of two epitopes on the dengue 4 virus capsid protein recognized by a serotype-specific and a panel of serotype-cross-reactive human CD4+ cytotoxic T-lymphocyte clones. *J Virol* 1996; *70:* 141–7.
37. Jacobs MG, Robinson PJ, Bletchly C, Mackenzie JM, Young PR. Dengue virus nonstructural protein-1 is expressed in a glycosyl-phosphatidylinositol-linked form that is capable of signal transduction. *FASEB J* 2000; *14:* 1603–10.
38. Islam MA, Ahmed MU, Begum N, Chowdhury NA, Khan AH, Parquet MC, *et al*. Molecular characterization and clinical evaluation of dengue outbreak in 2002 in Bangladesh. *Jpn J Infect Dis* 2006; *59:* 85–91.
39. Fitch WM, Bush RM, Bender CA, Cox NJ. Long-term trends in the evolution of H(3) HA1 human influenza type A. *Proc Natl Acad Sci USA* 1997; *94:* 7712–8.
40. Kukreti H, Chaudhary A, Rautela RS, Anand R, Mittal V, Chhabra M, *et al*. Emergence of an independent lineage of dengue virus type 1 (DENV-1) and its co-circulation with predominant DENV-3 during the 2006 dengue fever outbreak in Delhi. *Int J Infect Dis* 2008; *12:* 542–9.
41. Bennett SN, Holmes EC, Chirivella M, Rodriguez DM, Beltran M, Vorndam V, *et al*. Selection-driven evolution of emergent dengue virus. *Mol Biol Evol* 2003; *20:* 1650–8.
42. Twiddy SS, Holmes EC, Rambaut A. Inferring the rate and time scale of dengue virus evolution. *Mol Biol Evol* 2003; *20*: 122–9.
43. Patil JA, Cherian S, Walimbe AM, Bhagat A, Vallentyne J, Kakade M, *et al*. Influence of evolutionary events on the Indian subcontinent on the phylogeography of dengue type 3 and 4 viruses. *Infect Genet Evol* 2012; *12*: 1759–69.
44. Morris AJ, Davenport CR, Tolan DR. A lysine to arginine substitution at position 146 of rabbit aldolase A changes the rate-determining step to Schiff base formation. *Protein Eng* 1996; *9:* 61–7.
45. Jaseja M, Copié V, Starkey J. Conformational studies of antimetastatic laminin-1 derived peptides in different solvent systems, using solution NMR spectroscopy. *J Pept Res* 2003; *69:* 24–39.
46. Woelk CH, Holmes EC. Reduced positive selection in vector-borne RNA viruses. *Mol Biol Evol* 2002; *19:* 2333–6.

Correspondence to: Dr Sonika Bhatnagar, Computational and Structural Biology Laboratory, Division of Biotechnology, Netaji Subhash Institute of Technology, Dwarka, New Delhi–110 078, India.
E-mail: ecc999@gmail.com

Received: 12 February 2013 *Accepted in revised form:* 10 May 2013