Genome sequence analysis of Tamana bat virus and its relationship with the genus *Flavivirus*

X. de Lamballerie,[1] S. Crochu,[1] F. Billoir,[1] J. Neyts,[2] P. de Micco,[1] E. C. Holmes[3] and E. A. Gould[4]

[1] Unité des Virus Emergents, EA3292-IFR48, Université de la Méditerranée, Faculté de Médecine, 27 Bd J. Moulin, F13005 Marseille, France

[2] Rega Institute for Medical Research, Minderbroedersstraat 10, K. U. Leuven, B-3000 Leuven, Belgium

[3] Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

[4] Centre for Ecology and Hydrology, Mansfield Road, Oxford OX1 3SR, UK

Tamana bat virus (TABV, isolated from the bat *Pteronotus parnellii*) is currently classified as a tentative species in the genus *Flavivirus*. We report here the determination and analysis of its complete coding sequence. Low but significant similarity scores between TABV and member-viruses of the genus *Flavivirus* were identified in the amino acid sequences of the structural, NS3 and NS5 genes. A series of cysteines located in the envelope protein and the most important enzymatic domains of the virus helicase/NTPase, methyltransferase and RNA-dependent RNA polymerase were found to be highly conserved. In the serine-protease domain, the catalytic sites were conserved, but variations in sequence were found in the putative substrate-binding sites, implying possible differences in the protease specificity. In accordance with this finding, the putative cleavage sites of the TABV polyprotein by the virus protease are substantially different from those of flaviviruses. The phylogenetic position of TABV could not be determined precisely, probably due to the extremely significant genetic divergence from other member-viruses of the family *Flaviviridae*. However, analysis based on both genetic distances and maximum-likelihood confirmed that TABV is more closely related to the flaviviruses than to the other genera. These findings have implications for the evolutionary history and taxonomic classification of the family as a whole: (i) the possibility that flaviviruses were derived from viruses infecting mammals rather than from mosquito viruses cannot be excluded; (ii) using the current criteria for the definition of genera in the family *Flaviviridae*, TABV should be assigned to a new genus.

Introduction

Tamana bat virus (TABV) was isolated in 1973 by J. L. Price in Trinidad from the salivary glands, saliva and spleen of the insectivorous bat *Pteronotus parnellii* (Price, 1978). Some characteristics of the virus (sensitivity to ether, pathogenicity for suckling mice and ability to haemagglutinate goose erythrocytes) were those of (enveloped) 'arboviruses', i.e. viruses currently classified as flaviviruses or alphaviruses, but no evidence for the existence of an arthropod vector could be found. Despite extensive investigations, no serological relationship with known arboviruses or other viruses could be detected and TABV remained unclassified. More recently, Kuno *et al.* (1998) published electron micrographs of TABV propagated in Vero cells that showed virus particles with the typical size and morphology of flaviviruses. However, a combination of PCR primers designed by these authors with the ability to amplify a portion of the NS5 gene of a large variety of flaviviruses yielded no amplicon when applied to TABV. Therefore, after more than 25 years and both serological and molecular investigations, the taxonomic position of TABV remains undetermined and quite intriguing. Here, we present the complete coding sequence of this hitherto unclassified virus and report our attempts to elucidate the genetic structure and evolutionary relationships of TABV. This analysis allowed us to identify TABV as a new and divergent member of the family *Flaviviridae* and had implications for the

Author for correspondence: X. de Lamballerie.
Fax +33 4 91 32 44 95. e-mail xndl-virophdm@gulliver.fr

The GenBank accession number for the complete sequence of the TABV ORF is AF285080.

0001-8388 © 2002 SGM

evolutionary history and taxonomic classification of the family as a whole.

Methods

Virus strain.

TABV (labelled Tr127154) was kindly supplied by Robert Shope (University of Texas, Galveston) at the seventh mouse brain passage (dated 8–9 December 1974). After an additional passage in newborn mice, it was propagated in Vero cells cultured as described previously (Billoir *et al.*, 2000). Cells and supernatant medium were recovered separately at 4–5 days post-infection.

Preparation of viral RNAs and cDNAs.

RNA was extracted from infected cells and from the supernatant medium by using the RNA-Now kit (Biogentex). RNAs were reverse-transcribed using random hexaprimers (Roche Molecular Biochemicals) and MuMLV Superscript reverse transcriptase (Gibco BRL) under standard conditions.

Genomic amplification and sequencing

Amplification of conserved regions. Primers were designed in conserved regions of the flavivirus genomes, using sequences from databases. The first set [TABV-NS3-S, 5′ BYiRTiGCiCCiACiMGiGTNGT 3′ (i, inosine); TABV-NS3-R, 5′-RTTiGCiCCCATYTCiSWDAT 3′; hybridization temperature 50 °C] was designed from the alignment of NS3 gene sequences. The second set (TABV-NS5-S, 5′ ATGGGiAARMRiGARAARAA 3′; TABV-NS5-R, GTRTCCCAiCCiGCiGTRTCRTC 3′; hybridization temperature 45 °C) was designed from NS5 sequences. PCR amplifications were achieved under standard conditions using *Taq* polymerase (Gibco-BRL) and cDNAs prepared from infected cells. Amplicons were gel-purified (Geneclean, Bio101) and ligated into the pGEM-T vector system I (Promega). Recombinant plasmids were transfected into *E. coli* XL-Blue cells and sequenced using the M13 universal primers, the D-Rhodamine DNA sequencing kit and an ABI Prism 377 sequence analyser (Perkin Elmer).

Determination of the complete coding sequence. The complete coding sequence of the virus was determined by using cDNAs prepared from the supernatant medium and the anchored PCR method, with one specific primer designed from a previously characterized virus sequence and a combination of non-specific oligonucleotides, as described previously (Billoir *et al.*, 2000). PCR products were cloned and sequenced as described above. Specific primers were designed from the reconstructed coding sequence and used to generate 15 overlapping PCR products covering the entire ORF. These products were sequenced directly on both strands using the amplification primers.

Sequence analysis

Local databases. Nucleotide and amino acid sequences of complete flavivirus ORFs were obtained from GenBank. Abbreviations are those recommended in the 7th report of the International Committee on Taxonomy of Viruses (Heinz *et al.*, 2000) and accession numbers are the same as reported previously (Billoir *et al.*, 2000) except for MVEV (AF161266) and MODV (AJ242984).

Sequences from pestiviruses [*Border disease virus* (BDV), GenBank accession no. U70263; *Bovine viral diarrhoea virus 1* (BVDV-1), M31182; *Bovine viral diarrhoea virus 2* (BVDV-2), U18059; *Classical swine fever virus* (CSFV), M31768], hepaciviruses [*Hepatitis C virus* subtype 1a (HCV-1a), M62321; HCV-1b, D90208; HCV-1c, D14853; HCV-2a, D00944; HCV-2b, D01221; HCV-2c, D50409; HCV-3a, D17763; HCV-4a, Y11604; HCV-5a, Y13184; HCV-6a, Y12083; HCV-11a, D63822; *GB virus A* [GBV-A] nig, U22303; GBVA lab, U94421; GBV-A cal, AF023424; GBV-A tri, AF023425; GBV-C human, AB003292; GBV-C tro, AF070476], potyviruses [family *Potyviridae*: *Soybean mosaic virus*, NC 002634; *Pea seed-borne mosaic virus*, AJ252242; *Pepper mottle virus*, NC 001517; *Potato virus A*, NC 001649; *Sweet potato feathery mottle virus*, NC 001841; *Tobacco etch virus*, NC 001555; *Peanut mottle virus*, NC 002600; *Plum pox virus*, NC 001445; *Turnip mosaic virus*, NC 002509; Japanese yam mosaic virus, NC 000947; Ryegrass mosaic virus, NC 001814] and carmoviruses [family *Tombusviridae*: *Carnation mottle virus*, NC 001265; *Galinsoga mosaic virus*, NC 001818; *Hibiscus chlorotic ringspot virus*, X86448; Japanese iris virus, NC 002187; *Melon necrotic virus*, NC 001504; *Saguaro cactus virus*, NC 001780] were also obtained from GenBank and used, in addition to the flavivirus sequences, to build local nucleotide and amino acid sequence databases in the DNATools platform (version 5.2.014; S. W. Rasmussen, Carlsberg Institute, Copenhagen).

Alignments. A search for significant similarity between TABV sequences and sequences from GenBank was performed using the Local MDB (multiple database) BLAST program implemented in DNATools. This program was kindly written at our request by S. W. Rasmussen. It allows iterative BLAST searches to be performed against a series of databases, each made of a single sequence. This is useful for the detection of similarity between distantly related sequences and the delimitation of homologous regions. BLASTP (protein query–protein database comparison) and BLASTX (nucleotide sequence query–protein database comparison) algorithms were used.

A search for conserved amino acid domains within the polyprotein of TABV was performed using the program HMMPFAM implemented in the UK Human Genome Mapping Project computing platform (http://www.hgmp.mrc.ac.uk/).

Pairwise and multiple alignments of partial or complete amino acid sequences were generated by the program CLUSTAL W version 1.74 (Thompson *et al.*, 1994). Conserved motifs were used as a control of validity for alignments as reported previously (Billoir *et al.*, 2000).

Phylogenetic analysis. Due to large genetic distances and the presence of regions without significant sequence similarity, it proved difficult to include TABV in a phylogenetic analysis of complete flavivirus polyprotein sequences. However, in specific regions of the polyprotein, MDB-BLASTP identified significant similarity scores between TABV and other viruses (see details in Results). Partial homologous sequences (in the structural, NS3 and NS5 genes) were used to generate relevant amino acid sequence alignments with CLUSTAL W. Genetic distances between sequences were estimated with the program MEGA (version 2.0; Kumar *et al.*, 2001) using the gamma-distance statistic. The shape parameter α, describing the extent of among-site substitution rate variation, was estimated from the data by using the program PAML (Yang, 1997). Trees were constructed on these distance matrices by using the neighbour-joining method.

A maximum-likelihood (ML) analysis of the helicase and NS3 amino acid sequence alignments was also used to determine the evolutionary position of TABV. First, an initial maximum-parsimony tree for all sequences from both genes was estimated by using a heuristic search algorithm (program PAUP* version 4.0; Swofford, 2000). Next, starting from this initial tree, four model trees were constructed using the program TREEVIEW (Page, 1996) that differed in the placement of the TABV lineage (see Fig. 5). In tree 1, TABV was placed as a sister-group to a clade containing the genus *Flavivirus* and CFAV. In tree 2, the positions of TABV and CFAV were reversed, with TABV now more closely related to the genus *Flavivirus*. In tree 3, TABV and CFAV grouped together and then joined the genus *Flavivirus*. Finally, in a more extreme revision, TABV was positioned next to the pestiviruses. All other branches on the phylogenies were unchanged. The likelihood of these four model trees was estimated by using an ML method, assuming that amino acid positions changed according to the Jones–Taylor–Thornton substitution matrix but allowing rates of amino acid substitution to vary along the sequence alignment according to a gamma distribution with shape parameter α estimated from the data. This analysis was also performed by using the PAML package (Yang, 1997).

Hydropathy plots. Hydropathy plots of structural proteins or complete polyproteins were produced in Microsoft Excel using the amino acid hydropathy values determined by Kyte & Doolittle (1982), considering sliding windows from 11 to 25 amino acids. For ease of comparison between the hydropathy profiles exhibited by TABV and Kunjin virus (KUNV), aligned sequences were exported with all alignment-generated gaps maintained (with a hydropathy value of zero). This permitted the comparison of hydropathy profiles of amino acid sequences of unequal lengths.

Base composition and codon usage. The G + C content of the TABV genome was determined and compared with that of other flaviviruses by using the program CODON W (version 1.3). Among flaviviruses, the influence of the G + C content on the amino acid composition of polyproteins, codon usage and the length of ORFs was investigated using the same program.

Results

Analysis of PCR products in the NS3 and NS5 genes

Amplification products obtained using primer sets TABV-NS3-S/TABV-NS3-R and TABV-NS5-S/TABV-NS5-R were respectively 585 and 273 nt long. The corresponding sequences were tested using MDB-BLASTX against complete amino acid sequences of members of the family *Flaviviridae*. Significant identity scores were found with homologous sequences of flaviviruses in both the NS3 [best score with *Dengue virus* type 1 (DENV-1), $P = 2e-13$] and NS5 [best score with *Yellow fever virus* (YFV), $P = 7e-13$] regions, suggesting that TABV is genetically more closely related to the flaviviruses than it is to viruses in the other genera.

Analysis of the complete ORF sequence

The complete TABV ORF sequence (GenBank accession no. AF285080) was 10053 nt long (including the initial ATG and the terminal stop codon) and encoded a 3350 aa polyprotein. This is shorter than any of the flavivirus polyproteins described to date. Perhaps significantly, the polyproteins of non-vectored viruses [*Rio Bravo virus* (RBV), 3379 aa; *Modoc virus* (MODV), 3374 aa; *Apoi virus* (APOIV), 3371 aa; Cell fusing agent virus (CFAV), 3341 aa] are shorter than those of arboviruses, which range between 3386 (DENV-4) and 3415 [*Powassan virus* (POWV)] aa.

A comparison was made of the TABV polyprotein sequence with those deposited in databases using the program HMMPFAM. The top-scoring sequence families were the flavivirus RNA-directed RNA polymerase (RdRp) (E value $4{\cdot}1e-45$), the flavivirus helicase (E value $4{\cdot}3e-20$) and the flavivirus envelope glycoprotein (E value $9{\cdot}3e-10$; see Fig. 2e). The relatedness to the envelope protein of flaviviruses is worthy of emphasis, since sequence similarity between members of different genera in the family *Flaviviridae* has been observed in some non-structural genes, but never in structural genes. This is a persuasive argument for grouping TABV in the genus *Flavivirus*. Accordingly, further investigations were carried out to test the hypothesis that TABV is related most closely to the flaviviruses.

Table 1. Proposed cleavage sites in the TABV polyprotein

Possible cleavage sites were identified from alignment with flavivirus polyproteins. The type of protease that cleaves the flavivirus polyprotein is indicated for each putative site. The one-letter amino acid code is used. Abbreviations: VSP, viral serine protease; HS, host signalase; ?, unknown protease; VirC, mature virion C protein; CTHD, C-terminal hydrophobic domain; AnchC, anchored C protein (mature virion C protein + CTHD); NI, not identified.

Cleavage site	Protease	Amino acid sequence
VirC/CTHD	VSP	QKRQK/SSGGY
AnchC/prM*	HS	MVIFC/GYQSG
Pr/M	Furin	HRTRR/SVTET
M/E*	HS	ILVIA/QFYLAD
E/NS1	HS	EVVAA/DKYVL
NS1/NS2A	?	NVVKA/SKMNK
NS2A/NS2B	VSP	NI
NS2B/NS3	VSP	NLRDK/SKGLI
NS3/NS4A	VSP	PLVQR/VFSGI
NS4A/2K	VSP	GITQR/EKSTG
2K/NS4B	HS	YYILA/DGEIL
NS4B/NS5	VSP	KTTQR/FRSSI

* Other possible sites are discussed in the text.

Study of structural genes

In the family *Flaviviridae*, structural proteins are derived following processing of the N-terminal part of the polyprotein. The 800 N-terminal amino acids of the TABV polyprotein were compared with the available *Flaviviridae* sequences using MDB-BLASTP. No match was found with proteins of hepaciviruses or pestiviruses. In contrast, significant identity scores were observed with structural proteins of flaviviruses [best score with KUNV, $P = 2e-28$, 185/826 (22 %) identity]. These results were used as markers for the alignment of flavivirus and TABV sequences using CLUSTAL W. The structural proteins of TABV were characterized by comparing sequence alignments, hydropathy plots and amino acid patterns.

(i) VirC/C-terminal hydrophobic domain (CTHD) cleavage site. The mature capsid protein (VirC) of flaviviruses is a small, highly basic protein that is cleaved from the nascent capsid protein (AnchC) by the viral serine protease (VSP) after a dibasic amino acid sequence and before a CTHD. Sequence alignment suggests that the cleavage site for TABV is located after the amino acid at position 95. The proposed residues are Gln–Lys, a pattern never reported for flaviviruses (Table 1). However Gln at position -2 and Lys at position -1 are seen

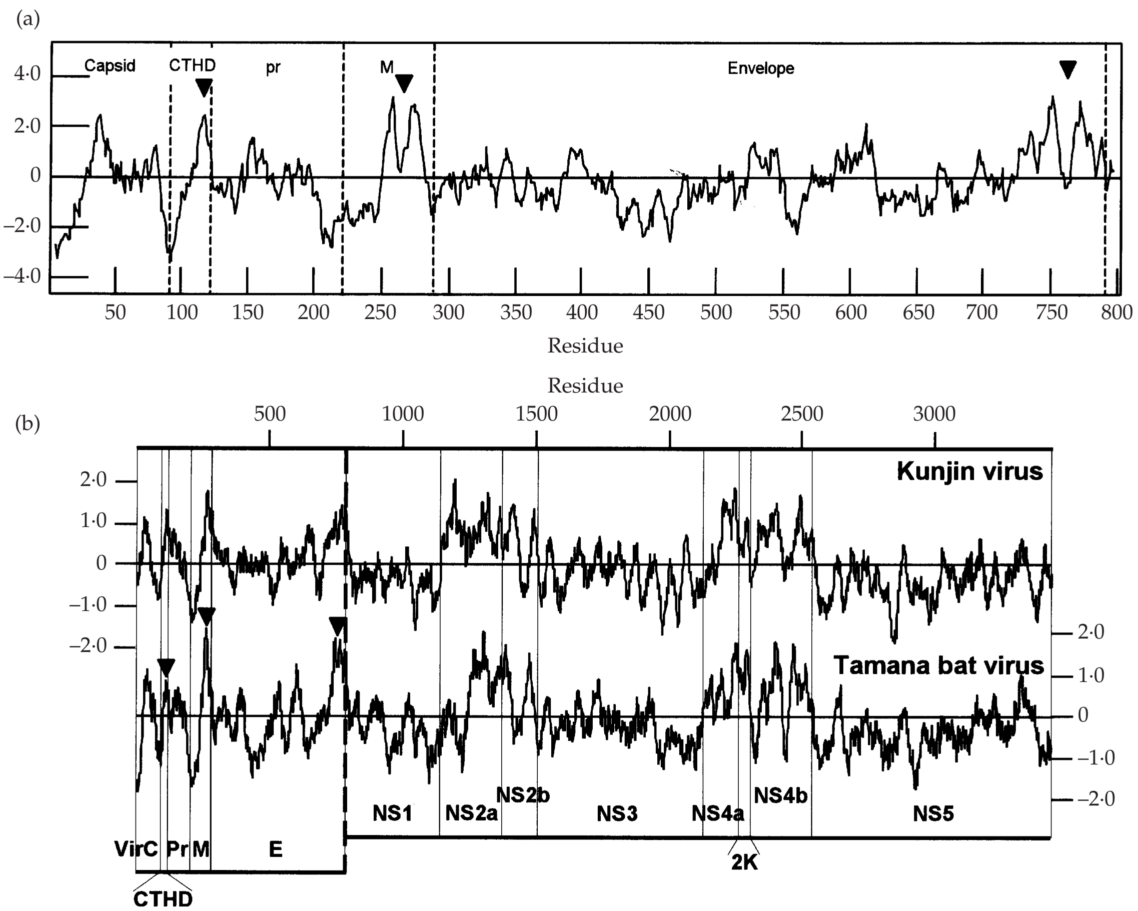


Fig. 1. Hydropathy plots. (a) Structural part of the polyprotein of TABV (sliding window, 11 aa; increment, 1 aa). (b) Comparison of the hydropathy profiles of the complete polyproteins of TABV and KUNV (sliding window, 25 aa; increment, 1 aa). Pointers (▼) show hydrophobic signal sequences that possibly act for the translocation of the pr, E and NS1 proteins in the lumen of the ER.

frequently in flavivirus cleavage sites. The amino acid content of VirC (rich in basic Lys and Arg residues) fits with a putative nucleoprotein function.

(ii) AnchC/prM cleavage site. The CTHD sequence of flaviviruses acts as a signal sequence for translocation of prM into the lumen of the ER. It is cleaved from the rest of the polyprotein by a host signalase (HS). In the case of TABV, the cleavage could occur after Cys^{121}, a situation comparable to that of CFAV, and consistent with the (−3, −1) rule of von Heijne (1984) and the presence of an upstream hydrophobic sequence (Table 1); alternatively, the cleavage could occur after Gly^{115} or Gln^{124}, which are also canonical sites for HS. The hydrophobic structure CTHD can be identified in Fig. 1 (a), which shows a detailed hydropathy plot of TABV structural proteins. The hydropathy profiles of the capsid proteins of TABV and KUNV are remarkably similar (Fig. 1b).

(iii) pr/M cleavage site. The prM protein of flaviviruses is the glycosylated precursor of the structural M protein. The prM cleavage is mediated by the host furin or an enzyme of similar specificity (Stadler *et al.*, 1997; Steiner *et al.*, 1992) and occurs at a site that conforms to the Arg–X–Arg/Lys–Arg pattern at positions −4 to −1 (Rice, 1996). This site can be identified after position 224 in the TABV polyprotein (Table 1). The deduced pr protein is 97 aa long, predominantly hydrophilic (Fig. 1a) and possesses two possible *N*-glycosylation sites and four Cys residues, of which two are conserved in flaviviruses.

(iv) M/Envelope cleavage site. This HS cleavage site might be situated after Ala^{277} (Table 1) or alternatively after Ala^{284} (but with a hydrophilic Gln residue at position −7). The M protein comprises a 32 aa ectodomain followed by two potential hydrophobic membrane-spanning domains (Fig. 1a), possibly acting as signal sequences for translocation of the E protein in the lumen of the ER (as reported for flaviviruses).

The prM hydropathy profile is very similar for TABV and KUNV. Hydrophobic membrane-spanning domains can be identified for both viruses (Fig. 1b). However, the suggested lengths of the pr and M proteins of TABV are respectively longer and shorter than those reported for flaviviruses.

(v) Envelope/NS1 cleavage site. The envelope protein of flaviviruses is cleaved from the non-structural part of the polyprotein by an HS. Alignments suggest position 786 as a possible site for TABV (Table 1), consistent with the rule of von Heijne and the presence of an upstream hydrophobic sequence. The deduced E protein consists of a long ectodomain followed by a C-terminal membrane anchor that might be implicated in translocation of the NS1 protein in the lumen of the ER (by reference to flaviviruses) and can be identified in hydropathy plots (Fig. 1). It contains no putative glycosylation site, but 16 Cys residues, of which 15 are in the ectodomain and 10 of these are at positions conserved in flaviviruses, suggesting similar folding of the molecule through disulphide bonds. Comparison of the TABV E protein with that of tick-borne viruses (Mandl *et al*., 1989) suggests that disulphide bonds could exist between cysteines 287 and 313, 356 and 387, 374 and 398 and 460 and 579 in domain A and 596 and 627 in domain B. A sequence homologous to the 'fusion peptide', a 14 aa motif thought to be involved in fusion (Roehrig *et al*., 1989), is present at positions 380–393. It conforms with the DRGWXX(G/H)CXXFGKG motif observed for all flaviviruses other than CFAV.

Study of non-structural genes – (i) NS3

The NS3 protein of flaviviruses is hydrophilic and is believed to be at least bifunctional. The N-terminal sequence contains four regions (boxes 1–4) that have significant similarity to serine proteases belonging to the trypsin superfamily (Bazan & Fletterick, 1989; Gorbalenya *et al*., 1989a). This protease activity was shown to be essential for the polyprotein processing of YFV, *West Nile virus* (WNV), *Murray Valley encephalitis virus* (MVEV), *Tick-borne encephalitis virus* (TBEV) and DENV-2 (Chambers *et al*., 1990; Wengler *et al*., 1991; Lobigs, 1992; Pugachev *et al*., 1993; Valle & Falgout, 1998; Zhang *et al*., 1992) and requires both the proteinase domain of NS3 and the NS2B protein (Arias *et al*., 1993; Chambers *et al*., 1990; Falgout *et al*., 1991). The C-terminal domain of the NS3 protein contains significant regions of similarity to the DEAD family of RNA helicases (Gorbalenya *et al*., 1989b) in seven conserved segments designated motifs I, IA, and II–VI. RNA-stimulated NTPase and RNA triphosphatase activities have been demonstrated for NS3 (Wengler & Wengler, 1993; Warrener *et al*., 1993), without identification of catalytic/substrate-binding residues.

NS2B/NS3 cleavage site. Based on sequence alignments, this site was proposed at position 1477, after the Asp–Lys pair (Table 1). The presence of the Asp residue at position −2 is surprising for a site that is supposed to be cleaved by a VSP, but it should be noted that the proposed homologous cleavage site of CFAV has an Asn residue at position −2.

NS3/NS4A cleavage site. This cleavage (also mediated by the VSP) may occur at position 2089, after the Gln–Arg pair (Table 1). The presence of a Val residue at position +1 (unusual after VSP dibasic sites in flaviviruses) is also observed after the proposed NS3/NS4A cleavage site of CFAV.

Using the NS3 gene sequence as defined above (aa 1478–2089) and MDB-BLAST, high matching scores were found with the NS3 sequences of all flaviviruses, including CFAV [best score with DENV-1, $P = 3e-46$, 160/527 (30%) identity], in both the serine protease and helicase/NTPase motifs (Fig. 2a, b). With regards to the protease, conserved motifs could easily be identified in boxes 1, 2 and 3 and, in particular, the three catalytic residues His, Asp and Ser are conserved (Fig. 2a). However, residue 1601, supposed to be a substrate-binding residue, is not Asp (acidic, as reported for all flaviviruses including CFAV), but Lys (basic). Moreover, the sequence similarity in box 4, which contains four additional substrate-binding residues, is very low. These important variations in sequence could possibly imply significant differences in the biological properties of the enzyme.

Sequence identity between TABV, CFAV and flaviviruses was also observed in all seven motifs of the helicase domain (Fig. 2b). In the DEXD pattern, X is Ala for all flaviviruses, Cys for CFAV and Ser for TABV. All three residues are amino acids with short side chains. In motif III, two different alignments can be proposed (Fig. 2b).

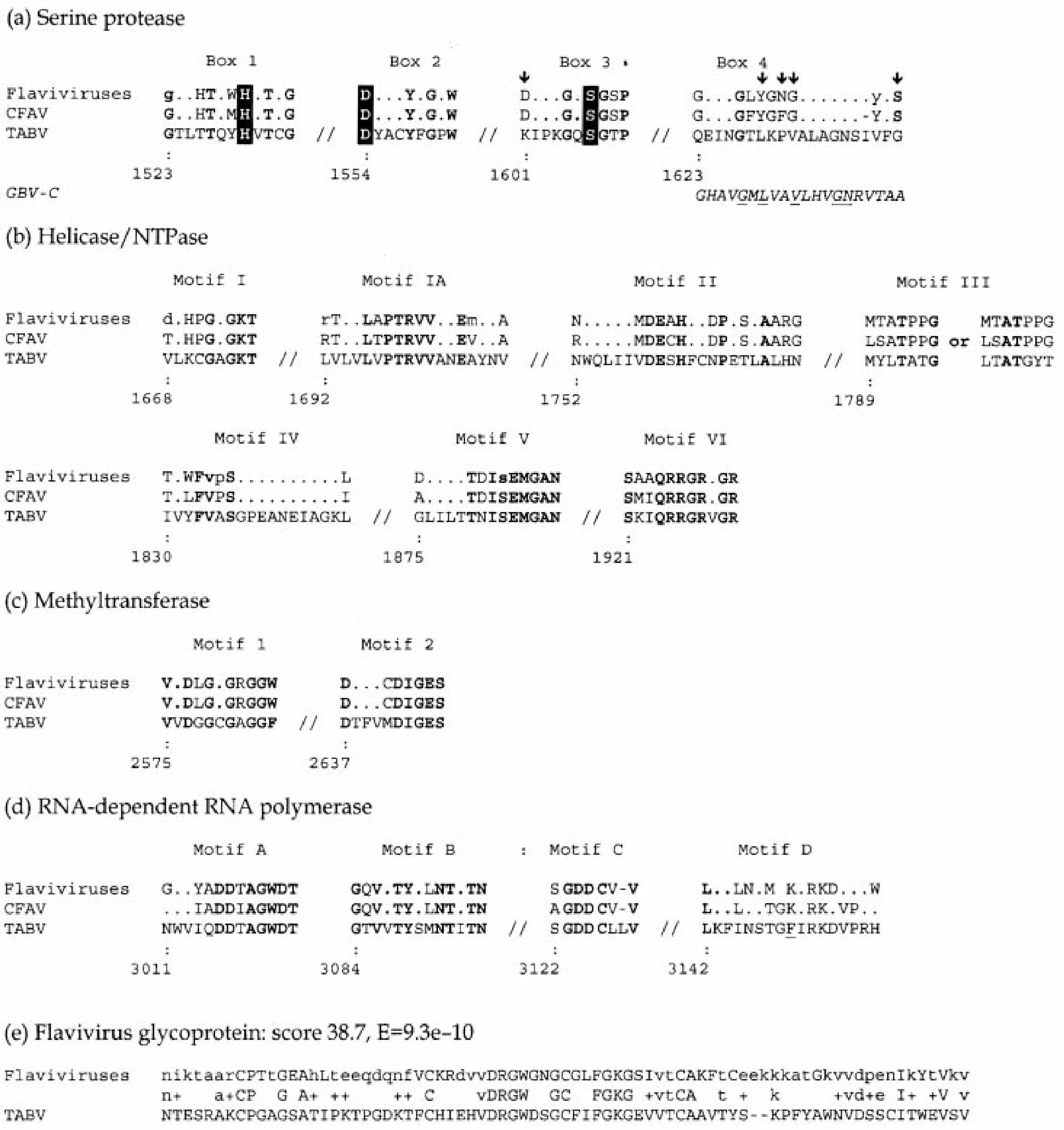
Significant but lower scores were also observed with the NS3 gene sequences of hepaciviruses [GBV-C, $P = 1e-9$, 71/294 (24%) identity; GBV-B, $P = 8e-8$, 106/457 (23%) identity; GBV-A, $P = 4e-5$, 66/298 (22%) identity; HCV, $P = 2e-5$, 69/297 (23%) identity] and pestiviruses [CSFV sequence, $P = 3e-17$, 108/479 (22%) identity]. For hepaciviruses, the best matches were found in the C-terminal part of the protein for motifs I, Ia, II, III, IV and VI of the helicase. In motif III, the Thr–Ala–Thr triad is conserved, corresponding to the second alignment proposed for this motif (Fig. 2b). Scores were lower for the protease domain, but His^{1530} (box 1), Asp^{1554} (box 2) and the GXSGXP motif (box 3) were conserved for all viruses. Therefore, all catalytic residues are common to hepaciviruses and TABV. Interestingly, the sequence corresponding to box 4 of the protease matched a homologous sequence in the GBV-C polyprotein (Fig. 2a).

For pestiviruses, all catalytic residues of the protease were conserved, but no strong similarity was found in box 4. In the helicase domain, conserved patterns were present in all motifs. In motif III, the Thr–Ala–Thr triad was conserved.

As reported previously for flaviviruses (Lain *et al*., 1989), significant identity scores were found with the CI sequence of potyviruses [best score with the sequence from *Soybean mosaic virus*, $P = 6e-13$, 89/358 (24%) identity]. These homologies were identified only in the helicase motifs I, Ia, II, IV, V and VI.

(ii) NS5

The NS5 protein of flaviviruses is a long, hydrophilic and basic protein that exhibits RdRp activity (Tan *et al*., 1996; Steffens *et al*., 1999). Four motifs (A–D) possess residues conserved in all virus RdRps (Poch *et al*., 1989) and, in particular, motif C (a core motif for catalytic RdRp activity), which includes the Gly–Asp–Asp conserved pattern. In the N-



(a) Serine protease

```
                      Box 1             Box 2             Box 3 •          Box 4
                                                        ↓                    ↓ ↓↓          ↓
Flaviviruses    g..HT.WH.T.G         D...Y.G.W         D...G.SGSP      G...GLYGNG.......y.S
CFAV            G..HT.MH.T.G         D...Y.G.W         D...G.SGSP      G...GFYGFG......-Y.S
TABV            GTLTTQYHVTCG   //    DYACYFGPW   //    KIPKGQSGTP  //  QEINGTLKPVALAGNSIVFG
                :                    :                 :               :
             1523                 1554              1601            1623
GBV-C                                                                  GHAVGMLVAVLHVGNRVTAA
```

(b) Helicase/NTPase

```
                    Motif I          Motif IA                 Motif II                    Motif III

Flaviviruses    d.HPG.GKT       rT..LAPTRVV..Em..A      N.....MDEAH..DP.S.AARG      MTATPPG    MTATPPG
CFAV            T.HPG.GKT       RT..LTPTRVV..EV..A      R.....MDECH..DP.S.AARG      LSATPPG or LSATPPG
TABV            VLKCGAGKT  //   LVLVLVPTRVVANEAYNV  //  NWQLIIVDESHFCNPETLALHN  //  MYLTATG    LTATGYT
                :               :                       :                           :
             1668            1692                    1752                        1789

                      Motif IV            Motif V          Motif VI

Flaviviruses    T.WFvpS..........L      D....TDIsEMGAN      SAAQRRGR.GR
CFAV            T.LFVPS..........I      A....TDISEMGAN      SMIQRRGR.GR
TABV            IVYFVASGPEANEIAGKL  //  GLILTTNISEMGAN  //  SKIQRRGRVGR
                :                       :                   :
             1830                    1875                1921
```

(c) Methyltransferase

```
                     Motif 1       Motif 2

Flaviviruses    V.DLG.GRGGW       D...CDIGES
CFAV            V.DLG.GRGGW       D...CDIGES
TABV            VVDGGCGAGGF  //   DTFVMDIGES
                :                 :
             2575              2637
```

(d) RNA-dependent RNA polymerase

```
                     Motif A          Motif B       :  Motif C          Motif D

Flaviviruses    G..YADDTAGWDT     GQV.TY.LNT.TN        SGDDCV-V      L..LN.M K.RKD...W
CFAV            ...IADDIAGWDT     GQV.TY.LNT.TN        AGDDCV-V      L..L..TGK.RK.VP..
TABV            NWVIQDDTAGWDT     GTVVTYSMNTITN   //   SGDDCLLV  //  LKFINSTGFIRKDVPRH
                :                 :                    :             :
             3011              3084                 3122          3142
```

(e) Flavivirus glycoprotein: score 38.7, E=9.3e–10

```
Flaviviruses    niktaarCPTtGEAhLteeqdqnfVCKRdvvDRGWGNGCGLFGKGSIvtCAKFtCeekkkatGkvvdpenIkYtVkv
                n+    a+CP  G A+ ++     ++ C     vDRGW  GC  FGKG +vtCA  t +   k      +vd+e I+ +V v
TABV            NTESRAKCPGAGSATIPKTPGDKTFCHIEHVDRGWDSGCFIFGKGEVVTCAAVTYS--KPFYAWNVDSSCITWEVSV
```

Fig. 2. Conserved motifs in the polyprotein of TABV. (a)–(d) Conserved enzymatic motifs in the proteins encoded by the NS3 and NS5 genes. Sequence alignments include amino acids completely (capitals) or nearly completely (lower-case) conserved among flaviviruses (Flaviviruses), the CFAV sequence (only residues common to flaviviruses or TABV are indicated) and the newly characterized TABV. This is in accordance with the current classification, in which CFAV and TABV are listed as tentative species in the genus *Flavivirus*. Residues are numbered by reference to the TABV sequence; dashes represent gaps, residues in bold are conserved positions and dots represent non-conserved amino acids. In (a), the sequence of GBV-C corresponding to box 4 is shown (see comments in text); arrows indicate putative substrate binding sites; white letters on black correspond to catalytic sites. (e) Conserved amino acid motifs detected in the structural part of the polyprotein of TABV using the program HMMPFAM.

terminal domain of NS5, two conserved motifs (1–2) are homologous to methyltransferases and might be implicated in *S*-adenosyl methionine binding (Koonin, 1993).

NS4B/NS5 VSP cleavage site. Based on sequence alignments, this site might be located at position 2495 of the TABV polyprotein (Table 1), following the Gln–Arg pair (with the unusual Phe residue at position +1). Using the deduced NS5 sequence of TABV, high matching scores were found with the NS5 protein of all flaviviruses, including CFAV [best score with KUNV, $P = 5e-93$, 260/884 (29%) identity]. Conserved patterns were identified in both the methyltransferase and RdRp domains (Fig. 2c–d). As observed previously (Poch *et al.*, 1989), the weakest identity scores were found in motif D of the RdRp. The Lys residue conserved in a large number of virus RdRps was found to be Phe in the TABV sequence. Interestingly, in the region homologous to the 37 aa interdomain of DENV-2, the Thr supposed to be the substrate of the CK2 Ser/Thr kinase and to be implicated in the nuclear localization of the NS5 (Forwood *et al.*, 1999) was conserved.

Significant scores were also observed with NS5 gene sequences of pestiviruses [CSFV, $P = 3e-10$, 83/356 (23 %) identity] and pestiviruses [GBV-A, $P = 0{\cdot}001$, 29/120 (24 %) identity; GBV-C, $P = 0{\cdot}071$, 14/49 (28 %) identity]. In these cases, significant identity scores were found only for the RdRp domain motifs A, B and C. In the case of HCV, a low matching score was found [$P = 0{\cdot}1$, 8/19 (42 %)] in the first motif of the methyltransferase domain.

Low scores were also observed with the RdRp motifs B and C in the polymerase sequence of carmoviruses [best score with *Galinsoga mosaic virus*, $P = 0{\cdot}085$, 22/93 (23 %) identity] and potyviruses [best score with *Tobacco etch virus*, $P = 0{\cdot}011$, 17/58 (29 %) identity]. Interestingly, the Phe at position 3150 in RdRp motif D of TABV was conserved in some carmoviruses.

(iii) NS1, NS2 and NS4

Using sequence alignments, attempts were made to identify the cleavage sites of these TABV proteins, by reference to those described for flaviviruses (Table 1).

The NS1/NS2A cleavage site is proposed at position 1130. In common with flaviviruses, it satisfies the (−3, −1) rule, but not the requirement for an upstream hydrophobic sequence (Rice & Strauss, 1990). The NS2A/NS2B cleavage site could not be identified. The NS4a/2K cleavage site (cleaved by the VSP in flaviviruses) is proposed for TABV at position 2229. The residues Gln and Arg (positions −2 and −1) are found in all flaviviruses, but the Glu residue at position +1 is unusual. The 2K/NS4B cleavage site, which may be cleaved by an HS, is proposed at position 2254 (consistent with the rule of von Heijne and the presence of an upstream hydrophobic sequence).

Using the NS1, NS2 and NS4 sequences of TABV and MDB-BLASTP, no significant match with *Flaviviridae* sequences was observed. In particular, it is notable that NS1 does not contain the very conserved series of cysteines found in all flaviviruses.

Hydropathy plots

The relationship of TABV with flaviviruses was further investigated by producing and comparing hydropathy plots of complete polyproteins. A comparison of the hydropathy profiles exhibited by TABV and KUNV polyproteins is presented in Fig. 1 (b). It shows striking similarities in both the structural and non-structural parts of the polyproteins. Such similarities exist not only in the genes in which significant identity scores were observed (see the profiles in the VirC, CTHD, Pr, M, NS3 and NS5 regions), but also in the NS2b, NS4a, 2K and NS4b regions, where no significant sequence identity was identified. This is extremely suggestive of shared biological properties and supports the evidence of a close evolutionary relationship between TABV and flaviviruses.

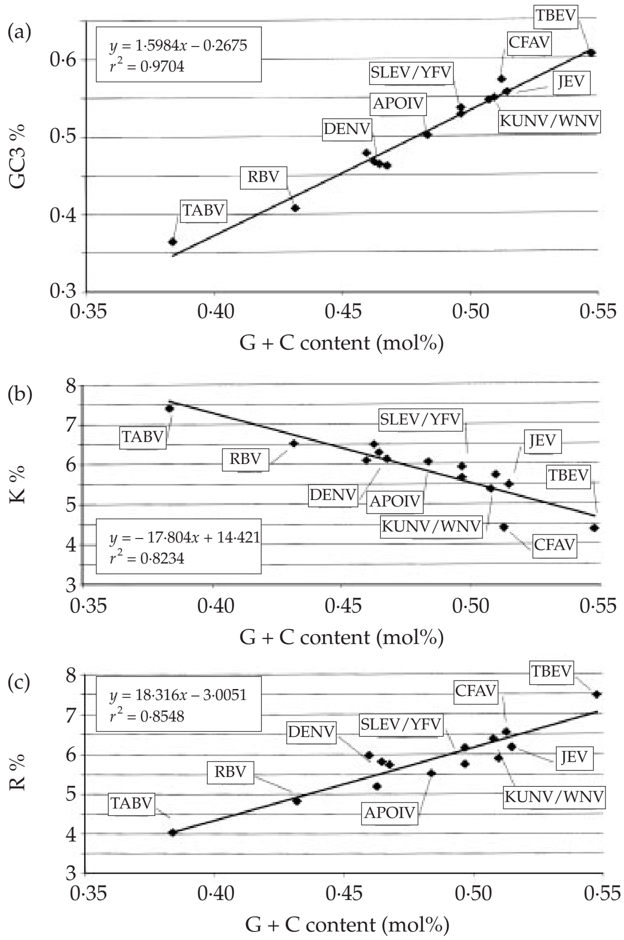


Fig. 3. Analysis of G+C content. (a) Relationship between global G+C content and G+C content at the third position of the codon (GC3 %) among viruses of the genus *Flavivirus*. (b)–(c) Relationship between global G+C content and the contents of lysine (K %) (b) and arginine (R%) (c) of the polyproteins among viruses of the genus *Flavivirus*. Abbreviations not already defined in the text are JEV (*Japanese encephalitis virus*) and SLEV (*St Louis encephalitis virus*).

Base composition and codon usage

A study of the base composition of the coding sequence of TABV showed that its G+C content was 38·4 mol %, a value that is lower than those observed for flaviviruses described to date (which are all above 43 mol %). The lowest value among flaviviruses is that of RBV (43·2 mol %), another virus isolated from bats. Previous studies of dsDNA genomes (Bellgard & Gojobori, 1999) have shown the existence of a linear relationship between G+C content at the third position (GC3 %) of the codons and the G+C content of all codon positions. This also appears to be true for flaviviruses, as shown in Fig. 3 (a), which reports the G+C contents and

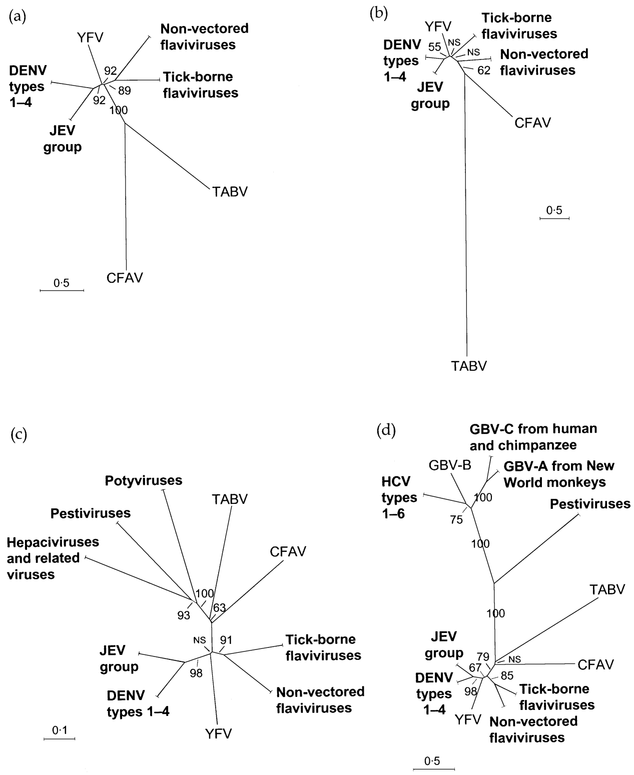


Fig. 4. Phylogenetic trees of the structural genes (a), the NS5 region (b), the helicase region (c) and the NS3 region (d) were produced from amino acid alignments between residues Cys^{176} and Cys^{460} (a) or the complete NS5 protein (b), the helicase region (c) or the protease and helicase regions (d) by using the neighbour-joining method. Genetic distances were calculated by the gamma-distance algorithm with the shape parameter α calculated from the data [$\alpha = 1{\cdot}3$ (a), 0·5 (b) or 1 (c, d)]. Bootstrap values are indicated; NS, not significant (< 50%).

GC3% values of representative members of this genus, including TABV.

The influence of this low G+C content on the properties of the virus polyprotein was investigated. It was expected that, as reported previously in dsDNA genomes (Nishizawa & Nishizawa, 1998), the low G+C content might be associated with an increase in Lys residues (encoded by AAA, AAG) and a decrease in Arg residues (CGN, AGA, AGG). This was verified when a comparison was made with other flaviviruses (Fig. 3b, c). The Arg content increased and the Lys content decreased as a linear function of the G+C content.

A low G+C content might also affect the codon usage of the virus. The percentage of amino acid residues encoded as predicted by random codon usage is low for TABV (79%) in comparison with flaviviruses and especially with CFAV (91%). However, this difference is not totally attributable to the low G+C content of TABV: after correction of the bias due to the G+C content (taking into account the frequency of each nucleotide in the complete sequence), only 62% of the Arg residues of TABV are encoded as expected statistically (versus 92% for CFAV). This suggests that, at least in the case of Arg residues, constraints other than the G+C content are implicated in the codon usage.

Phylogenetic analysis

In the structural genes, as noticed previously, alignments that included CFAV were difficult to produce (Cammisa-Parks *et al.*, 1992). The only region where MDB-BLASTP detected significant matches between TABV sequences and both CFAV and flaviviruses was located between Cys^{176} and Cys^{460}. A multiple alignment was produced using CLUSTAL W (with low alignment scores) and pairwise distances were calculated. In that region, CFAV was the most divergent virus, but relevant evolutionary information cannot easily be inferred from the study of genetic distances > 80% found between CFAV and TABV and also between CFAV and other flaviviruses. A phylogenetic tree for CFAV and representatives of the genus *Flavivirus* is presented in Fig. 4(a). TABV does not cluster with any of the recognized clades in the genus. It forms a new group, distantly related to both CFAV and flaviviruses. An NS3-like topology (Billoir *et al.*, 2000) is observed, with non-vectored flaviviruses in the same evolutionary group as tick-borne viruses (as observed previously in phylogenetic trees constructed from the NS3 sequences of flaviviruses). In the absence of a clear outgroup, this tree is unrooted for this group of viruses.

In the NS5 region, it proved impossible to produce significant alignments based on virus sequences from different genera. Therefore, complete NS5 sequences of flaviviruses, CFAV and TABV were aligned (Fig. 4b). Not unexpectedly, the unrooted tree obtained displayed the NS5-like topology described previously (Billoir *et al.*, 2000), with low bootstrap values.

In the NS3 region, MDB-BLASTP predicted no significant matching scores between TABV, potyviruses and hepaciviruses in the region encompassing the first 200 amino acids of the N terminus. This region includes the most important sites of the protease domain. Therefore, alignments including TABV, flavivirus, pestivirus, hepacivirus and potyvirus sequences were produced between Gly^{1670} and Arg^{1931}, in a region that includes the helicase motifs. The phylogenetic tree constructed from these data (Fig. 4c) displays an NS3-like topology (Billoir *et al.*, 2000) in the branches representing the flavivirus group. According to this tree, flaviviruses and CFAV have a common ancestor distinct from TABV, but the topology at the deepest nodes proved to be unstable when other methods were used for distance calculation or tree building. A similar analysis was then carried out using both the protease and helicase domains of TABV, flavivirus, pestivirus and hepacivirus sequences. The region studied extended from position 1523 of the TABV polyprotein (box 1 of the protease domain) to position 1931 (motif VI of the helicase domain). The topology observed (Fig. 4d) in the flavivirus lineage was the same as that observed with helicase sequences alone.

Table 3. ML analysis of the phylogenetic position of TABV

Model trees 1–4 are shown in Fig. 5. Tree parameters: −ln *L*, log likelihood; *δ*, difference in log likelihood from ML tree; *α*, shape parameter of a gamma distribution of rate variation among amino acid sites.

Tree	−ln *L*	*δ*	*α*
Helicase			
1. (Flaviviruses, CFAV), TABV	17129·274	1·593	1·2315
2. (Flaviviruses, TABV), CFAV	17127·681	ML tree	1·234
3. (CFAV, TABV), flaviviruses	17129·032	1·351	1·236
4. (Flaviviruses, CFAV), (pestiviruses, TABV)	17161·945	34·264	1·206
NS3			
1. (Flaviviruses, CFAV), TABV	12495·775	ML tree	1·240
2. (Flaviviruses, TABV), CFAV	12497·588	1·813	1·256
3. (CFAV, TABV), flaviviruses	12498·133	2·358	1·254
4. (Flaviviruses, CFAV), (pestiviruses, TABV)	12551·484	55·709	1·170

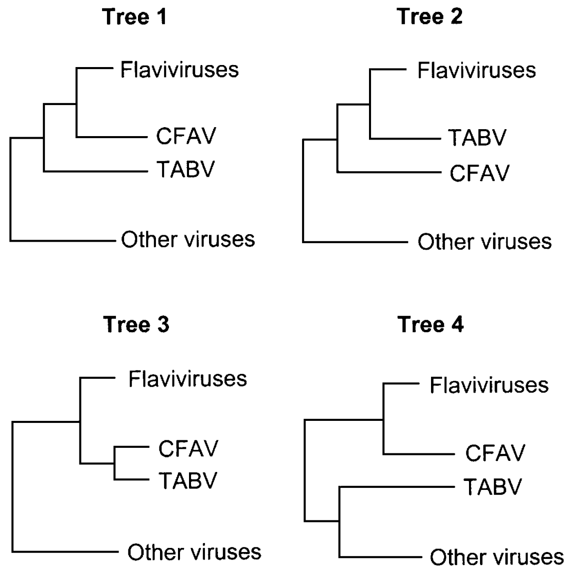


Fig. 5. Model trees that differ in the placement of the TABV lineage. Likelihood comparisons are given in Table 2.

ML analysis of these helicase and protease–helicase amino acid alignments also provided little phylogenetic resolution (Table 2). Minimal differences in likelihood were observed between trees in which TABV was depicted as either (i) the sister-group to the genus *Flavivirus* plus CFAV, (ii) more closely related to the genus *Flavivirus* than CFAV or (iii) forming a distinct clade with CFAV (Fig. 5, trees 1–3). Consequently, it is impossible to resolve the phylogenetic position of TABV on the basis of this set of data. However, a tree linking TABV with the pestiviruses, rather than with flaviviruses or CFAV (Fig. 5, tree 4), had a much lower likelihood in both the helicase and NS3 data sets, indicating that TABV is clearly more closely related to the genus *Flavivirus* and CFAV. This can be deduced from the *δ* value, as shown in Table 2: *δ* is > 30 for the topologies that group TABV together with pestiviruses and between 1 and 3 for all other topologies, indicating that the tree linking TABV with the pestiviruses is very unlikely.

Discussion

We have shown here that many characteristics of TABV deduced from the analysis of its coding sequence are shared with flaviviruses. The similar genomic organization, the unique polyprotein with conserved cleavage sites, the similarity of hydropathy plots, the highly conserved amino acid domains in the most important enzymes and in the structural proteins, the conserved cysteines presumably identifying sites for protein folding in the envelope protein and its phylogenetic relationships are all strong arguments to demonstrate that TABV belongs to the flavivirus evolutionary lineage. In contrast, the relatedness to the other lineages of the family *Flaviviridae* is less convincing, implying that TABV is not likely to be a member of the genera *Pestivirus* or *Hepacivirus*. This is compatible with findings from previous studies that reported common biological properties with enveloped arboviruses (Price, 1978) and a morphology similar to that of flaviviruses (Kuno *et al*., 1998). The same studies reported the absence of a serological relationship with flaviviruses and the fact that the TABV genome could not be amplified using various sets of primers specific for flaviviruses, suggesting that the TABV was quite different from 'classical' flaviviruses. This is also confirmed: TABV is genetically not closely related to any of the flaviviruses, either vectored or non-vectored, including CFAV. It constitutes a distinct genetic group within the flavivirus lineage, with some original characteristics, including the absence of conservation of cysteine residues in the NS1 protein and the predicted deviated substrate specificity of the VSP. The putative substrate-binding sites of this enzyme are substantially different from those of flaviviruses and four of the possible pairs preceding the putative cleavage sites are Gln–Arg residues. With the exception of the NS4A/2K cleavage site, where it is found for all flaviviruses, this pair is only found at the NS2B/NS3 site of DENV-1, -2 and -3, at the NS2A/NS2B site of APOIV and, interestingly, at both the NS2A/NS2B and NS2B/NS3 sites of RBV, another virus isolated from bats. Further studies are required to determine whether or not the fact that none of the VSP sites of the TABV polyprotein consists only of Arg and Lys residues correlates with specific enzymatic properties of the protease.

Another characteristic of the TABV genome is its low G + C content (38·4 mol%, lower than that of any flavivirus described to date). Low values (around 45 mol%) are also found in the genomes of viruses of the RBV serocomplex (Jenkins *et al*., 2001), but not for viruses with no known vector belonging to the YFV group [Sokoluk virus, *Yokose virus* and *Entebbe bat virus* (ENTV); Gaunt *et al*., 2001] or viruses within the genera *Pestivirus* and *Hepacivirus*. The relationship between G + C content, host specificity and/or phylogenetic origin of viruses is therefore still poorly understood in the family *Flaviviridae*. However, data presented here show that, within the flavivirus lineage, there is a correlation between the G + C content and the amino acid composition of the polyprotein.

It is as yet unclear why the phylogenetic position of TABV is so difficult to determine. There are two likely explanations: (i) that the diversification of TABV occurred very rapidly with respect to CFAV and flaviviruses or, more simply, (ii) that these data contain insufficient phylogenetic signal because of their extensive divergence. Unfortunately, given the great evolutionary distance between TABV and the flaviviruses, future phylogenetic resolution will require models of amino acid substitution specifically designed to deal with divergent RNA viruses. However, it must be noted that our data do not exclude the possibility that the evolutionary branch containing TABV diverged very early, possibly before the branch containing CFAV. Thus, one could imagine that persistent infection of mammals is an ancestral character of the whole family that has been conserved by hepaciviruses and pestiviruses and lost recently by some flaviviruses. This argument is supported by the fact that persistence is associated with (i) TABV, (ii) viruses in the RBV serocomplex (diverged from the deepest node of the flavivirus group in the NS5-like topology) (Kuno *et al*., 1998), (iii) viruses related to ENTV (the deepest divergence within the YFV group) (Kuno *et al*., 1998), (iv) viruses in the TBEV complex (Frolova *et al*., 1982, 1987) and (v) the more recently emergent *Saboya virus*. In addition, *Japanese encephalitis virus* (JEV) (Sulkin *et al*., 1970), WNV (Theiler & Downs, 1973), *St Louis encephalitis virus* (SLEV) (Sulkin *et al*., 1966) and DENV (Platt *et al*., 2000) have all been isolated from healthy bats, implying persistent infection. In other words, it cannot be excluded that flaviviruses were derived from viruses infecting mammals rather than from mosquito viruses, as has been proposed previously (Cammisa-Parks *et al*., 1992; Gubler, 1999; Porterfield, 1999).

In all cases, the molecular characterization of TABV is important for the taxonomic organization of the family *Flaviviridae*. CFAV and TABV have not yet been assigned to genera in the family *Flaviviridae*, although they have both been listed as tentative species in the genus *Flavivirus* (Heinz *et al*., 2000). Neither CFAV nor TABV satisfies other criteria listed in the ICTV scheme of classification for inclusion in existing genera within the family *Flaviviridae*. In particular, antigenic relationships have been used as a simple and efficient threshold for the delimitation of genera in this family. According to this criterion (and in accordance with the analysis of genetic distances), TBAV should be assigned to a second genus in the flavivirus lineage, for which the tentative name of genus Tamanavirus might be proposed. Following the same criteria, CFAV should be assigned to a third genus in the flavivirus lineage, and at least three new genera should be created within the hepaciviruses. Thus, the family *Flaviviridae* might include seven or more genera (three of them represented by single virus species at the present time). Alternatively, if the current taxonomic position is retained, i.e. with three large genera representative of the three evolutionary lineages, TABV and CFAV would belong to the genus *Flavivirus*.

Finally, the phylogenetic relationship observed between the helicase genes of members of the families *Flaviviridae* (including TABV) and *Potyviridae* is an intriguing feature. It should be noted that recombination events in various genes have been detected to date in viruses related to hepaciviruses (GBV-C; Worobey & Holmes, 2001), in flaviviruses (DENV; Tolou *et al*., 2001), in pestiviruses (Meyers & Thiel, 1996) and also very recently in potyviruses (Bousalem *et al*., 2000). Because relatedness between the families *Flaviviridae* and *Potyviridae* was not identified in genes other than the helicase, its origin might be the result of horizontal transfer of genetic information, possibly through genetic recombination (Goldbach, 1992), that occurred between ancient ancestors of these viruses. Whether or not these novel genetic exchanges have taken place in this manner will become known only when new and more sensitive methods for comparative analysis become available.

The authors are indebted to Dr S. W. Rasmussen for creating the MDB-BLAST program. The ‘Unité des Virus Emergents’ is an associated research unit of the Institut de Recherche pour le Développement (IRD). This study was supported in part by the IRD.

References

Arias, C. F., Preugschat, F. & Strauss, J. H. (1993). Dengue 2 virus NS2B and NS3 form a stable complex that can cleave NS3 within the helicase domain. *Virology* **193**, 888–899.

Bazan, J. F. & Fletterick, R. J. (1989). Detection of a trypsin-like serine protease domain in flaviviruses and pestiviruses. *Virology* **171**, 637–639.

Bellgard, M. I. & Gojobori, T. (1999). Significant differences between the G + C content of synonymous codons in orthologous genes and the genomic G + C content. *Gene* **238**, 33–37.

Billoir, F., de Chesse, R., Tolou, H., de Micco, P., Gould, E. A. & de Lamballerie, X. (2000). Phylogeny of the genus *Flavivirus* using complete coding sequences of arthropod-borne viruses and viruses with no known vector. *Journal of General Virology* **81**, 781–790.

Bousalem, M., Douzery, E. J. P. & Fargette, D. (2000). High genetic diversity, distant phylogenetic relationships and intraspecies recombination events among natural populations of *Yam mosaic virus*: a contribution to understanding potyvirus evolution. *Journal of General Virology* **81**, 243–255.

Cammisa-Parks, H., Cisar, L. A., Kane, A. & Stollar, V. (1992). The complete nucleotide sequence of cell fusing agent (CFA): homology between the nonstructural proteins encoded by CFA and the nonstructural proteins encoded by arthropod-borne flaviviruses. *Virology* **189**, 511–524.

Chambers, T. J., Weir, R. C., Grakoui, A., McCourt, D. W., Bazan, J. F., Fletterick, R. J. & Rice, C. M. (1990). Evidence that the N-terminal domain of nonstructural protein NS3 from yellow fever virus is a serine protease responsible for site-specific cleavages in the viral polyprotein. *Proceedings of the National Academy of Sciences, USA* **87**, 8898–8902.

Falgout, B., Pethel, M., Zhang, Y. M. & Lai, C. J. (1991). Both nonstructural proteins NS2B and NS3 are required for the proteolytic processing of dengue virus nonstructural proteins. *Journal of Virology* **65**, 2467–2475.

Forwood, J. K., Brooks, A., Briggs, L. J., Xiao, C. Y., Jans, D. A. & Vasudevan, S. G. (1999). The 37-amino-acid interdomain of dengue virus NS5 protein contains a functional NLS and inhibitory CK2 site. *Biochemical and Biophysical Research Communications* **257**, 731–737.

Frolova, T. V., Pogodina, V. V., Frolova, M. P. & Karmysheva, V. I. (1982). Characteristics of long-term persisting strains of tick-borne encephalitis virus in different forms of the chronic process in animals. *Voprosy Virusologii* **27**, 473–479 (in Russian).

Frolova, T. V., Frolova, M. P., Pogodona, V. V., Sobolev, S. G. & Karmysheva, V. I. (1987). Pathogenesis of persistent and chronic forms of tick-borne encephalitis (experimental study). *Zhurnal Nevrologii i Psikhiatrii Imeni S. S. Korsakova* **87**, 170–178 (in Russian).

Gaunt, M. W., Sall, A. A., de Lamballerie, X., Falconar, A. K. I., Dzhivanian, T. I. & Gould, E. A. (2001). Phylogenetic relationships of flaviviruses correlate with their epidemiology, disease association and biogeography. *Journal of General Virology* **82**, 1867–1876.

Goldbach, R. (1992). The recombinative nature of potyviruses: implications for setting up true phylogenetic taxonomy. *Archives of Virology Supplementum* **5**, 299–304.

Gorbalenya, A. E., Donchenko, A. P., Koonin, E. V. & Blinov, V. M. (1989a). N-terminal domains of putative helicases of flavi- and pestiviruses may be serine proteases. *Nucleic Acids Research* **17**, 3889–3897.

Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P. & Blinov, V. M. (1989b). Two related superfamilies of putative helicases involved in replication, recombination, repair and expression of DNA and RNA genomes. *Nucleic Acids Research* **17**, 4713–4730.

Gubler, D. J. (1999). Dengue viruses (*Flaviviridae*). In *Encyclopedia of Virology*, 2nd edn, pp. 375–384. Edited by A. Granoff & R. G. Webster. New York: Academic Press.

Heinz, F. X., Collett, M. S., Purcell, R. H., Gould, E. A., Howard, C. R., Houghton, M., Moormann, R. J. M., Rice, C. M. & Thiel, H.-J. (2000). *Flaviviridae*. In *Virus Taxonomy. Seventh Report of the International Committee on Taxonomy of Viruses*, pp. 859–878. Edited by M. H. V. van Regenmortel, C. M. Fauquet, D. H. L. Bishop, E. B. Carstens, M. K. Estes, S. M. Lemon, J. Maniloff, M. A. Mayo, D. J. McGeoch, C. R. Pringle & R. B. Wickner. San Diego: Academic Press.

Jenkins, G. M., Pagel, M., Gould, E. A., Zanotto, P. M. de A. & Holmes, E. C. (2001). Evolution of base composition and codon usage bias in the genus *Flavivirus*. *Journal of Molecular Evolution* **52**, 383–390.

Koonin, E. V. (1993). Computer-assisted identification of a putative methyltransferase domain in NS5 protein of flaviviruses and λ2 protein of reovirus. *Journal of General Virology* **74**, 733–740.

Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244–1245.

Kuno, G., Chang, G.-J. J., Tsuchiya, K. R., Karabatsos, N. & Cropp, C. B. (1998). Phylogeny of the genus *Flavivirus*. *Journal of Virology* **72**, 73–83.

Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**, 105–132.

Lain, S., Riechmann, J. L., Martin, M. T. & Garcia, J. A. (1989). Homologous potyvirus and flavivirus proteins belonging to a superfamily of helicase-like proteins. *Gene* **82**, 357–362.

Lobigs, M. (1992). Proteolytic processing of a Murray Valley encephalitis virus non-structural polyprotein segment containing the viral proteinase: accumulation of a NS3–4A precursor which requires mature NS3 for efficient processing. *Journal of General Virology* **73**, 2305–2312.

Mandl, C. W., Guirakhoo, F., Holzmann, H., Heinz, F. X. & Kunz, C. (1989). Antigenic structure of the flavivirus envelope protein E at the molecular level, using tick-borne encephalitis virus as a model. *Journal of Virology* **63**, 564–571.

Meyers, G. & Thiel, H.-J. (1996). Molecular characterization of pestiviruses. *Advances in Virus Research* **47**, 53–118.

Nishizawa, M. & Nishizawa, K. (1998). Biased usages of arginines and lysines in proteins are correlated with local-scale fluctuations of the G+C content of DNA sequences. *Journal of Molecular Evolution* **47**, 385–393.

Page, R. D. M. (1996). TreeView: an application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* **12**, 357–358.

Platt, K. B., Mangiafico, J. A., Rocha, O. J., Zaldivar, M. E., Mora, J., Trueba, G. & Rowley, W. A. (2000). Detection of dengue virus neutralizing antibodies in bats from Costa Rica and Ecuador. *Journal of Medical Entomology* **37**, 965–967.

Poch, O., Sauvaget, I., Delarue, M. & Tordo, N. (1989). Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *EMBO Journal* **8**, 3867–3874.

Porterfield, J. S. (1999). Encephalitis viruses (*Flaviviridae*): encephalitis viruses and related viruses causing hemorrhagic disease. In *Encyclopedia of Virology*, 2nd edn, pp. 424–430. Edited by A. Granoff & R. G. Webster. New York: Academic Press.

Price, J. L. (1978). Isolation of Rio Bravo and a hitherto undescribed agent, Tamana bat virus, from insectivorous bats in Trinidad, with serological evidence of infection in bats and man. *American Journal of Tropical Medicine and Hygiene* **27**, 153–161.

Pugachev, K. V., Nomokonova, N. Y., Dobrikova, E. Y. & Wolf, Y. I. (1993). Site-directed mutagenesis of the tick-borne encephalitis virus NS3 gene reveals the putative serine protease domain of the NS3 protein. *FEBS Letters* **9**, 115–118.

Rice, C. M. (1996). *Flaviviridae*: the viruses and their replication. In *Fields Virology*, 3rd edn, pp. 931–959. Edited by B. N. Fields, D. M. Knipe & P. M. Howley. Philadelphia: Lippincott–Raven.

Rice, C. M. & Strauss, J. H. (1990). Production of flavivirus polypeptides by proteolytic processing. *Seminars in Virology* **1**, 357–367.

Roehrig, J. T., Hunt, A. R., Johnson, A. J. & Hawkes, R. A. (1989). Synthetic peptides derived from the deduced amino acid sequence of the E-glycoprotein of Murray Valley encephalitis virus elicit antiviral antibody. *Virology* **171**, 49–60.

Stadler, K., Allison, S. L., Schalich, J. & Heinz, F. X. (1997). Proteolytic activation of tick-borne encephalitis virus by furin. *Journal of Virology* **71**, 8475–8481.

Steffens, S., Thiel, H.-J. & Behrens, S.-E. (1999). The RNA-dependent RNA polymerases of different members of the family *Flaviviridae* exhibit similar properties *in vitro*. *Journal of General Virology* **80**, 2583–2590.

Steiner, D. F., Smeekens, S. P., Ohagi, S. & Chan, S. J. (1992). The new enzymology of precursor processing endoproteases. *Journal of Biological Chemistry* **267**, 23435–23438.

Sulkin, S. E., Sims, R. A. & Allen, R. (1966). Isolation of St Louis encephalitis from bats (*Tadarida mexicana*) in Texas. *Science* **152**, 223–225.

Sulkin, S. E., Allen, R., Miura, T. & Toyokawa, K. (1970). Studies of arthropod-borne virus infections in Chiroptera. VI. Isolation of Japanese B encephalitis virus from naturally infected bats. *American Journal of Tropical Medicine and Hygiene* **19**, 77–87.

Swofford, D. L. (2000). PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods), version 4. Sunderland, MA: Sinauer.

Tan, B. H., Fu, J., Sugrue, R. J., Yap, E. H., Chan, Y. C. & Tan, Y. H. (1996). Recombinant dengue type 1 virus NS5 protein expressed in *Escherichia coli* exhibits RNA-dependent RNA polymerase activity. *Virology* **216**, 317–325.

Theiler, M. & Downs, W. G. (1973). *The Arthropod-borne Viruses of Vertebrates: an Account of the Rockefeller Foundation Virus Program (1951–1970).* London: Yale University Press.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.

Tolou, H. J. G., Couissinier-Paris, P., Durand, J.-P., Mercier, V., de Pina, J.-J., de Micco, P., Billoir, F., Charrel, R. N. & de Lamballerie, X. (2001). Evidence for recombination in natural populations of dengue virus type 1 based on the analysis of complete genome sequences. *Journal of General Virology* **82**, 1283–1290.

Valle, R. P. & Falgout, B. (1998). Mutagenesis of the NS3 protease of dengue virus type 2. *Journal of Virology* **72**, 624–632.

von Heijne, G. (1984). How signal sequences maintain cleavage specificity. *Journal of Molecular Biology* **173**, 243–251.

Warrener, P., Tamura, J. K. & Collett, M. S. (1993). RNA-stimulated NTPase activity associated with yellow fever virus NS3 protein expressed in bacteria. *Journal of Virology* **67**, 989–996.

Wengler, G. & Wengler, G. (1993). The NS3 nonstructural protein of flaviviruses contains an RNA triphosphatase activity. *Virology* **197**, 265–273.

Wengler, G., Czaya, G., Farber, P. M. & Hegemann, J. H. (1991). *In vitro* synthesis of West Nile virus proteins indicates that the amino-terminal segment of the NS3 protein contains the active centre of the protease which cleaves the viral polyprotein after multiple basic amino acids. *Journal of General Virology* **72**, 851–858.

Worobey, M. & Holmes, E. C. (2001). Homologous recombination in GB virus C/hepatitis G virus. *Molecular Biology and Evolution* **18**, 254–261.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**, 555–556.

Zhang, L., Mohan, P. M. & Padmanabhan, R. (1992). Processing and localization of Dengue virus type 2 polyprotein precursor NS3-NS4A-NS4B-NS5. *Journal of Virology* **66**, 7549–7554.

Received 7 February 2002; Accepted 3 April 2002