Complete genome characterization of Rocio virus (\textit{Flavivirus: Flaviviridae}), a Brazilian flavivirus isolated from a fatal case of encephalitis during an epidemic in São Paulo state

Daniele B. A. Medeiros,\textsuperscript{1} Márcio R. T. Nunes,\textsuperscript{1} Pedro F. C. Vasconcelos,\textsuperscript{1} Gwong-Jen J. Chang\textsuperscript{2} and Goro Kuno\textsuperscript{2}

\textsuperscript{1}Secção de Arbovirologia e Febres Hemorrágicas, Instituto Evandro Chagas, Secretaria de Vigilância em Saúde, Ministério da Saúde, Belém, Pará, Brasil
\textsuperscript{2}Arboviral Diseases Branch, Division of Vector-Borne Infectious Diseases, Centers for Disease Control and Prevention (CDC), Fort Collins, CO, USA

The flaviviruses of major medical importance in South American countries are yellow fever, dengue, Saint Louis encephalitis, West Nile and Rocio viruses. Rocio virus (ROCV) has been responsible for epidemics of severe encephalitis in Brazil with a case-fatality rate of 10\% and development of sequelae in 20\% of the survivors. We have sequenced and characterized the entire genome of ROCV for the first time, by determining the general traits of the open reading frame and the characteristics of viral genes including the potential cleavage sites, conserved or unique motifs, cysteine residues and potential glycosylation sites. The conserved sequences in the 3'-non-coding region were identified, and the predicted secondary structures during cyclization between 5’- and 3’-non-coding regions were studied. Multiple protein and phylogenetic analyses based on antigenically important and phylogenetically informative genes confirmed a close relationship between ROCV and Ilheus virus (ILHV), together constituting a unique and distinct phylogenetic subgroup as well as the genetic relationship of ROCV with several members of the Japanese encephalitis group. Although ROCV is phylogenetically related to ILHV, our study shows that it is still a virus distinct from the latter virus. This is the first flavivirus uniquely indigenous to Brazil that has been sequenced completely and the genome characterized. The data should be useful for further studies at the molecular level, including construction of infectious clone, identification of gene function, improved disease surveillance based on molecular diagnostic tools and vaccine development.

INTRODUCTION

In South America, until the mid-1970s, the two neurotropic flaviviruses had been Saint Louis encephalitis virus (SLEV) and Ilheus virus (ILHV). However, in contrast to North America, an epidemic of encephalitis in humans by SLEV in South America was not recorded for many years until the first outbreak in Argentina in 2005 (Diaz \textit{et al.}, 2006), despite accumulated records of virus isolation from mosquitoes and vertebrate hosts throughout the continent (Spence, 1980). As for the other virus, ILHV, human infection was sporadic and development of neurotropic manifestations has been even more infrequent. Thus, the sudden emergence of an outbreak of encephalitis caused by a flavivirus in 1974–75 in Brazil was unexpected.

Rocio virus (ROCV) belongs to the genus \textit{Flavivirus} in the family \textit{Flaviviridae}. Its prototype (strain SPH 34675) was isolated in Ribeira Valley in the south-eastern São Paulo state of Brazil in 1975 from the cerebellum of a fatal human case during an epidemic of encephalitis that could not be readily diagnosed with the reagents then available for known encephalitic viruses. The epidemic, which began in 1974, spread to more than 20 municipalities over the following 2 years causing approximately 1000 diagnosed cases.

Members of the genus \textit{Flavivirus} (hereafter called flaviviruses) are positive-sense, single-stranded RNA viruses that infect humans and other vertebrates. They are mainly transmitted by bites of haematophagous arthropods (either mosquitoes or ticks), except for a group of vertebrate...
viruses that do not need an arthropod-borne mode of transmission for survival (hereafter called no-vector viruses) and an increasing number of arthropod flaviviruses that replicate only in arthropods (mosquitoes). Currently, in this genus, more than 60 distinct viruses have been recognized worldwide with several others being classified as either tentative species or genotypes of the other viruses (Gould et al., 2003; ICTV, 2005). The genomic RNA consists of a unique open reading frame (ORF) that is flanked by a type 1 capped 5'-terminal non-coding region (NCR) and a 3'-terminal NCR that, together, form specific secondary stem–loop structures required for RNA translation, replication and/or expression of biological traits most likely including pathogenic determinants (Lindenbach & Rice, 2003; Rice et al., 1985). The flavivirus ORF encodes 10 proteins including three structural [capsid (C), pre-membrane (prM) and envelope (E)] and seven non-structural (NS1, NS2a, NS2b, NS3, NS4a, NS4b and NS5) proteins (Chambers et al., 1990).

Although ROCV was found to be antigenically related to ILHV and grouped in the Japanese encephalitis virus (JEV) antigenic complex (Chamberlain, 1980), a phylogenetic study based on partial NS5 gene revealed a close but distinct relation of ROCV and ILHV from the JEV complex viruses (Kuno et al., 1998). However, fundamental studies based on the full-genome sequence are desirable for identifying in more depth the phylogenetic relationship, genetic traits and other molecular traits of these neurotropic viruses. In this report, we selected ROCV as the first virus for such investigations. The data obtained from such studies should, in turn, facilitate vaccine preparation and designing specific primers and probes for a better diagnosis and surveillance of Rocío encephalitis. Recently, evidence of human infection by ROCV was again reported (Strattmann et al., 1997); and another outbreak in the future is highly possible. Thus, the study undertaken was timely because the diagnosis of flaviviral infections in South America has become more complicated due to the recent introduction of the West Nile virus (WNV) and the occurrence of unusual haemorrhagic manifestations in laboratory-confirmed SLE cases that resulted in confusion with dengue in Brazil (Mondini et al., 2007).

In order to meet such an epidemiological threat effectively, we have determined and characterized the full-length genome sequence of ROCV (strain SPH 34675) that was obtained from the cerebellum of a fatal encephalitis case in 1975 (Iversson, 1988; de Sousa Lopes et al., 1978, 1981).

METHODS

Virus. Sequencing of ROCV was conducted in two laboratories [CDC and Instituto Evandro Chagas (IEC)] as a collaborative research study, using the identical strain (SPH 34675) at similar mouse brain passage levels (level 4 in CDC and level 5 in IEC).

RNA extraction, reverse-transcription, PCR and nucleotide sequencing. Viral RNA was extracted directly from a suspension of infected mouse brain using the QIAamp Viral RNA Mini kit (Qiagen) according to the manufacturer’s instruction. A cDNA spanning from the 5’-terminus to the conserved sequence (CS2) in the 3’-NCR was prepared by reverse transcribing viral RNA using reverse transcriptase and a reverse primer VD8 (5’-GGGTCTCCTCTAACC-CTCTAGAAT-3’) previously designed from the CS2 sequence (Pierre et al., 1994). PCR was performed using the Expand Long Template PCR System kit (Roche Applied Science). The thermostatic programme set up in the Gene Amp PCR System 9600 thermocycler (Perkin-Elmer) was 1 cycle of 94 °C for 1 min, 50 °C for 5 min; 3 cycles of 94 °C for 20 s, 50 °C for 1 min, 68 °C for 4 min; 10 cycles of 94 °C for 20 s, 50 °C for 30 s, 68 °C for 4 min with an increment of 20 s per cycle; and a final cycle of extension at 68 °C for 7 min. PCR products (amplicons) were purified with a Centricep column (Princeton Separations), and aliquots of approximately 60–160 ng of the purified DNA templates were used for direct cycle sequencing by using a Prism DNA sequencing kit (Big Dye) for dye terminator cycle sequencing with Ampli-Taq FS enzyme (ABI) and a CEQ8000 Genetic Analysis System (Beckman Coulter).

Because the full account of the sequencing strategy based on primer walking as well as a list of degenerate primers used were described earlier (Kuno & Chang, 2007) and to avoid redundancy, only the essential features of the technique are briefly presented here. For sequencing with this cDNA, a number of degenerate primers were designed either in one or in both directions, based on the conserved amino acid sequences in the non-structural genes (NS), previously identified in a multiple sequence alignment of flaviviral sequences available in GenBank. The gene and amino acid sequences selected for such degenerate primers are as follows: NS3 (GTSGSPI, GLYGNG, LAPTRVV, DVMCHATE, MDEAHF, SIAARGY, MTATPPGO, ISEMGAN and SAAQRRG) and NS5 (DGCGCRG, SRNSTHEMY, NNMGGKREK, ADDTAGWD and WMTTEDML). Degenerate primers were also designed from less conserved genes: envelope (E) (DRGWGNGC, GLFGKGS, GHKLCRV, PFDGSYIV and DTAWDFGS), NS1 (GCWYWMEI and YGMEIRP). The three forward primers in the 5’-NCR that were empirically determined are 5’-ATG(A/T)CTAAA(A/G)AAACCAGGA3’-5’-AAACCAAGGA(A/G)A/G/ A(CCGG3’-3’-TCAATATGCT(A/G)AAACCAGGA3’, 5’-AAACCAAGGA(A/G)A/G/ A(CCGG3’-3’-TCAATATGCT(A/G)AAACCAGGA3’-5’-TCAGATGTGCT(A/G)AAACCAGGA3’-5’-CTCAATATGCT(A/G)AAACCAGGA3’).

In the first stage of sequencing, using pairs of forward and reverse degenerate primers, multiple amplicons spanning any regions within the cDNA were generated. Primers were selected so that the lengths of these amplicons were preferably less than 3.5 kb. The generated amplicons were sequenced in both directions: in the 5’- and 3’-ends of the genome were first poly-C-tailed or poly-A-tailed, respectively, and non-coding region (NCR) and a 3’-terminal NCR that, together, form specific secondary stem–loop structures required for RNA translation, replication and/or expression of biological traits most likely including pathogenic determinants (Lindenbach & Rice, 2003; Rice et al., 1985). The flavivirus ORF encodes 10 proteins including three structural [capsid (C), pre-membrane (prM) and envelope (E)] and seven non-structural (NS1, NS2a, NS2b, NS3, NS4a, NS4b and NS5) proteins (Chambers et al., 1990).

In the second stage, non-overlapping gaps between sequenced regions or between amplified genomic sequences obtained. Those internal primers were used for further sequencing of the amplicons were preferably less than 3.5 kb. The generated amplicons were sequenced in both directions: in the 5’- and 3’-ends of the genome were first poly-C-tailed or poly-A-tailed, respectively, and non-coding region (NCR) and a 3’-terminal NCR that, together, form specific secondary stem–loop structures required for RNA translation, replication and/or expression of biological traits most likely including pathogenic determinants (Lindenbach & Rice, 2003; Rice et al., 1985). The flavivirus ORF encodes 10 proteins including three structural [capsid (C), pre-membrane (prM) and envelope (E)] and seven non-structural (NS1, NS2a, NS2b, NS3, NS4a, NS4b and NS5) proteins (Chambers et al., 1990).

In the first stage of sequencing, using pairs of forward and reverse degenerate primers, multiple amplicons spanning any regions within the cDNA were generated. Primers were selected so that the lengths of these amplicons were preferably less than 3.5 kb. The generated amplicons were sequenced in both directions: in the 5’- and 3’-ends of the genome were first poly-C-tailed or poly-A-tailed, respectively, and then amplified using 5’- or 3’-RACE kits, according to the manufacturer’s instructions (Invitrogen).

Assembly of overlapping sequences for full-length genome and multiple sequence alignment. Partial sequences of ROCV were first inspected for quality and then assembled for generating a contiguous full-genome sequence, by using the SeqMan II program (v. 5.03) in LaserGene program (DNA Star). This software was also used to predict the entire ORF for the polyprotein of ROCV as well as to determine the 5’- and 3’-NCRs. The full-length genome sequence of ROCV was deposited in GenBank as revised version 1 (AY632542-R1) of the ORF sequence previously deposited (ROCV; AY632542), by including the 5’- and 3’-NCR sequences. Multiple sequence alignment was obtained first with the use of CLUSTAL W (Thompson et al., 1994), followed by manual adjustment with BioEdit (version 5.0.0) (Hall, 2001), and then realigned by applying the ‘removing gap...
function’ of MEGA align software (v. 5.03) in the LaserGene program (DNA Star) to generate correctly aligned multiple sequence files (for nucleotide and deduced amino acid sequences).

**Determination of cleavage sites, glycosylation sites, cysteine residues and conserved motifs.** Potential cleavage sites for ROCV polyprotein were primarily determined, according to the proteolytic processing cascade pattern for the flavivirus ORFs developed by Rice & Strauss (1990). The highest cleavage potential scores obtained by SignalP-NN computer program (http://www.cbs.dtu.dk/services/) (Chang et al., 2000) were used for determining the sites cleaved by the host cell-encoded signalase. Predicted glycosylation and cysteine residue sites were determined using the NetNGlyc (v.1.0) (http://www.cbs.dtu.dk/services/) and Proview (v. 5.03) of the LaserGene program (DNA Star), respectively. The MEGA align software was used to assess the presence of flaviviruses highly conserved motifs in the ROCV genome.

**Scanning of ROCV ORF and gene-by-gene analyses.** For the identification of related viral proteins, the amino acid sequence of ROCV ORF was scanned against other flavivirus sequences available in GenBank using the protein–protein BLAST program (http://www.ncbi.nlm.nih.gov/). Parameters were set as follows: database-filter, only virus sequences; matrix, BLOSUM62; gap cost, existence 11 and extension 1; expected [e-value threshold=0.005]. Scanning was performed on multiple amino acid sequence alignment prepared by the method above. Bootstrap values were obtained by generating 100 replicas, and the sequences most closely related to ROCV (query sequence) were displayed based on the lowest e-value (<0.005) and the highest p-score (p-score value that corresponds to a determined percentage identity). Sequences were represented by the lowest e-value and the highest genome p-score were used for multiple sequence alignments (gene-by-gene) using the CLUSTAL W software (Thompson et al., 1994), and percentage of similarities for nucleotide and deduced amino acid sequences were calculated with the MEGA align software (LaserGene program; DNA Star).

**Phylogeny.** For phylogenetic analyses, the sequences of the genes possessing antigenically important traits (E and NS1 genes) and conserved genes (NS3 and NS5) were concatenated in that order for ROCV, and other selected flaviviruses representing mosquito-borne, tick-borne, no-vector and arthropod (mosquito) virus groups. Neighbour-joining (NJ) (Saitou & Nei, 1987) and maximum-parsimony (MP) (Swofford, 1999) methods were performed. For NJ, a distance matrix calculated from the aligned sequences by Kimura two-parameter formula (Kimura, 1980) was used, and a weight of four for transitions versus one for transversion was selected. In MP, in order to obtain the most parsimonious tree, the heuristic algorithm method was performed; and for determining the reliability of tree topology bootstrap analysis (Felsenstein, 1995) was carried out on 1000 replicas. Bootstrap resampling technique (Efron, 1982) was then used to further evaluate the reliability of the bootstrap analysis with a confidence value of 0.95 (95%). Two distinctly related arthropod flaviviruses, cell fusing agent virus (CFAV) and Kamiti River virus (KRV), were used as outgroups. For further analysis of the relationship between ROCV and Ntaya virus (NTAV), the E-NS5 concatenated amino acid sequences of Bagaza virus (BAGV), NTAV and six members of the JEV group [Murray Valley encephalitis virus (MVEV), Usutu virus (USUV), JEV, WNV, Kunjin virus (KUNV) and SLEV)] were used. Confidence values used as criteria for group inclusion or exclusion were calculated based on the mean of the concatenated E-NS1-NS3-NS5 [71% ± 2 standard deviations (sd)] and E-NS5 (71.5% ± 2 sd) amino acid sequence identities within and among selected members of the JEV, dengue virus (DENV), tick-borne and no-vector flavivirus groups.

**Computer-generated secondary structures in the 3′-NCR and genome cyclization between the 5′- and 3′-NCRs.** The secondary structures in the 3′-NCR and cyclization between the 3′- and 5′-terminal regions were analysed with the mfold program (Zuker et al., 1999). The first 200 nt of the genome representing the entire 5′-NCR followed by a part of the C gene and the entire 3′-NCR were used for the analysis, as these two terminal regions of flaviviruses contain the domains for genome cyclization between them as well as highly conserved sequences in the 3′-NCR (Khromykh et al., 2001).

**RESULTS**

**Genome analyses**

The complete genome of ROCV was determined to be 10 794 nt in length with an ORF of 10 275 nt that is flanked by 5′- and 3′-NCRs of 92 and 427 nt, respectively. The complete ORF was predicted to encode a large polyprotein of 3425 aa, that, after co- and post-translational cleavage processes, gives rise to 10 viral proteins, typical of flaviviruses. The lengths and positions of the genes or genomic regions are shown in Table 1. To further examine the lengths of ROCV genes, we compared their nucleotide and amino acid sequences with selected mosquito-borne [ILHV, SLEV, WNv, JEV, DENV-1–4 and yellow fever virus (YFV)] and tick-borne (Langat virus, Powassan virus and Tick-borne encephalitis virus) flaviviruses, as well as with representative members of no-vector flaviviruses (Modoc virus and Rio Bravo virus). Although differences in genome length observed among flaviviruses are mainly due to size variability in the 3′-NCRs (Bryant et al., 2005; Wallner et al., 1995), size heterogeneity was also observed along the coding regions. Generally, ROCV genes are more similar in length with those of ILHV, SLEV, WNv and JEV than with DENV serotypes 1–4 and tick-borne flaviviruses, but the length differences with no-vector viruses were pronounced (Supplementary Table S1 available in JGV Online).

**Cleavage sites, glycosylation sites and cysteine residues**

The predicted pentapeptide cleavage sites of ROCV were compared with the corresponding sites of representative members of the JEV (WNV, SLEV, JEV and MVEV) and Ntaya (BAGV) groups as shown in Supplementary Table S2 (available in JGV Online). The comparison revealed more differences (two or more amino acids in either peptide) at VirC/AnchC, AnchC/pr, pr/M, NS1/NS2A, NS2A/NS2B and NS4B/NS5 potential cleavage sites where amino acid substitutions were found in the ROCV genome. Single residue substitutions were also observed in the M/E, E/NS1 and NS2B/NS3.

The numbers of N-linked potential glycosylation sites (N-LGlyS) in the prM, E and NS1 genes for ROCV were two, one and four, respectively. No N-LGlyS was found in the hydrophobic domain. This pattern of three site numbers is compared with two, one and three for Argentine 66 strain of SLEV; one, zero and two for JEV and one, one and three for WNv. However, it should be noted that strain variation in the number of sites has been reported. Thus, in the E
gene of SLEV, while the Argentine 66 has one, strain MSI-7 has two, but strain Tr 9464 has no glycosylation sites (Kuno & Chang, 2005; Vorndam et al., 1993). Regarding cysteine residues, as strictly required for stabilizing disulfide bridges in all mosquito-borne viruses, 12 residues were found in the ectodomain of E and in NS1 genes of ROCV. As previously observed among other flaviviruses, the six cysteine residues in prM are clustered in the N-terminal hydrophobic pr domain.

**Motifs**

In prM, a few peptides are conserved among flaviviruses. Although the amino acid W (at residue 68) of ROCV is shared by all mosquito-borne viruses, it is replaced with F in all tick-borne viruses. Many amino acid motifs in the E, NS3 and NS5 proteins are shared between ROCV and other mosquito-borne, tick-borne and no-vector flaviviruses. In the E protein, the tripeptide in the domain III region (amino acid residues 388–390), which corresponds to the RGD motif in JEV, Alfuy virus (ALFV), USUV and MVEV (Bakonyi et al., 2004; Hurrelbrink et al., 1999; May et al., 2006; Sumiyoshi et al., 1987), RGE in WNV and KUNV (Castle et al., 1985; Coia et al., 1988) and RGP in SLEV (Ciota et al., 2007), is TGP in ROCV. The TGP peptide is also found in Kedougou virus (KEDV) and is very similar to those of Kokobera virus (KOKV) and BAGV (TGE) (Kuno & Chang, 2007). Regarding the conserved motif corresponding to the fusion peptide region (residues 98–110), ROCV also shares with other mosquito-borne flaviviruses the same amino acid sequence, differing from tick-borne flavivirus members at the amino acid residue 104 (G versus H). In the NS3 gene, the catalytic triad (H-D-S) of protease is well preserved in ROCV. However, a notable difference from the mosquito-borne viruses is found in the hexapeptide motif MDEAHF (residues 284–289), where tick-borne and no-vector groups replace F by W. Among most mosquito-borne viruses including ROCV, the heptapeptide at residues 294–300 is SIAARGY, but Y is replaced by W among yellow fever group viruses (YFV, Sepik virus and Entebbe bat virus) and by H among tick-borne viruses. In addition, at residue 225 in the conserved motif LAPTRV (residues 222–228), a single amino acid change (T→S) was observed in ROCV. In the NS5 gene, like in NS3 gene, many conserved motifs are shared between ROCV and other flaviviruses. However, at the highly conserved peptide AMTDTTPFGQQRVFKEKVDT (residues 343–362) shared by all mosquito-borne viruses, including ROCV, P is replaced by A in both tick-borne and no-vector groups, while it is replaced by T for two insect flaviviruses (CFAV and KRV). At residues 550–551, the dipeptide amongst most mosquito-borne viruses, including ROCV, is NE; however, it is DE among KEDV, YFV, tick-borne, no-vector (except TBV) and arthropod flaviviruses.

**Genome scanning**

Protein–protein BLAST analysis using the entire ORF of ROCV as a query sequence revealed protein identity with several flaviviruses (e-values<0.0; p-scores ranging from 1083 to 5691). The highest similarity was observed with ILHV (p-score: 5681, 88 %) and members of the more broadly defined JEV group: SLEV (p-score: 4910; 81 %), WNV (p-score: 4799, 80 %), MVEV (p-score: 4784, 80 %), and JEV (p-score: 4784, 80 %). As shown in Table 2, the close genetic relationship between ROCV and ILHV determined by CLUSTAL W software (77.5 % for amino acid sequences and 71 % for nucleotides in ORF) is reflected in higher amino acid identities for NS2B (85.6 %), NS3 (85.4 %) and NS5 (79.9 %). The identities in the E, NS4A and NS4B proteins are slightly lower but still close to the level of NS5 protein. The reduced identity (55.9 %) found only with capsid protein of ILHV is peculiar because low-level identities below 55 % are found in multiple proteins

**Table 1.** Full-length genome organization of ROCV

<table>
<thead>
<tr>
<th>Proteins</th>
<th>Genes</th>
<th>Size (nt)</th>
<th>Genome position</th>
<th>Protein size (aa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural proteins</td>
<td>5’-NCR</td>
<td>92</td>
<td>1–92</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>303</td>
<td>93–395</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>AnchC</td>
<td>51</td>
<td>396–446</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>prM</td>
<td>276</td>
<td>447–722</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>pr</td>
<td>225</td>
<td>723–947</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>1 503</td>
<td>948–2450</td>
<td>501</td>
</tr>
<tr>
<td>Non-structural proteins</td>
<td>NS1</td>
<td>1 059</td>
<td>2 451–3 509</td>
<td>353</td>
</tr>
<tr>
<td></td>
<td>NS2A</td>
<td>681</td>
<td>3 510–4 190</td>
<td>227</td>
</tr>
<tr>
<td></td>
<td>NS2B</td>
<td>393</td>
<td>4 191–4 583</td>
<td>131</td>
</tr>
<tr>
<td></td>
<td>NS3</td>
<td>1 857</td>
<td>4 584–6 440</td>
<td>619</td>
</tr>
<tr>
<td></td>
<td>NS4A</td>
<td>378</td>
<td>6 441–6 818</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td>2K</td>
<td>69</td>
<td>6 819–6 887</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>NS4B</td>
<td>768</td>
<td>6 888–7 655</td>
<td>256</td>
</tr>
<tr>
<td></td>
<td>NS5</td>
<td>2 715</td>
<td>7 656–10 370</td>
<td>904</td>
</tr>
<tr>
<td></td>
<td>3’-NCR</td>
<td>424</td>
<td>10 371–10 794</td>
<td>–</td>
</tr>
</tbody>
</table>
besides capsid when four JEV complex viruses are compared (Table 2). The identity (67.1%) in prM of ILHV is also low but higher than that of the capsid protein. Another peculiarity of ROCV is that the identity in NS5 protein of ILHV is lower than those of NS2B and NS3 proteins, because generally NS5 gene is more conserved than the other genes among flaviviruses. Other than with ILHV, however, identities of ROCV NS3 and NS5 proteins with other viruses compared are generally similar.

Phylogeny

The phylogenetic analysis based on the concatenated, approximately 2300 aa E-NS1-NS3-NS5 amino acid sequences of ROCV and other flaviviruses using MP and NJ methods resulted in two trees with a similar topology, although the bootstrap values in the MP consensus tree were slightly lower than those in the NJ tree (data not shown). Accordingly, in this report, the NJ tree was selected to represent the phylogeny of the selected viruses. Besides the identity in NS5 protein of ILHV is lower than those of NS2B and NS3 proteins, because generally NS5 gene is more conserved than the other genes among flaviviruses. Other than with ILHV, the identity in ROCV NS3 and NS5 proteins with other viruses compared are generally similar.

Phylogeny

The phylogenetic analysis based on the concatenated, approximately 2300 aa E-NS1-NS3-NS5 amino acid sequences of ROCV and other flaviviruses using MP and NJ methods resulted in two trees with a similar topology, although the bootstrap values in the MP consensus tree were slightly lower than those in the NJ tree (data not shown). Accordingly, in this report, the NJ tree was selected to represent the phylogeny of the selected viruses. Based on reliable bootstrap values (>75%) and confidence values (≥71 and 71.5% ± 2 SD for E-NS1-NS3-NS5 and E-NS5, respectively) as the criteria for group inclusion, the phylogram clearly demonstrates major cluster segregation with 100% bootstrap values among the mosquito-borne, tick-borne and no-vector flaviviruses, and subclassification of the mosquito-borne cluster to nine subclades (Fig. 1a). In the mosquito-borne cluster, divergence of the two major clades (predominantly Culex-borne viruses and Aedes-borne viruses) is evident. As shown in Fig. 1(a), the encephalitogenic ROCV is found in the predominantly Culex-borne virus clade and is most closely affiliated with ILHV. When the concatenated sequence is limited to E and NS5, as in Fig. 1(b), segregation of the lineage consisting of ROCV and ILHV from JEV complex viruses and also from Ntaya group (BAGV and NTAV) is more clearly demonstrated.

5′- and 3′-NCRs

The analysis of the computer-generated secondary structures of ROCV revealed a similar pattern found in other mosquito-borne flaviviruses. The conserved sequences (CS3, RCS2, CS2 and CS1) and associated secondary structures (Chambers et al., 1990; Hahn et al., 1987; Proutski et al., 1997) are evident in Fig. 2. The predicted cyclization between the cyclization domain (5′-CYC) of the 5′-NCR and the complementary sequence (3′-CYC) of the 3′-NCR is also illustrated in Fig. 2. As expected, the pentanucleotide (CACAG) that is critically important for the replication of all flaviviruses is found in the loop of the long, stable hairpin structure (3′-LSH) near the 3′-terminal sequence (Fig. 2).

The CS1 (GCATATTGACACCTGGGAAAGAC) and CS2 (GGACTAGAGGTTAGAGGAGACCC) of ROCV are nearly identical to the corresponding conserved sequences among YFV, DENV and JEV (Chambers et al., 1990; Hahn et al., 1987; Khromykh et al., 2001). On the other hand, the 3′-terminal sequence of the CS3 of ROCV (CCCCAGGACGCTGGGGTTTTTCAAGGC) is slightly degenerate or different. The pattern of CS organization (in 5′→3′ direction) in the 3′-NCR of ROCV is CS3-RCS2-CS2-CS1. The structure illustrated in Fig. 2 suggests that CS2, RCS2 and CS3, as well as 3′-LSH, are involved in the secondary structure. Generally, a high level of stability is observed along the predicted secondary structures in the lower left region below the dashed line in Fig. 2, where minor variation in the number of bulges and/or small loops (SmLs) is noted at different holding energy levels. By contrast, the region in the upper right of Fig. 2 above the dashed line including the CS2, RCS2 and CS3, as well as the hairpin sequences (HS) (5′-CapHS, 3′-HS1–3), revealed a greater variation in possible folding patterns generated at different energy levels.

DISCUSSION

During the encephalitis outbreak in humans in the São Paulo state located in the south-eastern region of Brazil in the 1970s, an enormous public health impact of ROCV was felt, because the outbreak affected at least 23 municipalities in the Ribeira Valley area of the São Paulo state. The encephalitic disease was characterized by an acute onset of fever, headache, and several central nervous system (CNS) signs and symptoms, resulting in a case-fatality rate of over 10%, and, importantly, development of CNS sequelae in
Fig. 1. NJ trees of ROCV and selected flaviviruses based on amino acid sequences of (a) concatenated E-NS1-NS3-NS5 or (b) E-NS5. Two distant arthropod flaviviruses (KRV and CFAV) and KOKV were used as outgroups in (a) and (b), respectively. The roman numerals refer to subclades described in the Discussion. The numbers at nodes indicate bootstrap support values and arrows indicate high supporting bootstrap value for JEV (97%), ROCV (100%) and Ntaya (100%) groups. Confidence values used as criteria for group inclusion or exclusion are shown in parentheses. The horizontal scale bar represents 10 and 5% nucleotide sequence divergences for (a) and (b), respectively. GenBank accession nos are shown in square brackets. MVEV (Murrey valley encephalitis virus), USUV (Usutu virus), JEV (Japanese encephalitis virus), KUNV (Kunjin virus), WNV (West Nile virus), SLEV (Saint Louis encephalitis virus), ILHV (Ilheus virus), ROCV (Rocio virus), BAGV (Bagaza virus), NTAV (Ntaya virus), KOKV (Kokobera virus), KEDV (Kedougou virus), DENV-1 (dengue virus type 1), DENV-2 (dengue virus type 2), DENV-3 (dengue virus type 3), DENV-4 (dengue virus type 4), YFV (yellow fever virus), TBEV (tick-borne encephalitis virus), LIV (louping ill virus), OHFV (Omsk hemorrhagic fever virus), LGTV (Langat virus), ALKV (Alkhurma virus), POWV (Powassan virus), MMLV (Montana myotis leukoencephalitis virus), RBV (Rio Bravo virus), MODV (Modoc virus), APOIV (Apoi virus), KRV (Kamiti River virus) and CFAV (cell fusing agent virus).

Fig. 2. Predicted secondary structures of the 5'- and 3’-NCRs of ROCV in cyclization. The folding pattern was generated by the mfold program (energy level = -97.0). Abbreviations: CS, conserved sequences; RCS, repeated conserved sequence; SmLs, small loop; LSH, long stable hairpin; 3’-CYC, cycling sequence within CS1 of the 3’-NCR; 5’-CYC, cycling sequence within capsid (Cap) gene; HS, hairpin sequence and NCR, non-coding region.
about 20% of the survivors (de Sousa Lopes et al., 1978, 1981; Iversson, 1988). The most frequent CNS symptoms were meningeal irritation (57.3%), alteration of consciousness (51%) and motor abnormalities (49.65%) especially gait and impaired equilibrium, but blindness and deafness were also reported (Iversson, 1988). Accordingly, full genome characterization and determination of the exact phylogenetic and taxonomic relationships of ROCV with other related flaviviruses were undertaken, as such data will be essential for achieving the ultimate goals of designing better molecular probes and primers for improved surveillance and diagnosis, determination of the neurovirulence markers at a molecular level, and development of attenuated vaccine and antiviral drugs.

As a result of accelerated sequencing studies, more than 40 flaviviruses have been fully sequenced thus far (Kuno et al., 1998; Billoir et al., 2000; Gaunt et al., 2001; Grard et al., 2007). Unfortunately, with the exception of SLEV, the paucity of such full-length genomic sequences for the viruses indigenous to South America has hampered the understanding of the evolution of ROCV and its genetic characteristics. Thus, our gene-by-gene genome scanning and phylogenetic analysis provide information for a better understanding of the genetic relationship of ROCV among related flaviviruses.

The genomic organization of ROCV shares the same pattern found in all other flaviviruses (Rice et al., 1985; Chambers et al., 1990). This study also reveals that ROCV shares many of its genetic traits with other mosquito-borne flaviviruses. Thus, as expected, its cysteine residue distribution pattern in prM, E and NS1 genes (6, 12 and 12, respectively) is identical to the pattern observed in other flaviviruses, because of the critical importance in proper conformation afforded by disulfide bridges.

Regarding the presence of potential N-LGlyS along the prM-E-NS1 genes in ROCV, differences in the number of sites were observed in the E and NS1 genes of the mosquito-borne, tick-borne and no-vector flaviviruses. Furthermore, as mentioned earlier, strain variation exists at least for SLEV (Spence, 1980; Vorndam et al., 1993; Diaz et al., 2006). Thus, basically the function of the potential sites in ROCV remains unknown in the absence of experimental data; and, clearly, further experimental studies are necessary to clarify the functional significance. Here, it is important to mention contradictory findings in the past. In one study, the correlation between the presence of such sites and antigenic properties of flavivirus was not established because deglycosylated viruses could maintain the same antigenicity (Winkler et al., 1987). On the other hand, in other more recent studies, glycosylation was found to be very important for viral replication, virulence, maturation or release of viral-like particles (Beasley et al., 2005; Crabtree et al., 2005; Goto et al., 2005; Li et al., 2006).

Our analysis of the conserved motifs showed that ROCV shares numerous conserved amino acid and/or peptides with several flaviviruses across different groups, each with a varied host range. Interestingly, in the E protein, the tripeptide (TGP) of ROCV within domain III was found to be different from those observed in all other JEV members (RGX, where X is D, E or T). As the RGD motif is speculated to be involved in the virus adsorption to host cells for JEV and MVEV, two important JEV group flaviviruses (Bhardwaj et al., 2001; Lee & Lobigs, 2000) and as it was also found in ALFV and USUV, further experimental studies for ROCV are needed to investigate the role of the TGP tripeptide in the mechanisms that involve the virus–cell interaction.

The functional roles of the 3'-NCR of flaviviruses have been recently studied and the roles in the genomic cyclization between the 5'- and 3'-NCRs, viral replication and translation were elucidated. Moreover, the putative changes in the highly conserved regions related to important biological functions of the flaviviruses (survival and/or pathogenicity) were recognized (Bryant et al., 2005; Khromykh et al., 2001). Khromykh et al. (2001) demonstrated that at least one base-pairing between the 5’–3’ termini, rather than the nucleotide sequences per se, is essential for the replication of viral RNA of the mosquito-borne Kunjin virus, a member of the JEV group, and that more than one pair of cyclization sequences might be involved in the replication of the tick-borne flaviviruses.

In terms of CS organization in the 3'-NCR, our results revealed a CS3-CS2-RCS2-CS1 pattern for ROCV. This CS organization for ROCV is very similar to but distinct from that of the JEV group viruses (CS3-CS2-RCS2-CS1). Thus, regarding the CS organization in the 3'-NCR, the relationship of ROCV to JEV group viruses is similar to the previously established phylogenetic relationship in which ROCV is closer to the JEV complex viruses but is still distinct (Kuno et al., 1998). Although the CS pattern in the 3'-NCR has not been used for subgrouping within the mosquito-borne flaviviruses, as shown in this report, it deserves further studies to determine its utility as a useful supplementary taxonomic marker at viral population level rather than at individual strain level.

According to the recent taxonomic classification, ROCV is classified as a genotype of ILHV in the Ntaya virus group (ICTV, 2005). Numerous phylogenetic studies (Billoir et al., 2000; Cook & Holmes, 2005; Gould et al., 2003; Kuno et al., 1998, 2005) have been conducted using sequences of individual genes and/or ORF to investigate the flavivirus genetic relationship. These studies generated basically two contrasting phylogenies, NS5 gene tree and NS3/ORF tree. In this study, we tried to determine how the application of concatenated sequences of selected genes would affect phylogeny, by focusing only on antigenically important or phylogenetically more informative genes but by dropping highly variable genes, such as NS2 and NS4. The phylogeny of flaviviruses obtained using the concatenated sequences (E-NS1-NS3-NS5) is similar to the topology previously obtained with NS3 gene or ORF (Billoir et al., 2000; Cook & Holmes, 2005; Kuno & Chang, 2006). Our phylogenetic
results confirmed a close genetic relationship between ROCV and the encephalitogenic JEV group viruses. Using the criteria for group inclusion (amino acid identity $\geq 71\%$), ROCV did not show enough identity to be included either in the JEV or in the Ntaya group, which is represented in our study by NTAV and BAGV because of insufficient (69\%) identity. Furthermore, the phylogeny obtained in this study suggests a strong possibility that the two Brazilian flaviviruses (ROCV and ILHV) segregated earlier independently, suggesting that they represent a lineage distinct from the lineage leading to JEV and Ntaya groups.

Currently, ROCV is classified as a subtype of ILHV (ICTV, 2005). However, traditionally, ROCV has been readily distinguished from ILHV serologically, as the ratios of homologous versus heterologous titres obtained when ROCV was used as antigen or antibody (>640/<10 and 1280/<10, respectively) were far higher and significant than the minimal threshold (>fourfold) used to distinguish flaviviral species (Karabatsos, 1985). In terms of phylogenetic classification, between the two phylogenetically related viruses, ROCV and ILHV, the nucleotide sequence identity (72.5\%) of the full-length NS5 gene nucleotide sequence (Table 2) demonstrated in this study strongly supports recognition of those viruses as two distinct species, if the species criterion (species defined by >84\% nucleotide sequence identity based on the 3′-terminal 1 kb sequence of NS5) previously established for all flavivirus species (Kuno et al., 1998) could be applied to complete NS5 gene sequences as well. Actually, quantitative criteria based on evolutionary distance of a particular gene, such as above, have been used to improve classification of tick-borne flaviviruses (Charrel et al., 2001) and for demarcation of strains and species (Van Regenmortel et al., 1997). Also, biologically, although both ROCV and ILHV are neurotropic, the former has been often associated with epidemics of encephalitis, while the latter has been associated only with sporadic febrile illness with minor neurological abnormalities, clearly reflecting significant differences in neurovirulence and transmissibility (Vasconcelos et al., 1992, 1998). Accordingly, our conclusion differs from the assignment of ROCV as subtype (or genotype) of ILHV by the ICTV (2005) and adds to the list of recent studies in which taxonomic classification of a particular flavivirus was found to be different from that by the ICTV (Grard et al., 2007; Kuno & Chang, 2007; May et al., 2006).

The peculiar vector–virus relationships between Aedes vectors and the viscerotropic flaviviruses (such as DENV and YFV) and between Culex vectors and predominantly neurotropic viruses (such as JEV group viruses) were recognized early on by Sabin (1959). In a more recent phylogenetic study, the relationship was basically confirmed phylogenetically (Gaunt et al., 2001). However, exceptions to this generalization have been reported for several mosquito-borne flaviviruses (Mutebi et al., 2004). Although both ROCV and ILHV are neurotropic and together share a subclade in the phylogenetic tree, unlike the members of the neurotropic JEV group viruses transmitted by Culex mosquitoes, both viruses have been most often isolated from Psorophora mosquitoes (de Souza Lopes et al., 1981; Vasconcelos et al., 1992, 1998). Several other members of the Culex-borne virus clade that do not meet this specific vector–disease syndrome relationship (Gaunt et al., 2001) include Koutango and Saboya viruses that have been isolated more often from ticks and sandflies, respectively, and which are not known to be neurotropic. Nonetheless, the above generalization (Gaunt et al., 2001) can still be applied to ROCV, because it is a neurotropic virus that can replicate in Culex mosquitoes at least under laboratory conditions (Mitchell et al., 1981).

Complete genome sequencing is a pre-requisite to genomic characterization because it provides the critically needed information for an accurate viral classification, for a better choice of informative genes to conduct phylogenetic analysis, and for designing specific and/or broadly reactive diagnostic primers and/or probes in order to improve disease surveillance and diagnosis. Furthermore, the knowledge of the genomic sequence is indispensable for preparing infectious clones for further applications to such investigations as the identification of gene function and of virulence marker and development of antiviral drugs, as well as vaccine development. Our data on ROCV represent the first complete genomic characterization of an indigenous Brazilian flavivirus of public health importance and should be useful for elucidating the genetic relationship with other South American flaviviruses transmitted by mosquitoes, such as Cacipacore virus, ILHV, Bussuquara virus and Iguape virus.

**ACKNOWLEDGEMENTS**

This work is dedicated in memory of Dr Oscar de Souza Lopes who isolated ROCV and contributed greatly to our understanding of the medical importance of the virus and natural transmission cycle. This study was supported by the IEC/SVS and the CNPq grants 300460/2005-8 and 402859/2005-7.

**REFERENCES**


complete coding sequences of arthropod-borne viruses and viruses with no known vector. *J Gen Virol* 81, 781–790.


