J Mol Evol (2001) 52:383–390
DOI: 10.1007/s002390010168

JOURNAL OF MOLECULAR EVOLUTION

Evolution of Base Composition and Codon Usage Bias in the Genus *Flavivirus*

Gareth M. Jenkins,[1] Mark Pagel,[2] Ernest A. Gould,[3] Paolo M. de A. Zanotto,[4] Edward C. Holmes[1]

[1] Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, U.K.
[2] School of Animal and Microbial Sciences, University of Reading, Whiteknights, Reading RG6 6AJ, U.K.
[3] Natural Environment Research Council, Centre for Ecology and Hydrology, Mansfield Road, Oxford OX1 3SR, U.K.
[4] Laboratory of Molecular Evolution and Bioinformatics, Department of Microbiology, ICB II, Universidade de São Paulo, São Paulo, Brazil

Received: 29 August 2000 / Accepted: 19 December 2000

Abstract. The extent to which base composition and codon usage vary among RNA viruses, and the possible causes of this bias, is undetermined in most cases. A maximum-likelihood statistical method was used to test whether base composition and codon usage bias covary with arthropod association in the genus *Flavivirus*, a major source of disease in humans and animals. Flaviviruses are transmitted by mosquitoes, by ticks, or directly between vertebrate hosts. Those viruses associated with ticks were found to have a significantly lower G+C content than non-vector-borne flaviviruses and this difference was present throughout the genome at all amino acids and codon positions. In contrast, mosquito-borne viruses had an intermediate G+C content which was not significantly different from those of the other two groups. In addition, biases in dinucleotide and codon usage that were independent of base composition were detected in all flaviviruses, but these did not covary with arthropod association. However, the overall effect of these biases was slight, suggesting only weak selection at synonymous sites. A preliminary analysis of base composition, codon usage, and vector specificity in other RNA virus families also revealed a possible association between base composition and vector specificity, although with biases different from those seen in the *Flavivirus* genus.

Correspondence to: Dr. E.C. Holmes; e-mail: Edward.Holmes@zoo.ox.ac.uk

Key words: RNA virus — Flaviviruses — Base composition — G+C content — Codon usage

Introduction

It is well established that base composition and codon usage vary within and among organisms. For example, the genomic G+C content of prokaryotes ranges from 0.25 to 0.75, and, in some of these organisms, mutation pressure has been documented as the main determinant of this variation as well as differences in codon usage (Sharp et al. 1993). Similarly, codon usage in vertebrate genomes reflects the base composition of the local isochore, where the G+C content may vary from 0.3 to 0.65 (Bernardi et al. 1985), and therefore is also likely to reflect regional differences in mutation pressure (Wolfe et al. 1989). Yet, in most cases, the origin of such uneven base composition and codon usage bias is uncertain. In particular, codon usage may also be determined by translational selection rather than by mutation pressure alone. In *Escherichia coli*, *Saccharomyces cerevisiae*, and *Drosophila*, for example, highly expressed genes have a strong selective preference for codons with a high concentration of the corresponding acceptor tRNA molecule, whereas genes expressed at a lower level display a more uniform pattern of codon usage (Grantham et al. 1981; Gouy and Gautier 1982; Sharp et al. 1986; Powell and Moriyama 1997). This may itself affect base composition if all optimal codons terminate with the same nucleotide,

as is the case for *Drosophila*, where optimal codons end in G or C (Shields et al. 1988). Finally, base composition and codon usage may also be influenced by dinucleotide preferences. As a case in point, CpG doublets in vertebrates occur at one-fifth of the expected frequency due to methylation (Bird 1986), and this doublet is now known to play a major role in determining chromatin structure and gene activation (Kundu and Rao, 1999).

Far less attention has been paid to revealing the determinants of base composition and codon usage in viruses, particularly RNA viruses, although these may be useful indicators of the evolutionary forces which shape genomes. For example, human immunodeficiency virus (HIV) and other lentiviruses show a strong preference for the A nucleotide, which may reach frequencies of up to 40% (Hemert and Berkhout 1995). Inefficiently translated codons are also prevalent, possibly helping these viruses minimize their antigenicity and hence establish persistent infections (Haas et al. 1996). Furthermore, while pneumoviruses are associated with a strikingly low G+C content (Pringle and Easton 1997), a high G+C content is a shared property of viroids, satellite RNAs, and hepatitis delta virus (Branch et al. 1990). Also, nearly all RNA viruses are deficient in the dinucleotide CpG, although the reasons why are uncertain (Karlin et al. 1993).

A more detailed knowledge of base composition and codon usage in RNA viruses would be desirable for many reasons. First, it is relevant for determining whether synonymous substitutions are neutral in these infectious agents as has been proposed previously (Gojobori et al. 1990). If so, rates of synonymous substitution will reflect the underlying rates of mutation and replication rather than any selective requirements. Second, patterns of base composition and codon usage may provide information on viral pathogenesis. For example, HIV has a highly skewed codon composition in variable regions of the surface glycoprotein, which has been proposed to be caused by insertion of AAT triplets, thereby assisting the virus in escaping the host immune response (Bosch 1994). Third, such information may be useful in elucidating the evolutionary origins of viruses if specific patterns of codon usage and base composition can be associated with the patterns favored by particular host species. Fourth, codon usage may reveal strategies for regulating viral gene expression and is thus relevant to vaccine design (Haas et al. 1996). Finally, fluctuations in base composition may adversely affect phylogenetic reconstruction since most conventional methods assume a stationary base composition across the tree (Lockhart et al. 1994).

The focus of this study is the evolution of base composition and codon usage in the genus *Flavivirus*, a member of the family Flaviviridae. To date, this genus includes some 70 single-strand positive sense RNA viruses (~12 kb in length) which have a variety of geographical ranges and principal vertebrate hosts including rodents, bats, birds, and primates (Monath and Heinz 1996). Many are associated with human disease, most notably yellow fever virus, Japanese encephalitis virus, West Nile virus, and dengue virus. The genus *Flavivirus* can be divided into two monophyletic groups: non-vector-borne and vector-borne flaviviruses (Kuno 1998). The former are transmitted directly between vertebrates and replicate solely in these hosts. The latter are transmitted indirectly via mosquitoes or ticks and replicate in both insects and vertebrates; the mosquito-borne and tick-borne flaviviruses also form separate phylogenetic clusters. Given their diverse survival strategies, flaviviruses provide an ideal opportunity to study the possible effects of genetic and ecological factors on base composition and codon usage in viruses, particularly since insect and vertebrate host cells are likely to provide very different environments. Here, a comparative analysis was undertaken to test whether base composition and codon usage covary with arthropod association and to identify the possible causes of any such correlations. An analysis of base composition, codon usage, and vector specificity in other RNA virus families was also performed.

Materials and Methods

Sequences

Seventy-three partial NS5 gene sequences (~1020 bp) were used for the comparative analysis of base composition, codon usage, and arthropod association in the genus *Flavivirus*, (Kuno et al. 1998). Forty-two viruses were mosquito-borne, 18 tick-borne, and 13 non-vector-borne. Thirteen complete *Flavivirus* coding regions were also analyzed. Patterns of base composition and codon usage were also investigated in the structural polyprotein of 9 members of the Togaviridae and the M segment of 38 members of the Bunyaviridae, two other families of RNA viruses. A complete list of sequences and accession numbers is available at http://evolve.zoo.ox.ac.uk/alignments.

Phylogeny Reconstruction

Sequences were aligned using CLUSTALW (Thompson et al. 1994) and corrected manually. Third codon positions and a small region of extensive sequence divergence (nucleotide positions 530–560), which was difficult to align, were removed prior to phylogenetic analysis resulting in a final alignment of 682 bp. The maximum-likelihood phylogeny was reconstructed using a HKY-Γ model of nucleotide substitution (Hasegawa et al. 1985; Yang 1993) using PAUP* version 4 (Swofford 1998) (likelihood parameters available upon request). To assess the degree of support for key nodes on the tree, a bootstrap resampling analysis was undertaken using 1000 replicate neighbor-joining trees with input distances estimated under the maximum-likelihood substitution model.

Comparative Analysis

G+C content, vector specifity, codon usage bias, and dinucleotide bias were analyzed using a set of maximum-likelihood statistical methods

(Pagel 1998, 1999) that control for phylogenetic associations among viral lineages (the methods are implemented in the program CONTINUOUS, available from Mark Pagel upon request). These methods presume a constant-variance (Brownian motion) model of trait evolution.

Trait Evolution. The statistical methods allow us to estimate three parameters, λ, κ, and δ, that test important details about the tempo and mode of trait evolution on the phylogeny (Pagel 1998, 1999). The parameter λ tests whether the patterns of similarities and differences among viral lineages are those that would be expected from their phylogenetic associations. The default value of λ is 1.0 (where the evolution of the trait is dependent on phylogenetic history). However, if a trait has evolved such that it is independent of the phylogeny, then $\lambda = 0$, and little if any phylogenetic correction of the data is called for. The parameters κ and δ test whether the trait is evolving in direct proportion to the lengths of the branches of the phylogenetic tree. Values of $\kappa < 1$ indicate that the trait evolves proportionally more in shorter branches, whereas values of $\kappa > 1$ suggest that traits evolve proportionally more in longer branches. δ is estimated from the total path length from the root of the tree to each tip and is useful for testing whether the rate of evolution has changed throughout the history of the phylogeny. Values of $\delta < 1.0$ are indicative of adaptive radiations in which most evolutionary change occurs early in the phylogeny; values of $\delta > 1.0$ suggest that later, species-specific evolution has dominated. When these parameters differ from 1.0, scaling the branch lengths in accordance with their values improves the statistical test.

Differences Among Vector Groups. To test for differences in G+C content or other variables among the three vector groups, a standard technique in regression analysis known as "dummy" coding was used. This involved coding each virus using two binary variables to identify each vector group uniquely. For example, to test whether the base composition of non-vector-borne viruses is significantly different from those of the other two vector groups, all mosquito-borne viruses were coded as on the first variable and 0 on the second, tick-borne viruses as 0, 1, and non-vector-borne viruses as 0, 0. The partial correlation between G+C content and the first binary variable, holding constant the second variable, denoted $r_{G+C\,1\cdot 2}$, tests whether the group coded 1 on the first vector differs from the group coded 0 on this vector (here mosquito-borne versus non-vector). Similarly, the partial correlation between G+C content and the second binary variable, holding constant the first variable ($r_{G+C\,2\cdot 1}$) tests whether tick-borne viruses differ in G+C content from non-vector-borne viruses. To test whether the base compositions of the mosquito-borne and tick-borne groups differ from each other, the same procedure was repeated with a different coding (e.g., mosquito-borne viruses as 1, 0, tick-borne viruses as 0, 0, and non-vector-borne viruses as 0, 1).

Directional Trends in G+C Content. The possibility that G+C content has evolved directionally along the phylogeny was also tested. The CONTINUOUS method allows one to fit a "directional" model (Pagel, 1999) of trait evolution and compare it to the standard "random-walk" or nondirectional model of trait evolution using a likelihood-ratio test.

Ancestral State Reconstruction. Ancestral G+C content was estimated using a Brownian motion model implemented in the program ANCML (Schluter et al. 1997). This method assumes the parameters λ, κ, and δ to be 1.0, so where these parameters deviated from 1.0, the branch lengths of the phylogeny were adjusted accordingly. This model should perform well in the absence of any directional trends along the phylogeny, since the G+C content represents the average outcome to a large number of binomial events and would therefore be expected to evolve according to a random walk.

Dinucleotide and Codon Usage Analysis

The effective codon usage statistic, N_C, was used to measure overall codon bias (Wright 1990). The reported value of N_C is always between 20 (when only one codon is effectively used for each amino acid) and 61 (when codons are used randomly). Codon usage for individual amino acids was compared by measuring the G+C content at the third codon position for each degenerate amino acid separately, and the significance of differences between viruses was evaluated using a χ^2 statistic. To test whether uneven base composition accounts for codon usage bias in flaviviruses, the pattern of codon usage bias for each *Flavivirus* polyprotein was compared with 1000 randomized sequences of the same length, where each randomized sequence had the same overall base composition as that found at fourfold degenerate third codon (i.e., synonymous) positions in the original sequence. This approach was used because synonymous positions are better indicators of the extent of base composition bias than nonsynonymous positions, and the measure of codon usage bias employed is independent of the base composition at nonsynonymous positions since it corrects for uneven amino acid usage.

Dinucleotide frequencies were also compared with the values expected under the null hypothesis of no dinucleotide bias and the significance tested by approximating the null distribution for each doublet to a normal distribution with mean p and variance $p\,(1 - p)/n$, where p is the expected frequency and n the total number of doublets. Finally, the effect of dinucleotide bias on codon usage was assessed by comparing the effective codon usage statistic for each *Flavivirus* polyprotein with 1000 randomized sequences of the same length, where each randomized sequence had the same overall dinucleotide composition as that found in doublets terminating at the third codon position in the original sequence. Here, all third codon positions were used since fourfold third codon positions alone do not include all possible doublets.

Results

A preliminary analysis of base composition in the genus *Flavivirus* revealed a striking pattern of variation among different arthropod groups. Non-vector-borne viruses have a low G+C content (0.466 ± 0.006), mosquito-borne viruses an intermediate G+C content (0.502 ± 0.003), and tick-borne viruses a high G+C content (0.542 ± 0.003). A similar pattern was observed for G+C content at first and second codon positions alone ($G+C_{12}$) and at synonymous third codon positions ($G+C_{3S}$). The mean values of G+C, $G+C_{12}$, and $G+C_{3S}$ for each vector group were highly significantly different (t test: $P < 0.0001$), and the range of variation was greatest at synonymous third codon positions (0.394 ± 0.01 for non-vector-borne viruses, 0.575 ± 0.008 for tick-borne viruses).

Figure 1 shows the maximum-likelihood phylogeny of the genus *Flavivirus*. Since each vector group is monophyletic, and base composition would be expected to be more similar among closely related species, we then compared G+C content using a method which controls for phylogenetic nondependence.

First, to test the nature of G+C trait evolution in this group of viruses, the parameters λ, κ, and δ were estimated (Table 1). The high values of λ indicate that there exists a strong correlation between similarities in G+C content and phylogenetic relatedness. Similarly, the values of δ indicate that the rate of G+C content evolution has remained stable over the time scale represented by the phylogeny. However, the upper confidence limit for κ was less than 1.0 for $G+C_{12}$ and $G+C_{3S}$. This indicates

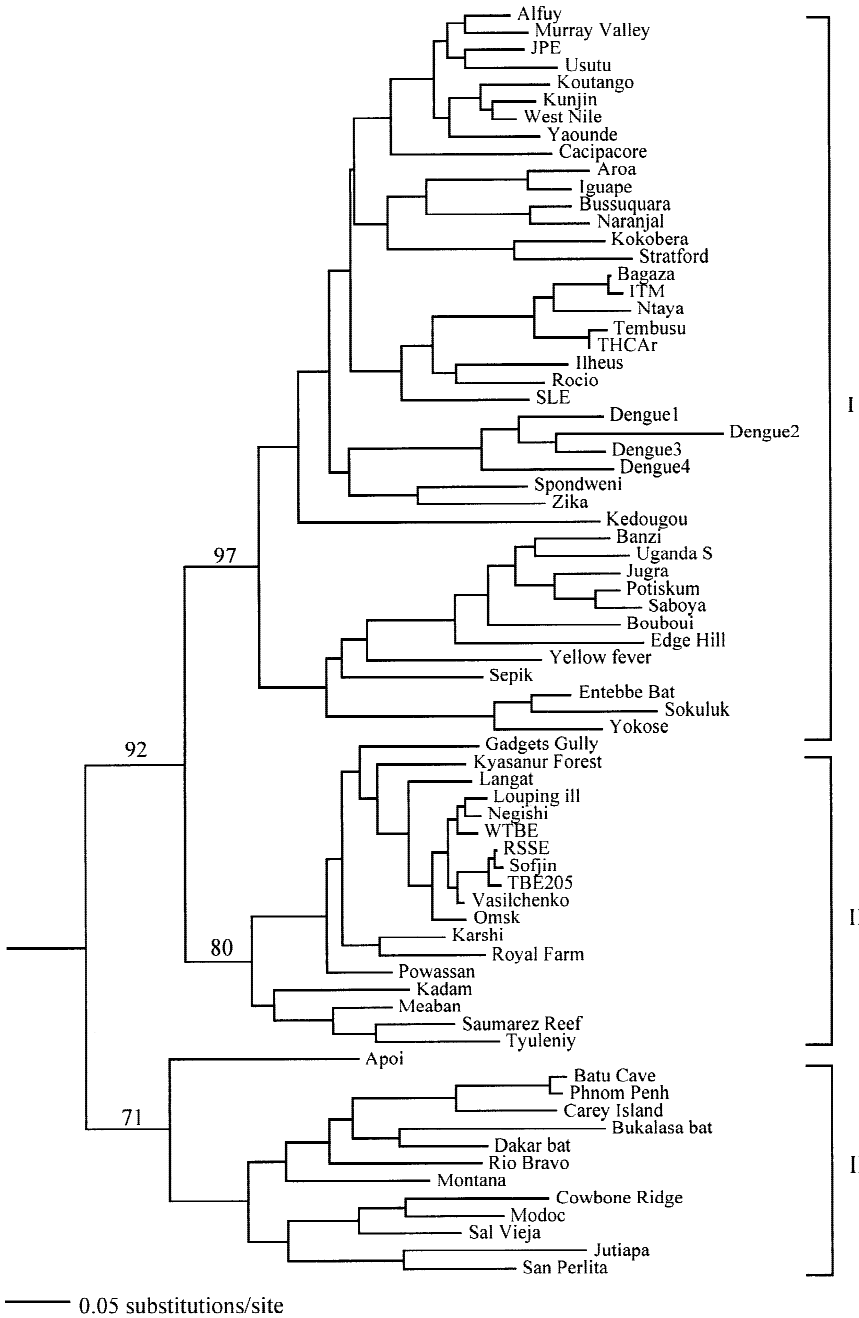

Fig. 1. Maximum-likelihood phylogenetic tree of the genus *Flavivirus* rooted using CFA (cell fusing agent) as an outgroup. Clades I, II, and III correspond to mosquito-borne, tick-borne, and non-vector-borne viruses, respectively, and numbers indicate neighbor-joining bootstrap values of critical nodes. JPE, Japanese encephalitis; ITM, Israel Turkey meningoencephalitis; SLE, St. Louis encephalitis; THCAr, isolate from Thailand; WTBE, Western tick-borne encephalitis; RSSE, Russian spring summer encephalitis; TBE205, tick-borne encephalitis strain 205.

a significant tendency for the G+C content to evolve disproportionately more in shorter branches for these site categories. This might arise if the G+C content quickly reaches an equilibrium value along internal branches of the tree following speciation events (or the changes in selective regimes they may bring about), then remains comparatively stable. We therefore scaled the branch lengths using κ in all of our tests of correlations between $G+C_{12}$ and $G+C_{3S}$ content and arthropod specificity and in our reconstructions of ancestral states.

The partial correlation tests found G+C and $G+C_{3S}$ content to be significantly different between non-vector-borne and tick-borne viruses ($P < 0.05$). None of the other comparisons were significant. To determine which groups gained or lost G+C content since their divergence from a common ancestor, the ancestral G+C contents of the complete *Flavivirus* genus and each individual vector group were estimated (Table 2). Unfortunately, since the confidence intervals for the ancestral G+C content of the whole *Flavivirus* genus overlap with those of each individual vector group, it is not possible to conclude which groups gained or lost G+C content. The only confidence intervals which do not overlap are the overall G+C content and $G+C_{3S}$ of the non-vector-borne and tick-borne groups, consistent with the previous findings that only these groups differ significantly in G+C content. The lack of any clear evidence from the ancestral state reconstructions for a directional trend of G+C content evo-

Table 1. Estimates of λ, κ, and δ for G+C, $G+C_{12}$, and $G+C_{3S}$[a]

Trait	λ	κ	δ
G+C	0.978 (0.876, 1.00)	0.642 (0.190, 1.07)	0.801 (0.169, 1.52)
$G+C_{12}$	0.967 (0.856, 1.00)	0.495 (0, 0.902)*	1.37 (0.471, 2.31)
$G+C_{3S}$	0.960 (0.807, 1.00)	0.589 (0.175, 0.973)*	0.798 (0.170, 1.51)

[a] Numbers in parentheses refer to 0.95 confidence intervals, and asterisks indicate values that do not overlap with 1.0.

lution is also supported by our test of directional trait evolution. The likelihood of the random-walk and directional models did not differ significantly for either the complete *Flavavirus* genus or each vector group separately ($P > 0.05$). We also obtained similar ancestral G+C estimates using the maximum-likelihood method of Galtier and Gouy (1998) (0.484, 0.475, 0.501, and 0.533 for the whole genus, non-vector-borne, mosquito-borne, and tick-borne viruses, respectively), although in this case it was only computationally possible to analyze each vector group separately and to use a pruned data set (40 taxa) to estimate the ancestral G+C content of the whole genus.

To identify possible causes for this variation in base composition, we first tested whether the pattern of base composition is unique to the NS5 gene or occurs throughout the whole *Flavivirus* polyprotein. Our analysis of 13 complete polyprotein sequences revealed the same pattern of base composition throughout the viral genome, although only a single non-vector-borne virus was available for analysis. Next, overall codon usage was compared among vector groups, as this is an obvious reflection of differences in base composition. The mean effective codon usage indices for non-vector-borne, mosquito-borne, and tick-borne viruses were 51.8 ± 0.6, 52.8 ± 0.4, and 54.3 ± 0.7, respectively. Unlike for G+C content, these differences were not significant controlling for phylogenetic nonindependence ($P > 0.05$).

Codon usage for individual amino acids was then compared between Rio Bravo virus, the only non-vector-borne virus for which a complete sequence is available, and the complete coding regions of other flaviviruses (Table 3). Whole polyproteins were used for this analysis since the NS5 gene is too short to compare codon usage for individual amino acids between different viruses. A comparison of Rio Bravo virus against four tick-borne viruses showed that the $G+C_{3S}$ content was less in the Rio Bravo genome for all amino acids encoded by two or more codons. Most of these differences were significant ($P < 0.05$), indicating that many synonymous codons contribute to the overall difference in G+C content between non-vector-borne and tick-borne flaviviruses. Similarly, when Rio Bravo was compared against eight mosquito-borne viruses, the majority of degenerate amino acids also had a lower G+C in the Rio Bravo genome, although fewer differences were significant, which is probably due to the lower overall difference in G+C content between these groups. A highly significant correlation was also found between $G+C_{12}$ and $G+C_{3S}$ in the NS5 gene considered alone and controlling for phylogenetic nonindependence ($r = 0.436$, $P < 0.0001$). Altogether, these results reveal that the variation in base composition is due to a general difference in codon usage across all triplets, rather than at one or a few specific codons.

We then asked whether codon usage bias in flaviviruses is simply due to uneven base composition? The effective codon usage for all complete *Flavivirus* polyproteins was significantly lower than in randomized sequences of the same synonymous base composition ($P < 0.05$) (Table 4), indicating that factors other than base composition influence the usage of synonymous codons in these viruses to some extent. Since CpG dinucleotide deficits have been reported in flaviviruses (Weaver et al. 1993; Karlin et al. 1994; Kuno et al. 1998), a comparison of actual and expected dinucleotide frequencies was also undertaken. Only the frequencies of CpG and UpG were significantly different at all codon positions for all complete polyproteins ($P < 0.05$), where CpG was present at 0.5 the expected frequency and UpG was in excess by a corresponding amount (Table 4). Furthermore, no significant variation in dinucleotide biases was detected among the three vector groups controlling for phylogenetic nonindependence ($P > 0.05$). Finally, we compared codon usage in flaviviruses with randomized sequences of the same dinucleotide composition. For the non-

Table 2. Ancestral G+C content for different arthropod groups[a]

Group	G+C	$G+C_{12}$	$G+C_{3S}$
All flaviviruses	0.497 (0.472, 0.522)	0.491 (0.477, 0.505)	0.478 (0.423, 0.533)
Mosquito-borne	0.507 (0.489, 0.524)	0.486 (0.476, 0.497)	0.513 (0.472, 0.555)
Tick-borne	0.526 (0.510, 0.543)	0.499 (0.489, 0.509)	0.543 (0.504, 0.582)
No known vector	0.485 (0.463, 0.506)	0.489 (0.478, 0.501)	0.445 (0.400, 0.491)

[a]Numbers in parentheses refer to 0.95 confidence intervals.

Table 3. Number of degenerate amino acids for which $G+C_{3S}$ is less in Rio Bravo, a non-vector-borne flavivirus, than in other selected vector-borne flaviviruses[a]

Virus	Number of amino acids
Tick-borne	
Louping ill	18 (16)
Vasilchenko	18 (17)
Western tick-borne encephalitis	18 (17)
Powassan	18 (13)
Mosquito-borne	
Dengue type 1	14 (8)
Dengue type 2	15 (4)
Dengue type 3	13 (6)
Dengue type 4	16 (10)
West Nile	18 (17)
Japanese encephalitis	18 (16)
Murray Valley	16 (8)
Yellow fever	18 (12)

[a] Numbers in parentheses indicate the number of amino acids for which the difference in $G+C_{3S}$ is significant ($P < 0.05$).

vector-borne and mosquito-borne viruses, the fraction of randomizations for which codon usage bias was higher than in the original sequence was still less than 0.05, indicating that uneven nucleotide and dinucleotide composition do not account for all of the codon usage bias in these viruses (Table 4).

Finally, to determine whether the relationship between base composition and vector specificity is general for RNA viruses, we conducted a preliminary analysis of G+C content, codon usage and arthropod association in other RNA virus families for which differences in vector specificity exist. A comparison of rubella virus, a non-vector-associated member of the Togaviridae, with alphaviruses, a genus of mosquito-transmitted togaviruses, showed the non-vector-associated virus to have a much higher $G+C_{3S}$ content (0.80) and, consequently, a relatively low effective codon usage (40.2) (Fig. 2). This contrasts with the genus *Flavivirus*, where a low $G+C_{3S}$ content was a characteristic of non-vector-borne viruses. Base composition and codon usage were also compared among bunyaviruses, nairoviruses, and hantaviruses, which are mosquito-borne, tick-borne, and non-vector-borne members of the family Bunyaviridae (Fig. 2). As in the genus *Flavivirus*, tick-borne viruses have a higher $G+C_{3S}$ content, although unlike before, the genomes of these Bunyaviridae are A+U rich. Overall, this analysis suggests a possible association between base composition and vector specificity in other RNA virus families, although the biases appear to be different to those observed in the genus *Flavivirus.*

Discussion

Our results document for the first time a relationship between base composition and vector specificity in RNA viruses in that non-vector-borne flaviviruses have a significantly lower G+C content than those transmitted by ticks. Since this difference appears across all amino acids and codon positions, it is unlikely to be due solely to biases in dinucleotide composition or selection for specific codons but, rather, to a more general difference in nucleotide usage. Despite this variation in base composition, we are confident in our reconstruction of the phylogenetic relationships of the flaviviruses, since only first

Table 4. Codon usage and dinucleotide bias

Virus	N_C[a]	N_C[b]	N_C[c]	CpG[d]	UpG[d]
No known vector					
Rio Bravo	50.22	54.62	51.88 (0.002)	0.35	1.49
Mosquito-borne					
Dengue type 1	50.05	54.64	52.97 (0)	0.45	1.40
Dengue type 2	48.25	54.12	51.45 (0)	0.41	1.43
Dengue type 3	49.63	55.12	52.56 (0)	0.40	1.38
Dengue type 4	50.78	56.93	53.44 (0)	0.37	1.38
West Nile	53.76	59.94	55.34 (0.006)	0.58	1.44
Japanese encephalitis	55.78	59.84	56.84 (0.017)	0.60	1.36
Murray Valley	53.63	59.41	56.06 (0)	0.50	1.45
Yellow fever	53.14	60.81	54.50 (0.011)	0.39	1.49
Tick-borne					
Powassan	54.86	60.22	55.64 (0.084)	0.51	1.42
Western tick-borne encephalitis	54.49	60.01	55.33 (0.073)	0.54	1.47
Louping ill	53.89	59.78	54.76 (0.087)	0.56	1.45
Western tick-borne encephalitis	54.37	59.21	55.10 (0.124)	0.57	1.47

[a] Actual codon usage.

[b,c] Expected codon usage correcting for uneven base and dinucleotide composition, respectively. Numbers in parentheses refer to the fraction of randomizations for which codon usage bias was greater than for the actual sequence. For N_C[b], all 1000 randomizations had a lower codon usage bias than the initial sequence.

[d] Actual/expected doublet frequency across the entire polyprotein ($P < 0.000001$).

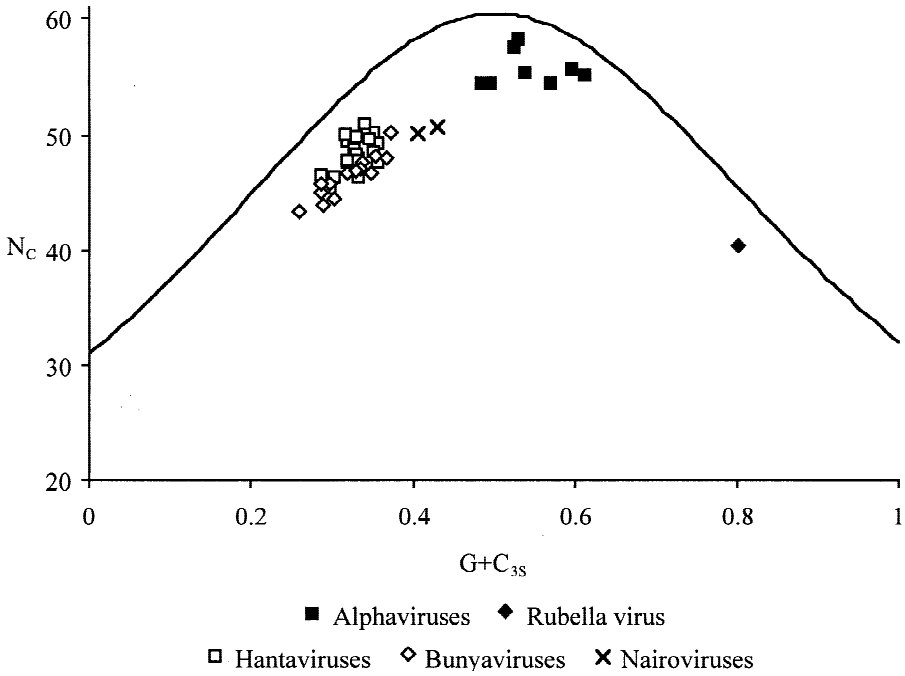

Fig. 2. G+C content at synonymous third codon positions ($G+C_{3S}$) and effective codon usage (N_C) in the families Togaviridae (alphaviruses and rubella virus) and Bunyaviridae (hantaviruses, bunyaviruses, and nairoviruses). The curve indicates the expected codon usage if G+C compositional constraints alone account for codon usage bias (Wright 1990).

and second position codons were used to build these and base composition fluctuates less at these sites. Indeed, a similar phylogeny was obtained using a LogDet method (Lockhart 1994), which avoids the problem of variable base composition (results available upon request).

Both neutralist and selectionist theories could be invoked to explain these taxon-specific differences in base composition. For example, the variation might reflect adaptation to particular environmental variables such as temperature, humidity, and chemical environments present in vertebrate hosts or vectors. It could also be that the vector groups have evolved different modes of pathogenesis (e.g., different sites of primary or secondary replication) in their natural vertebrate host which might affect base composition. However, as each vector group is associated with a similar range of natural vertebrate hosts, it is unlikely that the variation in base composition is due to differences in vertebrate host specificity. Furthermore, given that different biases in base composition are observed in other vector-borne viruses from the Togaviridae and Bunyaviridae (although these viral families are more genetically diverse than the flaviviruses), it seems equally unlikely that adaptation to vector cells can explain the patterns seen. Consequently, differences in mutation bias among flaviviruses are the most likely explanation for variable base composition given current evidence.

Since all three families of RNA viruses replicate in the cytosol using their own polymerase, distinct mutational biases could be explained by differences in their viral polymerase or in the cytosolic ribonucleotide pools of their target cells. Unfortunately, such critical biochemical information is unavailable. In contrast, mutational biases in DNA viruses would be more dependent on host enzymes and nuclear deoxyribonucleotide pools, since these viruses generally replicate in the nucleus using host-derived polymerases. Similarly, mutational biases in retroviruses such as HIV could be due to both cytoplasmic and nuclear nucleotide pools, or to both viral and host enzymes, since these viruses replicate first in the cytoplasm using virus-encoded reverse transcriptase and then in the nucleus with cellular RNA polymerase II.

Our study also revealed biases in dinucleotide and codon usage, although these did not covary with arthropod association. CpG doublets are notably deficient, and UpG doublets correspondingly in excess in all flaviviruses, although CpA doublets, which are also in excess in vertebrate DNA due to the double-strand symmetry of DNA, were not in excess here. One explanation for these dinucleotide biases is methylation mutation of CpG to UpG, analogous to the process that occurs in vertebrate DNA, where CpG mutates to TpG. This raises the possibility that CpG methylation takes place in RNA as well as in DNA, although methylation of RNA is a barely explored area. Since CpG methylation does not occur to the same extent in invertebrate DNA, it is interesting that no significant variation in CpG frequency was detected among vector and non-vector-borne flaviviruses. A critical question in this context is therefore whether CpG suppression or UpG overrepresentation is an adaptive or passive process. Furthermore, given that uneven base and dinucleotide composition does not account for all of the codon usage bias in these viruses, it is possible that some weak selection for optimal translation may also occur. This is clearly a possibility that needs to be explored further, although our understanding of the importance of the host cell in determining the evolution of these viruses is limited by an incomplete knowledge of

codon usage in different species of mosquitoes, ticks, and several of their vertebrate hosts.

While it is clear that synonymous sites in flaviviruses are subject to a variety of constraints, effective codon usage is only approximately one-quarter less than its maximum value for all flaviviruses, so the overall effect of the biases we have documented must be slight. Even rubella virus, which has a G+C content of 0.80 at synonymous third codon positions, still effectively uses 40.2 of a possible 61 codons, although RNA secondary structure may also constrain synonymous sites in some cases (Simmonds and Smith 1999). Clearly, a more comprehensive analysis is needed to reveal the true extent of codon usage bias variation within and among RNA viruses and what factors are responsible, including the influence of factors such as cell tropism, principal host species, method of transmission, and viral genetic structure. Such information would then allow us to judge more precisely the relative importance of mutation pressure versus natural selection in determining base composition and codon usage in these pathogens.

Acknowledgments. This work was supported by research grants from the Royal Society, the Wellcome Trust, and CNPq of Brazil.

References

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Bird AP (1986) CpG-rich islands and function of DNA methylation. Nature 321:209–213

Bosch ML, Andeweg AC, Schipper R, Kenter M (1994) Insertion of N-linked glycosylation sites in the variable regions of the human immunodeficiency virus type 1 surface glycoprotein through AAT triplet reiteration. J Virol 68:7566–7567

Branch AD, Levine BJ, Robertson HD (1990) The brotherhood of circular RNA pathogens, viroids, circular satellites, and the delta agent. Semin Virol 1:143–152

Galtier N, Gouy M (1998) Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol Biol Evol 15: 871–879

Gojobori T, Moriyama EN, Kimura M (1990) Molecular clock of viral evolution, and the neutral theory. Proc Natl Acad Sci USA 87: 10015–10018

Gouy M, Gautier C (1982) Codon usage in bacteria: Correlation with expressivity. Nucleic Acids Res 12:539–549

Grantham R, Gautier C, Gouy M, Jacabzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 10:7055–7074

Hasegawa M, Kinshino H, Yano TA (1985) Dating of the human-ape splitting by a molecular clock of a mitochondrial DNA. J Mol Evol 22:160–174

Hass J, Park E, Seed B (1996) Codon usage limitation in the expression of HIV-1 envelope glycoprotein. Curr Biol 6:315–324

Hemert FJ, Berkhout B (1995) The tendency of lentiviral open reading frames to become A-rich, constraints imposed by viral genome organisation and cellular tRNA availability. J Mol Evol 41:132–140

Johnston J (1963) Econometric methods. McGraw–Hill, New York

Karlin S, Doerfler W, Cardon LR (1994) Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not those of large eukaryotic viruses? J Virol 68:2889–2897

Kundu TK, Rao MRS (1999) CpG islands in chromatin organisation and gene expression. J Biochem 125:217–222

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

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

Monath TP, Heinz FX (1996) Flaviviruses. In: Fields BN, Knipe DM (eds) Virology. Raven Press, New York, pp 961–1034

Pagel M (1998) Inferring evolutionary processes from phylogenies. Zool Scripta 26:331–348

Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401:877–884

Powell JR, Moriyama EN (1997) Evolution of codon usage bias in *Drosophila.* Proc Natl Acad Sci USA 94:7784–7790

Pringle CR, Easton AJ (1997) Monopartite negative strand RNA genomes. Semin Virol 8:49–57

Schluter D, Price T, Mooers AØ, Ludwig D (1997) Likelihood of ancestor states in adaptive radiation. Evolution 51:1699–1711

Sharp PM, Touhy TMF, Mosurski KR (1986) Codon usage in yeast, cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 14:5125–5143

Sharp PM, Stenico M, Peden JF, Lloyd AT (1993) Codon usage: Mutational bias, translational selection, or both? Biochem Soc Trans 21:835–841

Shields DC, Sharp PM, Higgins DG, Wright F (1988) 'Silent' sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Simmonds S, Smith DB (1999) Structural constraints on RNA virus evolution. J Virol 73:5787–5794

Swofford DL (1998) PAUP*. Phylogenetic analysis using parsimony (* and other methods), Version 4. Sinauer Associates, Sunderland, MA

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Weaver SC, Hagenbragh A, Bellew LA, Netesov SV, Volchkov VE, Chang GJ, Clarke DK, Gousset L, Scott TW, Trent DW, Holland JJ (1993) A comparison of the nucleotide sequences of eastern and western equine encephalomyelitis viruses with those of other alphaviruses and related RNA viruses. Virology 197:375–390

Wolfe K, Sharp PM, Li WH (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Wright F (1990) The effective number of codons used in a gene. Gene 87:23–29

Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396–1401