G+C content differs in conserved and variable amino acid residues of flaviviruses and other evolutionary groups.

Raphaëlle Klitting, Ernest Gould, Xavier De Lamballerie

To cite this version:
Research paper

G + C content differs in conserved and variable amino acid residues of flaviviruses and other evolutionary groups

Raphaëlle Klitting a,⁎, Ernest Andrew Gould a, Xavier de Lamballerie a,b

a UMR "Emergence des Pathologies Virales" (EPV: Aix-Marseille University - IRD 190 - Inserm 1207 - EHESP), 27 bd Jean Moulin, 13385 Marseille, France
b Institut Hospitalo-Universitaire Méditerranée-Infection, 27 bd Jean Moulin, 13385 Marseille, France

Article history:
Received 8 July 2016
Received in revised form 1 September 2016
Accepted 19 September 2016
Available online 20 September 2016

Keywords:
Flavivirus
Evolution
G + C content

Abstract
Flaviviruses are small RNA viruses that exhibit genetic and ecological diversity and a wide range of G + C content (GC%). We discovered that, amongst flaviviruses, the GC% of nucleotides encoding conserved amino acid (AA) residues was consistently higher than that of nucleotides encoding variable AAs. This intriguing phenomenon was also identified for a wide range of other viruses, and some non-viral evolutionary groups. Here, we analyse the possible mechanisms underlying this imbalanced nucleotide content (in particular the role of the specific G content and the AA composition in flaviviral genomes) and discuss its evolutionary implications. Our findings suggest that one of the most simple characteristics of the genetic code (i.e., the G or G + C content of codons) is linked with the evolutionary behavior of the corresponding encoded AAs.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Flaviviruses constitute a genetically and ecologically diverse group of viruses some of which are arthropod-borne human pathogens (arbo-viruses) including yellow fever virus (YFV), dengue virus (DENV), West Nile virus (WNV), Japanese encephalitis virus (JEV), tick-borne encephalitis virus (TBEV) and recently emerging flaviviruses such as Zika virus. Other flaviviruses are transmitted by different and not necessarily defined routes, some of which exclusively infect arthropods. Flaviviruses are intimately linked with the history of virology, because YFV and DENV were amongst the first human pathogenic viruses of vertebrates to be isolated (Calisher and Gould, 2003). Moreover, the YFV live-attenuated 17D vaccine which was first used as early as 1938 in Brazil (Smith et al., 1938), and modified versions of this vaccine, are now some of the safest, most efficacious and successful viral vaccines ever produced.

Thus, in view of their importance as human pathogens and their genetic diversity their evolution has been extensively studied, taking advantage of the availability of a large corpus of genomic sequences. A wide range of genomic G + C content (GC%) has been reported within the genus Flavivirus (Kuno et al., 1998; Gaunt et al., 2001; Jenkins et al., 2001; de Lamballerie et al., 2002; Jenkins and Holmes, 2003; Cook and Holmes, 2006; Lobo et al., 2009; Schubert and Putonti, 2010; Kitchen et al., 2011; Belalov and Lukashev, 2013; Moureau et al., 2015). At an early stage of investigation, a possible association between nucleotide base composition and vector/host specificity was identified (Jenkins et al., 2001; de Lamballerie et al., 2002), and lineage-specific co-evolutionary processes between the viral and host/vector groups were proposed (Lobo et al., 2009). However, little is known about the mechanisms that drive or constrain flaviviral genetic evolution or even the factors that determine their GC% at the molecular level.

Here, in an initial molecular analysis of dengue 2 virus (DENV2), we detected higher GC% in sylvatic than non-sylvatic DENV2 isolates. By looking in detail at the differences between sylvatic and non-sylvatic sequences, we made a remarkable observation: when compared with variable amino acid residues, G + C content was strikingly higher in sequence encoding conserved amino acid residues in both sylvatic and non-sylvatic isolates.

***

Abbreviations: AA, Amino Acid; APOIV, Apoi virus; CA-Cons, Complete Alignment Conserved; CA-Var, Complete Alignment Variable; CDS, complete coding sequences; CXPV, Culex-borne flavivirus; DENV, dengue virus; GC%, G + C content; ISFV, insect specific flaviviruses; JEV, Japanese encephalitis virus; MBFV, mosquito-borne flaviviruses; MOOD, Modoc virus; MMLV, Montana Myotis Leukoencephalitis virus; NWV, viruses with no known vector; ORF, open reading frame; PW, pairwise; RBV, Rio Bravo virus; TBEV, tick-borne encephalitis virus; TBFV, tick-borne flaviviruses; WNV, West Nile virus; YFV, yellow fever virus.

⁎ Corresponding author.
E-mail addresses: raphaelle.klitting@posteo.de (R. Klitting), eag@ceh.ac.uk (E.A. Gould), xavier.de-lamballerie@univ-amu.fr (X. de Lamballerie).

http://dx.doi.org/10.1016/j.meegid.2016.09.017
1567-1348/© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
residues of dengue virus, other flaviviruses and subsequently viruses from other taxonomic groups and eukaryotic and prokaryotic cells.

2. Results

2.1. GC% in genomic coding sequences of representative flaviviruses

It was previously observed (Jenkins et al., 2001; de Lamballerie et al., 2002) that GC% in flaviviruses, within a species, has a limited range. However, this range varied significantly (p-value = 0.01529, Wilcoxon-rank sum test, R software (Team, 2013)) between flavivirus species, from e.g. 38.4% for Tamana bat virus to 54.3% for TBEV. This is illustrated in Fig. 1, which shows the GC% distribution of representative viruses in the different evolutionary groups of the genus Flavivirus. Thus, the range of GC% within a given species is narrow: WNV (median 51.2; min 50.5–max 52.7), YFV (49.7; 48.9–50.1), Culex-borne flavivirus (CXFV: 53.0; 52.8–53.3), TBEV (54.3; 53.8–55.2), Aipoi virus (APOIV: 48.4; 48.4–48.4), Modoc virus (MODV: 45.2; 45.2–45.2), Montana Myotis Leukoencephalitis virus (MMLV: 43.9; 43.9–43.9) and Rio Bravo virus (RBV: 43.1; 43.1–43.2). However, for the major evolutionary groups there are wide differences in GC%: in the case of viruses with no known vector (NKV) the GC% ranges from 43.1 to 48.4. Similarly, the observed GC% ranges within the tick-borne flaviviruses (TBFV: 54.2; 48.6–55.2), mosquito-borne flaviviruses (MBFV: 46.3; 46.6–53.3) and insect specific flaviviruses (ISFV: 52.8; 46.6–53.3) are much wider than for individual species.

The wide range of GC% for the DENV can be explained by the clinical and historical criteria that prevailed for the delineation of the taxon (Sabin, 1952; Hammon et al., 1960; Calisher et al., 1989). Antigenic and genetic diversity of each serotype is mostly comparable with that of other flavivirus species, and this is compatible with the different GC% ranges observed for serotype 1 (DENV1: 46.3; 46.1–46.6), serotype 2 (DENV2: 45.7; 45.2–47.2), serotype 3 (DENV3: 46.2; 46.0–46.6) and serotype 4 (DENV4: 46.9; 46.7–47.2).

Amongst these dengue serotypes, DENV2 shows the most notable GC% heterogeneity. The DENV2 Asian/American genotype includes strains with the lowest GC% (45.2–46.3%) whereas the sylvatic genotype includes those with the highest GC% (46.4–47.3%).

2.2. Differences in mean GC% between conserved and variable amino acid residues in DENV

In order to investigate the underlying basis for the higher GC% observed in DENV2 sylvatic strains, pairwise (PW) comparison of two complete coding sequences (CDS), one from the sylvatic genotype (AN: EF105385; GC% = 46.5) and one from the Asian/American genotype (AN: EU482750; GC% = 45.5), was performed. Amino acid sequences were aligned and residues classified as “conserved”, i.e. identical in both sequences, otherwise as “variable”. GC% was estimated in both strains for the nucleotide triplets encoding either conserved or variable AA (Fig. 1 in supplemental data). It was strikingly higher in triplets corresponding to conserved amino acids for both the sylvatic (mean GC%: 46.9 in conserved residues vs 41.3 in variable residues) and the Asian/American strain (46.0 vs 38.2). This observation was repeated ten times using strains from different genotypes with similar results (data not shown).

An alignment of 250 DENV2 CDS was performed using Clustal W implemented in MEGA6 (Tamura et al., 2013) according to AA sequences. Alignment building methods and a comprehensive listing of all the sequences included in the analysis are provided as supplemental data.

We analysed GC% in AA residues conserved in the complete alignment (Complete Alignment Conserved (CA-Cons) residues), and in those not fully conserved (Complete Alignment Variable (CA-Var) residues). Altogether, for DENV2 the GCs of all CA-Cons residues was 53% and that of all CA-Var residues was 43% (See Fig. 2A and Fig. 2 in supplemental data). We defined a “HiGC-Cons” score as the proportion of cases in which the GC% of CA-Cons residues was higher than that of CA-Var residues when all sequences of an alignment were compared two by two (each pair generates two values: one for A vs B and another for B vs A). Importantly, we observed that the GC% of CA-Cons residues was higher than that of CA-Var residues for 100% of pairs compared in our DENV2 dataset (CA-HiGC-Cons score of 100%, Table 1). In Fig. 2A, CA-Cons and CA-Var GC% values were plotted for each of the 250 coding sequences. The difference between the two sets was found to be statistically significant (Wilcoxon-rank sum test, p-value = 2.2 e–16).

Similar analyses were achieved for dengue serotypes 1, 3 and 4 using alignments of 250, 246 and 122 sequences, respectively. For each serotype, the GC% of all CA-Cons residues (47%, 47% and 48%, for serotypes 1, 3 and 4, respectively) was higher than that of all CA-Var residues (43%, 40% and 40%, respectively) and 100% CA-HiGC-Cons scores were observed (Table 1).

Systematic pairwise comparison of all DENV2 coding sequences was then performed, using an in-house software programme. Here, the PW-Cons residues correspond to those residues conserved between the two individual sequences compared. The PW-Var residues correspond to those residues that are not conserved in the two individual sequences compared. Therefore, the numbers of DENV2 PW-Cons residues are globally higher than the numbers of DENV2 CA-Cons residues, and in very closely related sequences the number of DENV2 PW-Var residues is close to 100%. For 96% of pairwise comparisons, GC% of PW-Cons residues was found to be higher than that of PW-Var residues (PW-HiGC-Cons score of 96%, Table 1). When similar analyses were achieved for dengue serotypes 1, 3 and 4, PW-HiGC-Cons scores of 90, 91 and 80% were obtained, respectively (Table 1).

2.3. Extension of the observation to the genus Flavivirus

A new dataset, extending analysis to the genus Flavivirus, was created. It comprised 131 CDS and involved 47 species, 11 tentative species and 10 unclassified species (see supplemental data for a comprehensive listing of these sequences), all recognized as members of the genus Flavivirus (ninth ICTV report (2014)). The amino acid sequences of these viruses were aligned as previously described (Moureau et al., 2015) and GC% was estimated for CA-Cons/Var residues and PW-Cons/Var residues, with a distinction between 1st/2nd and 3rd codon positions (Table 1). For the complete open reading frame (ORF) sequences, we observed 100% CA-HiGC-Cons scores (Table 1). The mean GC% of all CA-Cons residues (56%) was significantly higher than that of all CA-Var residues (49%) (Supplemental data Fig. 2). For convenience, the AA residues corresponding to CA-Cons residues in the flavivirus alignment will be called the “Core AA residues” from now on. Pairwise comparison also revealed 100% PW-HiGC-Cons scores (Table 1). The same analyses performed using an NS5 gene dataset identified CA-HiGC-Cons and PW-HiGC-Cons scores of 99% and 97% respectively.

Within the genus Flavivirus, seven additional alignments including sequences selected at various taxonomic levels (species, serotype, genotype) were prepared. At the species level, we observed 100% CA-HiGC-Cons scores and PW-HiGC-Cons scores of 98% and 96% for DENV and TBEV, respectively (Table 1).

One hundred percent CA-HiGC-Cons scores were observed for all DENV serotypes and PW-HiGC-Cons scores of 90%, 96%, 91% and 80% were obtained for serotypes 1, 2, 3 and 4, respectively (Table 1). For DENV2 genotype Asian/American and DENV3, genotype III, CA-HiGC-Cons scores reached 100% for both genotypes and PW-HiGC-Cons scores were 89% and 69% respectively (Table 1).

In summary, within the genus Flavivirus, CA-HiGC-Cons scores were systematically 100% regardless of the taxonomic level and PW-HiGC-Cons scores increased with the taxonomic level and genetic diversity.
Fig. 1. GC\% distribution within the Flavivirus genus. The number of sequences corresponding to each genomic GC content value is reported for a set of viral species (WNV, TBEV, DENV1/2/3/4, YFV, CxFV, MODV, MMLV, RBV, APOIV) from the genus Flavivirus. The proportion of sequences from the species corresponding to a given GC content value is presented on a separate graph for each species (or vector group in the case of the NKV group). The phylogenetic relationships between all species are illustrated in a phylogenetic tree of the genus Flavivirus. The tree includes 131 sequences corresponding to 47 different recognized species in the genus Flavivirus. A Maximum Likelihood (1000 bootstrap replicates) tree was prepared using a General Time Reversible Model with invariant sites and a gamma distribution of rates across site.
2.4. Extension of the study to other virus evolutionary groups

Amino acid alignments were produced for representatives of species Hepatitis C virus (genus Hepacivirus), Classical swine fever virus (genus Pestivirus), Chikungunya virus (genus Alphavirus) and Enterovirus C (genus Picornavirus). Results were comparable with those reported above for the genus Flavivirus (CA-HiGC-Cons scores 100% and PW-HiGC-Cons scores above 95%). However, this was not the case for levivirus bacteriophages: CA-HiGC-Cons scores were commonly high, but PW-HiGC-Cons scores were below 50% when an alignment of concatenated ORFs and the highly conserved replicase ORF were analysed. In the case of retroviruses, a CA-HiGC-Cons score of 86% was obtained and the PW-HiGC-Cons score was 73%.

In a selection of negative-stranded RNA viruses, (genus Flavivirus) and species Influenzavirus A and B, and species Lassa mammavirus, 100% CA-HiGC-Cons scores were recorded and PW-HiGC-Cons scores ranged between 86% and 93%.
from 60 to 93%. In contrast, species Zaire Ebolavirus recorded a strikingly low CA-HiGC-Cons score when alignments of concatenated ORFs and the highly conserved polymerase ORF were analysed (Table 1).

In double-stranded RNA viruses, CA-HiGC-Cons scores were almost always 100%. However, different results were recorded in pairwise comparisons of Bluetongue virus (PW-HiGC-Cons score = 82%) and Rotavirus A (PW-HiGC-Cons score = 39%).

We extended our analysis to DNA viruses. For the single-stranded Human Parvovirus B19 species, CA-HiGC-Cons scores were higher than 75% and PW-HiGC-Cons scores were above 60%. For the double-
stranded mastadenoviruses, high HiGC-Cons scores were observed with sequences of species Human Mastadenovirus D, but not when analysis was extended to sequences from the complete genus (Table 1). Overall, these results suggest that, within a given evolutionary group, high HiGC-Cons scores can be observed for a large range of genetic distances (e.g. in the case of flavirviruses) or in a more limited range (e.g., at the genus level for Filoviridae, at the species level for the species Human Mastadenovirus D).

When we investigated alignments within the RpoB bacterial gene, 100% CA-HiGC-Cons scores were observed but PW-HiGC-Cons scores varied with the proportion of conserved AA residues, i.e., from 50% in an alignment of Mycobacterium tuberculosis genes (98% AA identity) to 96% in an alignment of different eubacterial genes (32% AA identity).

In alignments from primate genes (Apobec 3G and concatenated ORFs in 54 nuclear gene regions (Perelman et al., 2011), high CA-HiGC-Cons scores were observed, but PW-HiGC-Cons scores were close to equilibrium (50%).

### 2.5. Relationship between G + C content of conserved AA residues and AA distance

We performed pairwise comparisons between DENV2 sequence EU482750 and the other 249 sequences included in our DENV2 dataset, and plotted the GC% of PW-Cons AA residues as a function of the AA distance. The Spearman’s rank-sum correlation test (R software) identified a positive correlation (p-value = 2.2 e-16), with minimum and maximum values of GC% corresponding to minimum and maximum AA distances respectively (Fig. 2B).

A similar analysis was performed using the DENV2 sequence EU482750 and the other 130 sequences included in our flavivirus dataset (supplemental data Fig. 4). Again, a positive correlation (p-value = 7.6e-7) was identified between the GC% of PW-Cons AA residues and AA distance.

When comparing DENV2 sequence EU482750 with increasingly distant flavirviruses, the number of conserved AA residues decreases whilst their corresponding GC% increases. Therefore, the observed parallel decrease of GC% and AA distance may be due to a simple dilution of the high GC% “core” conserved residues along the nodes of the evolutionary tree (from root to terminal branches). To investigate this issue, we generated a flavivirus phylogenetic tree (Fig. 2C) and created an alignment using all the sequences from each node between the root and the DENV2 EU482750 leaf. The AA distance at the node was set as the number of AA residues of the DENV2 sequence EU482750 which were not conserved in the complete alignment generated at the node. The corresponding GC% of CA-Cons AA was calculated from the CA-Cons AA residues of DENV2 sequence EU482750 in this alignment.

During the evolutionary progression from the root of the tree to its terminal branches, at each node, the number of conserved residues increased as the result of new “node-specific” conserved AA residues. If the GC% of these “node-specific” conserved AA residues is higher than that of the variable residues at the node, then a simple dilution of the high GC% “core” conserved residues cannot alone explain the decrease of GC% in conserved residues observed with the decrease of genetic diversity. If this is not the case, the dilution hypothesis would be valid. To test this hypothesis, the GC% of “node-specific” CA-Cons AA residues was calculated from the DENV2 sequence EU482750 as follows:

\[
(GC_n \times L_n - GC_{dn} \times L_{dn})/(L_n - L_{dn})
\]

where GCn and Ln represent GC% and number of conserved AA residues at the considered node, respectively; and GCdn and Ldn represent GC% and the number of conserved AA residues at the first deeper node. Fig. 2D shows that, for the vast majority of nodes, the GC% of “node-specific” CA-Cons AA residues is higher than that of CA-Var residues. Therefore, a pure dilution effect cannot account for the observed higher GC% of CA-Cons versus CA-Var residues observed at the most peripheral nodes of the tree.

### 2.6. HiGC-cons score: which nucleotide frequency accounts for different G + C content in conserved and variable AA residues?

Nucleotide content within DENV2 sequence EU482750 was measured as follows (i) for the complete sequence (all residues), (ii) for AA residues fully conserved in the DENV2 alignment (S2 CA-Cons residues), (iii) for AA residues fully conserved in the flavivirus alignment (Core CA-Cons residues). For both S2 CA-Cons and Core CA-Cons AA residues, increased GC% and, to a lesser extent, decreased A% accounted for the observed increase of GC% in conserved AA residues (Table 2).

We repeated pairwise comparison for the 250 sequences of the DENV2 dataset and considered the four nucleotides separately. The most important variations between PW-Cons and PW-Var AA residues were observed for the G and A content (Table 3). Results were even more clear-cut when the flavivirus dataset was analysed, with respectively 100% and 0% of pairs for which the GC% (respectively A%) of PW-Cons AA residues was higher than that of PW-Var residues.

### 2.7. Extended analysis of G + C content of conserved residues in relation to AA distances

The AA composition of a protein sequence accounts for ca. two thirds of the nucleotide composition of the sequence by which it is encoded. Accordingly, for S2 CA-Cons/Var AA residues and for Core CA-Cons/Var AA residues, (i) nucleotide content was analysed for each of the 20 AA, separately for conserved and variable residues; (ii) the proportion of each of the 20 AA was calculated separately for conserved and variable residues.

Analysis was performed for all four nucleotides (supplemental data). To increase legibility, results below are provided for G content only (Table 4). Each AA type whose GC% is higher in Core CA-Cons or S2 CA-Cons parts of the sequence can potentially contribute to the increase in GC%. Similarly, any AA type whose GC% is intrinsically “high” (>25%) and which is found in higher proportion in Core CA-Cons or S2 CA-Cons versus CA-Var residues observed at the most peripheral nodes of the tree.
Cons residues can also contribute to increase their G%. We calculated a participation score that accounts for these two phenomena using the following formula:

\[
\text{(mean G\% in sequence) } \times (\text{AA proportion in sequence}) / \text{total no. of AA in sequence}.
\]

In the above formula, “sequence” refers either to Core CA-Cons or S2 CA-Cons residues in DENV2 EU482750, according to whether it refers to Flavivirus or DENV2 alignments. This score enables evaluation of the participation of each AA type to the G content.

Overall, the AA types that contribute the most to the high G content of conserved residues are Glycine, Tryptophan, and Cysteine (Table 4).

Therefore, the difference in G content between conserved and variable residues is due to AA composition rather than to a “high G” encoding of the considered sequence. For example, Glycine residues participate in the high G content of conserved residues and yet are not necessarily encoded with a higher G% in these residues: the mean G% of Glycine residues is similar in S2 CA-Cons residues and S2 CA-Var residues (73% vs 72%), and even lower in Core CA-Cons residues than in Core CA-Var residues (70% vs 73%).

Overall, we conclude that the high G content in conserved residues of the DENV2 EU482750 sequence is due to the limited number of AA for which the G% is intrinsically high (Glycine, Tryptophan and Cysteine) and which are found in higher proportion amongst conserved residues of the sequence.

3. Discussion

Our study was stimulated by the original concept that flavivirus evolutionary groups are endowed with high genetic variability and specific G + C content (Jenkins et al., 2001; de Lamballerie et al., 2002). An extreme example is provided by the comparison of two sister evolutionary groups: viruses with no known vector are characterised by low G + C content (e.g., RBV, GC% = 43%) whereas tick-borne flaviviruses are characterised by high G + C content (e.g., TBEV, GC% = 54%). This is first explained by strong differences in the AA content of viral polyproteins, that have a higher proportion of AA residues with intrinsically high GC% (namely alanine, glycine and arginine) in TBEV sequence and a lower proportion of AA residues with intrinsically low GC% (isoleucine, lysine and asparagine). According to Sueoka (Sueoka, 1988) and Singer and Hickey (Singer and Hickey, 2000) “the most parsimonious explanation of the observed patterns of amino acid composition in these genomes (would be) an underlying mutational bias that varies between lineages”. In addition, the encoding of AAs also contributes to the difference observed, since, as previously reported, the G + C content at the third position of codons faithfully follows that of the complete polyprotein (Bellgard and Gojobori, 1999a,b; de Lamballerie et al., 2002).

In previous studies, relationship between G + C content and AA sequence has been extensively analysed. It has been shown that AAs encoded by GC-rich (and GC-poor) codons are more frequent in sequences with high (and low) G + C content (Singer and Hickey, 2000; Banerjee et al., 2005; Li et al., 2015). Authors have investigated the relationship between G + C content and the frequency of hydrophobic, hydrophilic or amphipatic AAs, the function and the structure of the protein (e.g., membrane proteins, protein secondary structures) (Lobry, 1997; Gu et al., 1998; Banerjee et al., 2005). Studies were performed on a genome wide scale and on genes (Lobry, 1997; Gu et al., 1998; Singer and Hickey, 2000; Banerjee et al., 2005), or on fragments of coding regions (Li et al., 2015), but not at the scale of individual AA residues.

Remarkably, whilst analysing the characteristics of G + C content amongst flaviviruses, we made a founding observation: in an alignment of flaviviral sequences, G + C content is strikingly different in conserved and variable AA residues: higher values are observed in conserved than in variable residues. Within the genus flavivirus, this observation was valid for the complete range of observed genetic distances, and for both complete alignments- and pairwise- conserved residues. For
example, in an alignment of two closely related dengue virus sequences (i.e., including a large majority of conserved residues), or an alignment of two distantly related flaviviral species (i.e., including a large majority of variable residues), G + C content was consistently higher in the conserved residues than in the variable residues. Moreover, in the case of flaviviruses, this difference is directly attributable to the specific G content in the conserved and variable AA residues.

The association between G + C or G content and conservation of AA residues was remarkably strong. In a dataset of 250 DENV2 sequences, the conservation rate in AA residues with a 100% G + C content (~6% of all AAs) was ~85% and in residues with a 100% G + C content for codon positions 1 and 2 (~21% of all AAs), it was ~84%. In other words, without considering the nature of the AAs or their position in codon positions 1 and 2 (~21% of all AAs), it was ~84%. In other words, without considering the nature of the AAs or their position in codon positions 1 and 2 (~21% of all AAs), it was ~84%. In other words, without considering the nature of the AAs or their position in codon positions 1 and 2 (~21% of all AAs), it was ~84%. In other words, without considering the nature of the AAs or their position in codon positions 1 and 2 (~21% of all AAs), it was ~84%. In other words, without considering the nature of the AAs or their position in codon positions 1 and 2 (~21% of all AAs), it was ~84% certainty the conservation amongst all DENV2 sequences of ~21% of AA residues. Furthermore, when the G content alone is considered, a similar trend is observed in glycine residues, i.e. the only AA with a 100% G content (Supplemental Data).

Indirect support to our findings can be found in previous analyses of prokaryotic genomes. Singer and Hickey (Singer and Hickey, 2000) analysed the relationship between the degree of AA bias and sequence divergence in Mycobacterium tuberculosis and Borrelia burgdorferi. Interestingly, they reported higher frequency of GARP amino acids (encoded by GC-rich codons) in the conserved genes of Borrelia burgdorferi, and higher frequency of FYMINK amino acids (encoded by GC-poor codons) in the variable genes. Several studies have highlighted a relationship between hydropathy of AAs and G + C content (D’Onofrio et al., 1999; Jabbari et al., 2003; Banerjee et al., 2005). Our finding may therefore derive from a higher conservation of hydrophobic AA residues. However, AAs which are at least two times more frequent amongst flavivirus core-conserved residues (i.e., residues shared by all flaviviral species) than amongst variable residues include glycine, arginine, tryptophan and cysteine. Glycine, arginine and tryptophan have an intrinsically high G content and are either amphipathic (glycine and tryptophan) or hydrophilic (arginine). Cysteine is the only hydrophobic AA in the list and its G content is not intrinsically high. Hence, the link between nucleotide content and conservation of AA is not a consequence of the hydropathy of the latters.

The next issue will be to resolve whether AAs such as glycine, arginine, tryptophan and cysteine are conserved because of their function, their intrinsic properties or their high G content. If G content drives AA conservation, then a high G% should be fostered on the ancestral sequence and that most non-synonymous substitutions occurred over time within this pool, in accordance with the concept that selective pressure maintains most AA frequencies close to an optimal DNA base composition (Sueoka, 1962). However, the fact that the most conserved AA residues have a specifically high G + C content does not necessarily imply that the complete ancestral sequence had a similarly high G + C content. Rather, we can imagine that a pool of residues with a lower G + C content existed in the ancestral sequence and that most non-synonymous substitutions occurred over time within this pool. Within this framework, the evolutionary mechanisms tend to replace an AA by another with a lower G + C content. Previously, Jenkins and collaborators (Jenkins et al., 2001) used phylogenetic methods to propose estimates of the G + C content of progenitors for the genus Flavivirus or its main evolutionary branches. Their results provided no clear evidence that the G + C content at the deepest nodes of the phylogenetic reconstruction was higher than that of extant species. Using similar methods but a more recent and larger dataset of flavivirus sequences, we also failed to identify a trend for G + C content decrease over time; similarly, when we compared the sequences of the first isolates of the West-Nile virus made in the United States of America (1999) with late isolates in the 2010–2012 period, or when we tried to apply to flaviviruses the approach proposed by Gojobori for estimating the G + C content of ancestral mycobacterial sequences (Bellgard and Gojobori, 1999a,b), we found no evidence of such a trend (data not shown). Finally, we tested but did not confirm, the hypothesis that the evolution of the G + C content with respect to genetic distance may be explained by the progressive dilution of an ancestral core of AAs with a high G + C content during the course of virus evolution.

Our founding observation, that G + C content of AAs is higher in conserved than in variable residues applies to the whole genus Flavivirus but its scope is likely to be wider. Pairwise and complete alignments were assembled in 11 different virus families and a “HiGC-Cons” score was defined as the proportion of asymmetric pairs for which the G + C content was higher in conserved than in variable residues (Table 1). Altogether, these results indicate that association between G + C content and AA conservation is encountered in a large number of viral and possibly, non-viral- evolutionary groups. High HiGC-Cons scores were observed in RNA viruses whose genomic structure is close to that of flaviviruses, e.g., alphaviruses and enteroviruses (short single-stranded RNA genomes of positive polarity). For more distant viral genera including retroviruses, negative and double-stranded RNA viruses, or DNA viruses, the association between G + C content and AA conservation may be limited to a specific range of genetic distances: our observation is not valid for a number of RNA and DNA datasets that include either highly conserved (sp. Zaire Ebolavirus) or poorly conserved sequences (Mastadenovirus and Leviviruses). Further investigations on non-viral evolutionary groups gave us reasonable grounds to believe that our observation may also apply to some prokaryotic sequences but we did not identify relevant datasets from which to draw a robust conclusion in the case of high eukaryotes. In sum, our investigations should be regarded as a first detailed description of a completely new phenomenon that should serve as a starting point for further studies of specific evolutionary groups, based on robust and validated genomic datasets, and associating nucleotide content of codons with the biological properties of the corresponding AAs by both analytical and experimental means.

Finally, our study sheds new light on the characteristics of the genetic code. It is clear that the distribution of codons assigned to AAs is not randomly determined (Taylor and Coates, 1989; Freeland and Hurst, 1998) since AAs with similar structures or properties commonly have adjacent codons or share the first base of their codon (Wong, 1980; Di Giulio, 1989). Our findings would additionally suggest a relationship between simple characteristics of the genetic code (i.e., the G or G + C content of codons) and the evolutionary behavior of the corresponding encoded AA residues. The mechanisms behind this association remain to be determined. Since the most obvious parameter associated with G + C content is the strength of molecular hybridization of polynucleotide duplexes, fundamental driving forces, such as thermodynamic constraints (England, 2013) may be operating.

4. Conclusion

Within the genus Flavivirus, G + C content is strikingly higher in conserved than in AA residues, and this observation is valid for a wide range of genetic distances, and for both complete alignments-
pairwise-conserved residues. We observed that the G + C content of conserved residues increases with genetic distance, but could not provide any evidence of an evolutionary mechanism that would tend to replace an AA by another with a lower G + C content.

The association between G + C content and AA conservation is also encountered in a large number of viral and possibly, non-viral-evolutionary groups. It is specifically strong in RNA viruses whose genomic structure is close to that of flaviviruses (short single-stranded RNA genomes of positive polarity) and may be limited to specific ranges of genetic distances in more distantly related virus groups. In the case of flaviviruses, this association is directly attributable to the increased frequency of AAs with an intrinsically high G content (namely glycine, arginine and tryptophan) amongst conserved residues, and not to the biochemical characteristics of the AAs (e.g. their hydrophathy). Beyond the specific case of flaviviruses, the intriguing discovery of a widely encountered relationship between the nucleotide content and the conservation of AA residues obviously requires detailed analysis in an enlarged variety of evolutionary branches of the tree of life.

Competing interests

The authors have declared that no financial and non-financial competing interests exist.

Acknowledgements

We thank Morgan Seston for software development.

This work was supported by the European Virus Archive goes Global, http://global.europeanvirus-archive.com/ [European Union’s Horizon 2020 research and innovation programme under grant agreement no. 653316]; and the Agence Nationale de la Recherche, http://www.agence-nationale-recherche.fr/ (grant ANR-14-CE14-0001 RNA Vaccines). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.meegid.2016.09.017.

References