

Analysis of synonymous codon usage bias and nucleotide and amino acid composition in 13 species of Flaviviridae

Fatemeh Moosavi¹, Hassan Mohabatkar^{2*} and Sasan Mohsenzadeh¹

¹ Department of Biology, College of Sciences, Shiraz University, Shiraz 71454, Iran

² Department of Biotechnology, Faculty of Advanced Sciences and Technologies, University of Isfahan, Isfahan, Iran

Received 29 May 2011

Accepted 5 July 2011

Abstract

The family, Flaviviridae includes viruses which cause several diseases including Dengue fever, Japanese encephalitis, Murray Valley encephalitis, Tick-borne encephalitis, West Nile encephalitis, Yellow fever and Hepatitis C virus infection. Members of this family have monopartite, linear, single-stranded RNA genomes of positive polarity with 9.6-12.3 kb in length. Here, we analyzed the codon usage of 13 species of this family by using gene infinity package. Base analysis was performed by CAIcal server and amino acid composition was calculated by PseAAC web-server. The results showed that the highest number of A, G and C bases were seen in the RNA genome of *Dengue virus 2*, *Tick borne encephalitis virus* and *Hepatitis C virus*, respectively. Although the number of U bases used in RNA genomes was very close, the highest U nucleotide amount was 23.77% in *Wesselsbron virus*. The lowest number of C, G, U and A bases was seen in *Bovine viral diarrhea virus*, *Dengue virus 2*, *Tick borne encephalitis virus* and *Hepatitis C virus*, respectively. In this study, it was found that the complete genome of *classical swine fever virus* has a lower GC content and genome of *Tick borne encephalitis virus*, *Hepatitis C virus* and *Powassan virus* have the highest GC content among other examined species. We also classified the amino acids as rare (Phenylalanine, Cysteine, Histidine, Methionine, Asparagine, Glutamine, Tryptophan and Tyrosine), frequent (Alanine, Glutamic acid, Glycine, Leucine, Valine and Threonine), and intermediate (all others). The highest numbers of preferred codons exist in *Wesselsbron virus* and the lowest in *West Nile virus*.

Keywords: Flaviviridae, codon usage bias, amino acid composition, nucleotide composition

Introduction

Each amino acid is coded either by one or more codon(s). Therefore, genes and species may utilize different sets of codons (Vicario et al., 2007). It has been demonstrated that codon usage pattern is nonrandom and species-specific, and the inter-genomic variation of the codon usage pattern is a common phenomenon (Grantham et al., 1981). For example, a range of minimal to extreme codon bias is present in unicellular organisms and in *Drosophila* species (Sharp and Li, 1987; Powell and Moriyama, 1997). In addition, it is well identified that different genes have special codon usage patterns in a same organism (Zhuo-Cheng et al., 2003). Some factors such as translational selection (Bennetzen et al. 1982), mutation (Sueoka et al., 2000), compositional constraints (Hou et al., 2002), physical location of the gene on chromosome (Kerr et al., 1997), replication-translational selection (Naya et al., 2001), hydrophobicity of each gene

(Romero et al., 2000), gene length (Sau et al., 2005), and CpG island (Shackelton et al., 2006) were found to influence codon usage in animal viruses and phages. Mutational bias was found as the main determinant factor (XU et al., 2008).

Flaviviridae are mainly classified into three genera: *Flavivirus*, *Hepacivirus* and *Pestivirus*. Members of this family have monopartite, linear, single-stranded RNA genomes of positive polarity with 9.6-12.3 kb in length (Regenmortel et al., 2000). Major diseases caused by the *Flavivirus* and *Hepacivirus* genera include: Dengue fever, Japanese encephalitis, Murray Valley encephalitis, Tick-borne encephalitis, West Nile encephalitis, Yellow fever and Hepatitis C Virus Infection while *Pestiviruses* are usually pathogens of non-human mammals (Gould et al., 2001). Analysis of codon usage patterns of Flaviviridae would provide a foundation for understanding the related mechanism for biased usage of synonymous codons, the evolution and pathogenesis of Flaviviridae. It would be helpful for choosing appropriate host expression systems for an

*Corresponding author E-mail:
h.mohabatkar@ast.ui.ac.ir

optimized expression of target genes. Here, using the codon usage database, the codon usage and base composition of thirteen species of Flaviviridae were analyzed, and the patterns of preferred codons for each individual amino acid in each species were identified.

Materials and Methods

Genomic sequence retrieval

Complete (or nearly complete) genomic RNA sequences for 13 Flaviviridae species and the amino acids of polyprotein sequences were retrieved from NCBI GenBank (<http://www.ncbi.nlm.nih.gov/>). The accession numbers of these proteins are AY835999 (*Dengue virus 1*), GQ868604 (*Dengue virus 2*), DQ863638 (*Dengue virus 3*), AF045551 (*Japanese encephalitis virus*), NC_000943 (*Murray Valley encephalitis virus*), GQ228395 (*Tick-borne encephalitis virus*), AY640589 (*Yellow fever virus*), FJ407092 (*Hepatitis C virus*), AY842931 (*West Nile virus*), AF526381 (*Bovine viral diarrhea virus*), NC_012735 (*Wesselsbron virus*), GQ923951 (*Classical swine fever virus*) and EU770575 (*Powassan virus*). Polyprotein sequences were obtained from NCBI, including those of *Japanese encephalitis virus* (AAA21436), *Murray Valley encephalitis virus* (NP_051124), *West Nile virus* (AAV54504), *Tick-borne encephalitis virus* (NP_043135), *Powassan virus* (NP_620099), *Yellow fever virus* (AAT58050), *Wesselsbron virus* (YP_002922020), *Hepatitis C virus* (BAB32877), *Classical swine fever virus* (ACL80334), *Bovine viral diarrhea virus* (NP_040937), *Dengue virus 1* (AAV97946),

Dengue virus 2 (ACW82981) and *Dengue virus 3* (ACW83004).

Sequence analysis

The nucleotide compositions were determined by using CAIcal server. This server is available at <http://genomes.urv.es/CAIcal>. Codon usage data and effective number of codons (ENC) was analyzed by gene infinity package (<http://www.geneinfinity.org>). PseAAC web-server was used to demonstrate amino acid composition of proteins (Shen and Chou, 2008).

Results

Base composition analysis

In order to examine the base composition variation among different complete RNA genome of several species of 13 Flaviviridae genera, their nucleotide composition was calculated.

As the results in table 1 indicate, strong nucleotide biases were observed in all of the examined genomes. The highest base percentage was 33.38 % A in *Dengue virus 2*, 31.57% G in *Tick borne encephalitis virus* and 28.56% C in *Hepatitis C virus*. Although the number of U base used in RNA genomes was very close, the highest U nucleotide amount was 23.77% in *Wesselsbron virus*. The lowest number of bases was 19.85% C in *Bovine viral diarrhea virus 1*; 25.17% G in *Dengue virus 2*; 20.70 U in *Tick borne encephalitis virus*, and 21.14 % A in *Hepatitis C virus*. The Adenine percentage was clearly the most frequent and nucleotide among these Flaviviridae members (ranging from 21.14 to 33.38%).

Table 1. Nucleotide composition of Flaviviridae.

Virus	Nucleotide				
	G%	% A	U%	C%	Length (bp)
<i>Dengue virus 1</i>	25.72	32.14	21.38	20.74	10727
<i>Dengue virus 2</i>	25.17	33.38	21.23	20.33	10679
<i>Dengue virus 3</i>	25.98	32.04	21.20	20.77	10707
<i>Japanese encephalitis virus</i>	28.44	27.53	21.00	23.01	10963
<i>Murray Valley encephalitis virus</i>	27.51	28.89	22.40	21.20	11014
<i>West Nile virus</i>	28.76	27.34	21.49	22.40	11029
<i>Tick borne encephalitis virus</i>	31.57	24.95	20.70	22.76	11096
<i>Powassan virus</i>	31.19	25.20	21.49	22.12	10839
<i>Yellow fever virus</i>	28.33	27.40	22.92	21.35	10862
<i>Wesselsbron virus</i>	26.62	28.58	23.77	21.03	10812
<i>Hepatitis C virus</i>	27.75	21.14	22.55	28.56	9432
<i>Classical swine fever virus</i>	26.41	30.92	21.65	21.00	12296
<i>Bovine viral diarrhea virus 1</i>	26.31	31.89	21.79	19.85	12220

Table 2 shows that the complete genome of *Classical swine fever virus* has a lower GC content (42.9%) and genomes of *Tick borne encephalitis virus*, *Hepatitis C virus* and *Powassan virus* have a higher GC content than other species (While the corresponding percentages were 56.52, 56.31 and 53.32). No clear difference in GC content was seen among other genomes. On the other hand, differences in GC content at different synonymous positions of codons were obvious.

For example, GC content at the synonymous first position of codons for the *Classical swine fever virus* and *Yellow fever virus* are 37.25 and 36.02 %, respectively, while these values for *Tick borne encephalitis*, *Bovine viral diarrhea* and *Japanese encephalitis* viruses are 28.27, 28.33 and 29.00 %, respectively. The GC content at the synonymous second position of codons is calculated 28.72 % for *Powassan virus* but for the *Tick borne encephalitis virus* is 36.35 %. GC content at the synonymous third position of codons is computed 28.07 and 28.65 % for *Classical swine fever virus* and *Yellow fever virus*, respectively, while for the *West Nile virus* and *Dengue virus 2* are 36.32 and 36.23 %. The most variable genome in GC content at the different synonymous positions of codons is *Classical swine fever virus* (ranging from 28.07 to 37.25%) but least variation was seen in RNA genome of *Wesselsbron virus*.

Amino acid composition

In this study, the frequency of the 20 amino acids in the polyproteins of different species of Flaviviridae was calculated (table 3). The amino acids can be classified as rare (Phenylalanine, Cysteine, Histidine, Methionine, Asparagine, Glutamine, Tryptophan and Tyrosine), frequent (Alanine, Glutamic acid, Glycine, Leucine, Valine and Threonine), and intermediary (all others).

GCN codon encodes Alanine and CCN codon encodes Proline residues. Interestingly, *Hepatitis C*

virus with the highest C and G content has the largest number of Alanine and Proline residues. Lysine amino acid is encoded by two AAG and AAA codons. The Largest number of this amino acid was present in *Pestivirus* that had high A and low C content.

Codon usage patterns

Commonly, codons which were used more than twice as frequently as host consensus codons are regarded as preferred codons of heterologous genes (Lu et al., 2005). In this work, compared with *Homo sapiens* (selected host for these viruses), several preferred codons were found in each species (table 4). In *Dengue virus 1, 3* and *Hepatitis C virus* five preferred codons and in *Bovine viral diarrhea virus* and *Dengue virus 2* three preferred codons were demonstrated. The highest numbers of preferred codons exist in *Wesselsbron virus* and the lowest in *West Nile virus*. Six and seven preferred codons were found in *Yellow fever virus* and *Classical swine fever virus* while only two preferred codons were found in *Murray Valley encephalitis* and *Powassan viruses*. Four preferred codons were also characterized for *Japanese encephalitis virus*. AGA coding for Arg was preferred in most of the species. The ATA coding for Ile, CAC for His and TTG for Leu are preferred only in *Bovine viral diarrhea virus*, *Hepatitis C virus* and *Murray Valley encephalitis virus*. For Leu, RNA genomes of various species did not have any preferred codons while for Gly, Ser and Arg, they used two preferred codons in some species. The preferentially used codons were A-ended, T-ended, and G-ended. Flaviviridae can be classified in five species groups based on their genome signature (U-rich, G-rich, A-rich, C-rich, or relatively unbiased). Results are summarized in table 5.

Table 2. GC, A, T, G and C contents at different codon positions of complete genome of Flaviviridae

<i>N%</i>	<i>Dengue virus 1</i>	<i>Dengue virus 2</i>	<i>Dengue virus 3</i>	<i>Japanese encephalitis virus</i>	<i>Murray Valley encephalitis virus</i>	<i>West Nile virus</i>	<i>Tick borne encephalitis virus</i>	<i>Powassan virus</i>	<i>Yellow fever virus</i>	<i>Wesselsbron virus</i>	<i>Hepatitis C virus</i>	<i>Classical swine fever virus</i>	<i>Bovine viral diarrhea virus 1</i>
A1	35.60	28.77	34.12	27.37	27.32	30.66	26.04	27.01	23.59	29.28	18.06	27.72	32.01
A2	32.60	36.84	33.31	25.15	28.49	27.06	21.74	25.85	30.46	29.27	22.48	33.23	30.30
A3	28.22	34.53	28.69	30.10	30.86	24.31	27.09	22.72	28.12	27.10	22.86	31.82	33.36
C1	21.31	37.43	21.60	22.74	22.88	18.39	21.58	19.68	24.31	20.03	33.43	25.01	20.20
C2	17.87	21.01	17.31	27.61	23.61	22.49	26.31	21.56	18.23	20.06	24.01	16.71	22.21
C3	23.05	17.18	23.39	18.69	17.10	26.25	20.39	25.13	21.49	22.94	28.24	21.30	17.11
U1	18.68	22.84	18.66	27.85	28.05	16.27	27.85	16.00	22.70	21.50	20.86	19.27	28.75
U2	17.06	18.90	16.81	19.07	22.25	28.26	18.98	28.20	16.87	23.64	15.07	17.44	20.23
U3	28.41	15.99	28.13	16.06	16.86	19.91	15.27	20.26	29.17	26.11	22.30	28.23	16.77
G1	24.39	19.98	25.60	22.03	21.74	34.63	24.53	37.30	29.39	29.17	27.64	27.99	19.02
G2	32.80	23.24	32.56	28.16	25.61	22.17	32.96	24.38	34.91	27.02	33.33	32.60	27.15
G3	20.31	32.31	19.78	35.14	35.17	29.51	37.24	31.88	21.21	23.83	22.96	18.64	32.75
GC	46.46	45.5	46.45	51.45	48.71	51.16	54.33	53.32	49.68	47.65	56.31	42.9	46.15
GC1	32.78	36.23	33.66	29.00	30.53	30.53	28.27	35.62	36.02	34.51	36.14	37.25	28.33
GC2	36.13	32.41	32.41	36.13	33.70	34.57	36.35	28.72	35.65	32.94	33.94	34.66	35.66
GC3	31.09	31.36	35.56	34.87	35.76	36.32	35.34	35.64	28.65	32.72	29.91	28.07	36.01

Table 3. Amino acid composition of polyprotein sequences from Flaviviridae

Amino acid	Dengue virus 1	Dengue virus 2	Dengue virus 3	Japanese encephalitis virus	Murray Valley encephalitis virus	West Nile virus	Tick borne encephalitis virus	Powassan virus	Yellow fever virus	Wesselsbron virus	Hepatitis C virus	Classical swine fever virus	Bovine viral diarrhea virus 1
A Alanine	7.017	6.812	6.991	8.450	8.416	7.981	8.202	7.760	7.183	7.137	9.199	6.466	5.843
C Cysteine	1.739	1.681	1.711	1.632	1.660	1.806	1.904	1.816	1.876	1.821	2.868	1.858	1.906
D Aspartic acid	4.245	3.863	4.012	4.720	4.688	4.515	4.511	4.890	4.632	5.140	4.187	4.710	4.514
E Glutamic acid	6.515	7.166	6.667	6.031	6.057	6.175	6.678	6.296	6.157	5.374	3.825	6.441	6.795
F Phenylalanine	3.096	2.978	2.891	3.205	3.262	3.059	2.841	2.723	3.254	3.231	2.868	3.182	3.009
G Glycine	8.196	8.169	8.348	8.741	8.532	8.535	9.402	9.605	8.854	8.341	8.473	7.510	7.523
H Histidine	2.152	2.005	2.094	2.098	1.951	1.864	2.402	2.401	2.375	2.144	1.879	2.088	2.106
I Isoleucine	5.896	6.547	5.988	4.837	5.446	5.098	3.486	3.895	5.277	5.521	4.616	5.448	6.194
K Lysine	6.073	6.399	6.519	5.536	5.737	5.476	4.628	4.802	5.717	5.962	3.363	7.434	7.447
L Leucine	9.493	9.112	9.322	8.800	8.969	9.059	9.783	9.810	9.147	8.899	9.891	9.623	9.855
M Methionine	3.774	3.716	3.658	3.234	2.999	3.321	3.310	3.455	3.811	3.671	2.341	2.240	2.482
N Asparagine	3.685	3.981	4.012	3.613	3.611	3.641	2.665	2.577	3.635	3.554	2.671	3.742	3.837
P Proline	4.098	4.276	4.484	4.138	4.193	4.107	3.837	4.100	4.046	4.229	7.056	4.404	4.413
Q Glutamine	3.243	2.978	3.274	2.535	2.504	2.651	2.695	2.577	2.756	2.526	3.066	3.131	3.260
R Arginine	5.660	5.751	5.074	2.593	5.999	6.234	7.147	7.145	6.039	5.551	5.605	5.321	5.291
S Serine	5.955	5.633	5.192	6.119	5.708	5.768	5.712	5.798	6.332	6.432	7.583	4.964	5.191
T Threonine	7.459	7.815	8.112	5.624	7.164	7.311	6.503	6.120	5.863	6.843	7.319	7.612	7.146
V Valine	6.692	6.370	6.637	7.139	7.833	8.127	9.168	9.048	8.209	8.253	7.616	8.070	7.447
W Tryptophan	2.830	2.654	2.802	8.217	2.679	2.709	2.988	2.958	2.492	2.761	2.275	1.629	1.580
Y Tyrosine	2.182	2.094	2.212	2.739	2.592	2.563	2.138	2.225	2.345	2.614	3.297	4.124	4.162

Table 4. Codon usage data in 13 species of Flaviviridae. Dark gray codons are the preferred codons in Flaviviridae. Triplets in bold face indicate a high frequency in coding the amino acid. Light gray codons appear during low frequency coding of the amino acid.

aa	Codon	Dengue1		Dengue 2		Dengue 3		Japanese encephalitis		Murray Valley encephalitis		Human
		1/1000	Fract	1/1000	Fract	1/1000	Fract	1/1000	Fract	1/1000	Fract	
Arg	CGC	6.15	0.06	3.37	0.05	5.60	0.05	8.21	0.09	4.90	0.07	10.68
	AGG	35.80	0.35	15.45	0.25	30.54	0.30	22.44	0.26	18.80	0.27	11.71
	AGA	36.92	0.36	25.85	0.41	41.19	0.40	26.82	0.31	20.43	0.30	11.72
	CGG	6.43	0.06	7.59	0.12	7.00	0.07	12.86	0.15	12.26	0.18	11.65
	CGA	8.67	0.08	6.46	0.10	8.41	0.08	10.13	0.12	7.35	0.11	6.24
Leu	CGT	9.79	0.09	3.65	0.06	9.53	0.09	6.02	0.07	4.63	0.07	4.63
	TTG	8.95	0.16	20.23	0.26	7.28	0.13	22.44	0.28	34.87	0.35	12.75
	TTA	4.20	0.08	12.36	0.16	5.04	0.09	8.76	0.11	11.44	0.11	7.43
	CTG	9.51	0.17	15.17	0.19	12.05	0.22	19.16	0.24	23.97	0.24	40.13
	CTA	6.71	0.12	10.96	0.14	5.60	0.10	7.94	0.10	8.72	0.09	7.04
	CTT	15.10	0.28	10.12	0.13	17.37	0.32	12.04	0.15	10.90	0.11	13.01
	CTC	10.35	0.19	10.12	0.13	7.57	0.14	10.67	0.13	10.08	0.10	19.67
Ser	AGT	26.01	0.28	7.31	0.09	23.82	0.27	9.30	0.09	9.81	0.12	12.05
	AGC	30.49	0.33	7.87	0.10	32.78	0.37	12.04	0.12	11.71	0.15	19.45
	TCG	3.08	0.03	8.71	0.11	1.12	0.01	18.88	0.18	7.90	0.10	4.47
	TCA	10.35	0.11	29.22	0.36	9.53	0.11	34.76	0.33	26.42	0.33	12.00
	TCT	13.43	0.14	11.24	0.14	11.21	0.13	13.14	0.13	10.62	0.13	14.89
	TCC	9.79	0.11	15.73	0.20	10.65	0.12	15.87	0.15	13.89	0.17	17.63
Ala	GCG	2.80	0.08	5.06	0.16	3.08	0.08	8.76	0.17	8.99	0.19	7.55
	GCA	8.39	0.23	14.89	0.48	15.97	0.39	19.43	0.38	19.89	0.42	16.00
	GCT	18.46	0.50	6.74	0.22	13.45	0.33	12.59	0.25	8.72	0.18	18.57
Gly	GCC	6.99	0.19	4.50	0.14	8.13	0.20	9.85	0.19	9.81	0.21	28.28
	GGG	14.27	0.18	15.17	0.28	17.65	0.21	24.08	0.35	15.80	0.26	16.48
	GGA	31.61	0.40	21.35	0.40	31.94	0.37	21.62	0.32	25.88	0.43	16.42
	GGT	14.55	0.18	8.99	0.17	18.21	0.21	10.67	0.16	10.62	0.18	10.80
Pro	GGC	19.58	0.24	8.43	0.16	17.93	0.21	11.49	0.17	7.90	0.13	22.56
	CCG	3.36	0.08	7.59	0.15	2.24	0.05	13.68	0.20	10.62	0.17	7.04
	CCA	17.34	0.39	23.04	0.46	16.25	0.38	30.38	0.45	32.42	0.51	16.84
	CCT	14.83	0.33	8.43	0.17	14.85	0.35	11.77	0.17	11.44	0.18	17.42
Thr	CCC	8.95	0.20	11.52	0.23	9.25	0.22	12.04	0.18	9.26	0.15	20.03
	ACG	6.15	0.10	8.71	0.14	5.88	0.10	16.15	0.22	10.62	0.16	6.15
	ACA	18.46	0.30	28.66	0.45	13.45	0.23	29.28	0.39	29.69	0.45	14.91
	ACT	20.70	0.34	12.64	0.20	19.89	0.35	13.68	0.18	12.80	0.19	13.01
Val	ACC	15.66	0.26	13.49	0.21	18.21	0.32	15.87	0.21	13.08	0.20	19.09
	GTG	10.63	0.27	14.89	0.44	11.21	0.28	15.60	0.45	19.61	0.52	28.56
	GTA	5.31	0.13	7.59	0.23	4.48	0.11	6.29	0.18	5.18	0.14	7.06
	GTT	15.38	0.39	7.02	0.21	16.25	0.40	7.94	0.23	6.27	0.17	10.98
Asn	GTC	8.11	0.21	4.21	0.13	8.69	0.21	5.20	0.15	6.54	0.17	14.63
	AAT	28.81	0.52	12.36	0.39	28.02	0.48	9.30	0.37	11.71	0.43	16.72
	AAC	26.29	0.48	19.67	0.61	30.26	0.52	15.60	0.63	15.53	0.57	19.17
Asp	GAT	22.38	0.54	10.40	0.50	21.29	0.57	9.58	0.60	9.81	0.54	21.98
	GAC	18.74	0.46	10.40	0.50	16.25	0.43	6.29	0.40	8.44	0.46	25.50
Cys	TGT	16.22	0.54	15.17	0.53	15.13	0.51	14.50	0.40	17.16	0.47	10.31
	TGC	13.99	0.46	13.49	0.47	14.57	0.49	22.17	0.60	19.61	0.53	12.55
His	CAT	33.29	0.58	17.42	0.46	32.50	0.59	9.58	0.42	13.35	0.48	10.69
	CAC	24.34	0.42	20.51	0.54	22.98	0.41	13.14	0.58	14.44	0.52	15.03
Phe	TTT	13.43	0.59	11.52	0.54	14.85	0.58	7.94	0.38	13.25	0.56	17.16
	TTC	9.51	0.41	9.83	0.46	10.93	0.42	12.86	0.62	10.35	0.44	20.39
Tyr	TAT	8.67	0.61	8.43	0.47	10.37	0.54	4.93	0.50	5.18	0.44	12.09
	TAC	5.59	0.39	9.55	0.53	8.97	0.46	4.93	0.50	6.54	0.56	15.41
Gln	CAG	15.94	0.42	41.30	0.57	16.25	0.36	30.10	0.60	33.78	0.62	34.39
	CAA	22.38	0.58	31.19	0.43	28.58	0.64	19.70	0.40	20.70	0.38	12.03
Glu	GAG	14.83	0.32	29.22	0.49	15.97	0.31	27.64	0.54	33.23	0.62	39.98
	GAA	31.89	0.68	30.91	0.51	35.58	0.69	23.26	0.46	20.70	0.38	28.92
Lys	AAG	23.22	0.40	42.43	0.53	17.93	0.36	29.56	0.60	31.60	0.55	32.22
	AAA	34.13	0.60	38.21	0.47	32.50	0.64	19.70	0.40	26.15	0.45	24.04
Ile	ATA	6.99	0.19	10.96	0.39	6.16	0.19	5.75	0.24	7.90	0.27	7.28
	ATT	13.15	0.36	8.43	0.30	14.57	0.46	7.66	0.32	11.99	0.41	15.79
	ATC	15.94	0.44	8.99	0.32	11.21	0.35	10.67	0.44	9.26	0.32	21.07

aa	Codon	West Nile		Tick-borne encephalitis		Powassan_virus		Yellow fever virus		Wesselsbron_virus		Human
		1/1000	Fract	1/1000	Fract	1/1000	Fract	1/1000	Fract	1/1000	Fract	1/1000
Arg	CGC	9.52	0.14	7.30	0.08	10.24	0.15	8.56	0.11	5.83	0.07	10.68
	AGG	15.78	0.24	32.18	0.34	16.88	0.24	24.03	0.30	24.97	0.29	11.71
	AGA	20.95	0.32	24.88	0.26	20.20	0.29	24.86	0.31	30.24	0.35	11.72
	CGG	8.16	0.12	18.66	0.20	9.41	0.13	9.12	0.12	8.60	0.10	11.65
	CGA	5.98	0.09	6.22	0.07	6.37	0.09	5.25	0.07	6.38	0.07	6.24
Leu	CGT	5.71	0.09	6.22	0.07	7.20	0.10	7.18	0.09	9.71	0.11	4.63
	TTG	21.49	0.24	30.29	0.36	23.25	0.24	8.01	0.12	12.21	0.19	12.75
	TTA	4.08	0.05	4.87	0.06	2.77	0.03	5.25	0.08	6.94	0.11	7.43
	CTG	27.20	0.30	22.71	0.27	27.95	0.29	12.15	0.19	11.10	0.17	40.13
	CTA	9.79	0.11	5.41	0.06	8.30	0.09	7.73	0.12	8.60	0.13	7.04
Ser	CTT	10.34	0.11	7.84	0.09	17.16	0.18	16.57	0.26	16.65	0.25	13.01
	CTC	17.68	0.20	12.44	0.15	16.33	0.17	14.64	0.23	9.99	0.15	19.67
	AGT	12.51	0.21	12.17	0.12	11.90	0.20	23.76	0.29	24.14	0.27	12.05
	AGC	13.06	0.22	14.60	0.15	13.84	0.23	18.78	0.23	22.48	0.25	19.45
	TCG	5.71	0.10	11.90	0.12	5.54	0.09	3.04	0.04	4.16	0.05	4.47
Ala	TCA	14.96	0.25	27.31	0.28	12.73	0.21	9.12	0.11	7.49	0.08	12.00
	TCT	6.53	0.11	13.79	0.14	5.54	0.09	15.19	0.19	15.82	0.18	14.89
	TCC	5.98	0.10	18.66	0.19	10.52	0.18	12.15	0.15	15.26	0.17	17.63
	GCG	9.79	0.13	10.55	0.19	10.79	0.14	4.14	0.09	5.55	0.12	7.55
	GCA	19.04	0.25	20.55	0.37	19.65	0.25	14.09	0.30	12.76	0.28	16.00
Gly	GCT	25.30	0.33	12.98	0.23	21.04	0.27	18.78	0.40	18.31	0.40	18.57
	GCC	22.85	0.30	11.63	0.21	26.02	0.34	9.67	0.21	9.43	0.20	28.28
	GGG	14.96	0.18	23.80	0.30	24.91	0.25	16.30	0.16	15.26	0.18	16.48
	GGA	42.17	0.50	29.20	0.37	39.58	0.41	35.08	0.34	28.02	0.32	16.42
	GGT	9.25	0.11	11.36	0.14	12.73	0.13	26.80	0.26	22.75	0.26	10.80
Pro	GGC	17.41	0.21	14.87	0.19	20.48	0.21	24.86	0.24	20.26	0.23	22.56
	CCG	4.62	0.11	11.36	0.17	5.26	0.13	5.80	0.10	8.05	0.17	7.04
	CCA	19.04	0.45	24.34	0.37	17.16	0.41	19.06	0.33	15.54	0.33	16.84
	CCT	8.71	0.20	15.95	0.24	7.75	0.19	22.10	0.38	13.87	0.29	17.42
	CCC	10.34	0.24	13.52	0.21	11.62	0.28	11.33	0.19	9.71	0.21	20.03
Thr	ACG	11.43	0.16	11.90	0.17	13.29	0.21	5.80	0.15	7.49	0.18	6.15
	ACA	22.58	0.31	28.12	0.40	19.93	0.32	11.05	0.29	11.65	0.28	14.91
	ACT	15.23	0.21	14.06	0.20	11.90	0.19	11.05	0.29	13.04	0.31	13.01
	ACC	22.85	0.32	16.50	0.23	16.88	0.27	9.94	0.26	9.43	0.23	19.09
Val	GTG	37.81	0.46	23.53	0.60	42.35	0.48	14.36	0.33	8.60	0.22	28.56
	GTA	6.80	0.08	3.52	0.09	8.03	0.09	5.25	0.12	6.94	0.17	7.06
	GTT	15.78	0.19	5.14	0.13	18.27	0.21	15.19	0.34	15.26	0.38	10.98
	GTC	21.22	0.26	6.76	0.17	19.65	0.22	9.39	0.21	9.16	0.23	14.63
Asn	AAT	13.33	0.37	6.49	0.33	11.62	0.46	19.34	0.58	21.09	0.50	16.72
	AAC	22.85	0.63	12.98	0.67	13.84	0.54	13.81	0.42	20.81	0.50	19.17
Asp	GAT	20.40	0.47	5.95	0.28	18.82	0.39	23.48	0.55	21.09	0.55	21.98
	GAC	23.39	0.53	15.68	0.73	29.06	0.61	19.06	0.45	16.93	0.45	25.50
Cys	TGT	8.43	0.45	16.50	0.41	11.62	0.60	23.20	0.57	23.03	0.52	10.31
	TGC	10.34	0.55	23.53	0.59	7.75	0.40	17.40	0.43	21.64	0.48	12.55
His	CAT	7.62	0.38	7.03	0.35	12.73	0.51	35.91	0.65	36.63	0.59	10.69
	CAC	12.51	0.62	13.25	0.65	12.18	0.49	19.61	0.35	25.25	0.41	15.03
Phe	TTT	13.06	0.45	7.57	0.39	13.84	0.51	14.09	0.56	15.82	0.56	17.16
	TTC	16.05	0.55	11.63	0.61	13.56	0.49	11.05	0.44	12.21	0.44	20.39
Tyr	TAT	9.52	0.38	3.24	0.44	9.41	0.45	8.56	0.58	9.43	0.58	12.09
	TAC	15.23	0.62	4.06	0.56	11.62	0.55	6.08	0.42	6.94	0.42	15.41
Gln	CAG	14.69	0.54	25.15	0.58	17.71	0.65	19.89	0.41	14.71	0.34	34.39
	CAA	12.51	0.46	18.39	0.42	9.41	0.35	28.18	0.59	28.02	0.66	12.03
Glu	GAG	31.01	0.52	32.45	0.65	34.32	0.56	20.44	0.36	13.60	0.35	39.98
	GAA	29.11	0.48	17.31	0.35	27.40	0.44	37.02	0.64	25.80	0.65	28.92
Lys	AAG	31.83	0.58	21.36	0.56	24.91	0.51	14.92	0.39	15.54	0.33	32.22
	AAA	22.85	0.42	17.04	0.44	24.36	0.49	23.48	0.61	31.08	0.67	24.04
Ile	ATA	11.15	0.22	5.41	0.29	8.58	0.23	6.35	0.25	5.27	0.19	7.28
	ATT	21.22	0.43	6.49	0.35	11.07	0.30	10.50	0.41	9.43	0.34	15.79
	ATC	17.41	0.35	6.49	0.35	17.71	0.47	8.56	0.34	13.32	0.48	21.07

aa	Codon	Hepatitis C virus		Classical swine fever virus		Bovine viral diarrhea virus 1		Human
		1/1000	Fract	1/1000	Fract	1/1000	Fract	1/1000
Arg	CGC	20.99	0.19	6.34	0.06	2.70	0.03	10.68
	AGG	16.54	0.15	27.82	0.27	27.50	0.35	11.71
	AGA	12.72	0.12	36.12	0.35	28.97	0.37	11.72
	CGG	22.58	0.21	10.74	0.10	8.84	0.11	11.65
	CGA	17.18	0.16	12.45	0.12	6.87	0.09	6.24
	CGT	18.45	0.17	10.49	0.10	3.19	0.04	4.63
Leu	TTG	10.18	0.11	7.81	0.11	21.85	0.27	12.75
	TTA	7.95	0.09	9.52	0.14	13.75	0.17	7.43
	CTG	21.95	0.24	12.20	0.18	20.87	0.26	40.13
	CTA	9.22	0.10	13.91	0.20	13.50	0.17	7.04
	CTT	22.58	0.24	19.03	0.27	7.12	0.09	13.01
	CTC	20.67	0.22	6.83	0.10	3.68	0.05	19.67
Ser	AGT	15.90	0.20	25.87	0.36	10.80	0.14	12.05
	AGC	21.31	0.27	21.96	0.31	14.73	0.19	19.45
	TCG	9.22	0.12	0.98	0.01	7.86	0.10	4.47
	TCA	7.95	0.10	7.08	0.10	19.40	0.25	12.00
	TCT	12.72	0.16	10.25	0.14	12.77	0.16	14.89
	TCC	13.04	0.16	5.37	0.08	12.77	0.16	17.63
Ala	GCG	13.99	0.22	1.95	0.05	5.65	0.16	7.55
	GCA	13.99	0.22	11.96	0.30	16.45	0.46	16.00
	GCT	22.58	0.35	9.03	0.23	6.14	0.17	18.57
	GCC	13.99	0.22	16.59	0.42	7.37	0.21	28.28
Gly	GGG	27.99	0.28	17.81	0.18	23.82	0.33	16.48
	GGA	24.49	0.25	30.99	0.32	21.85	0.30	16.42
	GGT	20.99	0.21	26.11	0.27	13.50	0.19	10.80
	GGC	25.45	0.26	22.69	0.23	12.52	0.17	22.56
Pro	CCG	12.40	0.15	1.46	0.03	11.05	0.20	7.04
	CCA	19.72	0.23	17.33	0.31	23.57	0.42	16.84
	CCT	26.40	0.31	20.99	0.38	10.07	0.18	17.42
	CCC	26.08	0.31	15.86	0.29	11.78	0.21	20.03
Thr	ACG	7.00	0.15	1.71	0.04	9.82	0.13	6.15
	ACA	12.72	0.26	15.86	0.33	28.73	0.37	14.91
	ACT	13.68	0.28	16.59	0.34	17.43	0.23	13.01
	ACC	14.63	0.30	14.15	0.29	21.36	0.28	19.09
Val	GTG	14.31	0.25	9.27	0.21	15.71	0.39	28.56
	GTA	11.45	0.20	11.71	0.26	12.77	0.31	7.06
	GTT	17.18	0.30	16.59	0.37	7.12	0.17	10.98
Asn	GTC	14.95	0.26	6.83	0.15	5.16	0.13	14.63
	AAT	10.81	0.53	15.62	0.44	12.52	0.44	16.72
	AAC	9.54	0.47	20.01	0.56	16.20	0.56	19.17
Asp	GAT	14.95	0.49	25.62	0.55	4.91	0.42	21.98
	GAC	15.59	0.51	21.23	0.45	6.87	0.58	25.50
Cys	TGT	20.36	0.46	14.15	0.53	14.24	0.52	10.31
	TGC	23.54	0.54	12.69	0.47	13.01	0.48	12.55
His	CAT	23.85	0.44	22.45	0.50	13.01	0.51	10.69
	CAC	30.85	0.56	22.69	0.50	12.28	0.49	15.03
Phe	TTT	9.54	0.45	12.45	0.70	8.59	0.66	17.16
	TTC	11.45	0.55	5.37	0.30	4.42	0.34	20.39
Tyr	TAT	10.18	0.48	12.45	0.47	14.73	0.48	12.09
	TAC	10.81	0.52	13.91	0.53	15.71	0.52	15.41
Gln	CAG	17.49	0.42	22.21	0.39	30.69	0.57	34.39
	CAA	23.85	0.58	35.14	0.61	22.83	0.43	12.03
Glu	GAG	8.27	0.34	19.77	0.38	17.43	0.57	39.98
	GAA	16.22	0.66	31.72	0.62	13.01	0.43	28.92
Lys	AAG	5.41	0.35	13.67	0.35	32.65	0.52	32.22
	AAA	10.18	0.65	24.89	0.65	30.69	0.48	24.04
Ile	ATA	9.54	0.43	9.27	0.27	20.38	0.48	7.28
	ATT	5.73	0.26	17.08	0.50	11.54	0.27	15.79
	ATC	7.00	0.31	8.05	0.23	10.56	0.25	21.07

Table 5. Genome signature of various genera of Flaviviridae.

Genus	Subgenus	Species	Nucleotide signature	
			High	low
Flavivirus	Dengue virus group	Dengue virus 1	A	T,C
		Dengue virus 2		
		Dengue virus 3		
	Japanese encephalitis virus group	Japanese encephalitis virus Species: Murray Valley encephalitis virus West_Nile_virus	A,G	T,C
Hepacivirus	Yellow fever virus group	Yellow fever virus		
	Tick-borne encephalitis virus group	Wesselsbron virus Tick-borne encephalitis virus Powassan virus	G	T,C
Pestivirus		Hepatitis C virus	C,G	A,T
		Classical swine fever virus	A	C
		Bovine viral diarrhea virus 1		

Discussion

Flaviviruses are transmitted by mosquitoes, ticks, or directly between vertebrate hosts. It was demonstrated (Jenkins et al., 2001) that those viruses associated with ticks have a significantly lower G+C content than non-vector-borne flaviviruses. In contrast, mosquito-borne viruses had an intermediate G+C content which was not significantly different from those of the other two groups.

It has been proposed by Wright (1990) that most variable ENC is a more direct measure of synonymous codon bias. The ENC values of different Flaviviridae genes vary from 51.84 to 58.65, with a mean of 54.004 and S.D. of 1.850. We found that all the ENC values of these strains were more than 50. The ENC values range from 20 to 61. In an extremely biased gene where only one codon is used for each amino acid, this value would be 20; in an unbiased gene, it would be 61. We conclude that there is not much codon usage bias in Flaviviridae genome.

It has already been shown that Cysteine (Lobry et al., 1994; Rodríguez-Maseda et al., 1994; Musto et al., 1997) and Leucine are the least and the most frequent amino acids among all of the species of Flaviviridae. *Hepacivirus* genus and *Tick-borne*

encephalitis virus group subgenus show major exceptions at amino acid composition in polyprotein sequences. For example, Glutamic acid and Lysine are the lowest number in *Hepacivirus* genus while have the highest number in other species. CGN and AGN codons encode Arginine. *Tick-borne encephalitis* virus group subgenus with the highest amount of G has the highest number of Arginine residue. This group also has a lesser amount of Asparagine, Isoleucine, Lysine and Phenylalanine whilst has higher number of Valine, Arginine and Glycine residues. Moreover, it is observed that Tryptophan has a maximum in *Japanese encephalitis* virus species and a minimum number in *Pestivirus* genus. Also in *Pestivirus* genus, Tyrosine has the highest number between all of the species.

We have found that Arg, Leu and Ser have six-fold coding degeneracy. All the examined species, excluding *Hepatitis C* virus, use AGA for Arg most frequently, while *Hepatitis C* virus most commonly uses CGG, CGC, CGA, CGT and CGG codons are less frequent in the majority of members. The most and the least commonly used codons for Leu and Ser are different.

Ala, Gly, Pro, Thr and Val have four fold coding degeneracy. For coding Ala, GCG is used less frequently while GCA and GCT are used more in

many of the species. For Gly, *Japanese encephalitis*, *Hepatitis C* virus, and *Bovine viral diarrhea* virus mostly use GGG, while other species mostly use GGA. ACG (for Thr) and CCG (for Pro) have the lowest frequency in most of species. GTG is used for Val most often in *Bovine viral diarrhea* virus, *West Nile* virus, *Tick-borne encephalitis* virus, *Powassan* virus, *Dengue 2* virus, *Japanese encephalitis* virus and *Murray Valley encephalitis* virus while in other species GTT is mainly used.

Asn, Asp, Cys, His, Phe and Tyr have two-fold codon degeneracy. Mostly for Asn, Cys and Phe, the usage of codons with T ending is higher than codons with C ending. For Asp, Gln, Glu, Lys and His, the usage of two codons is approximately the same. Ile is the only amino acid which has three-fold codon degeneracy. ATA, ATT and ATC are the least commonly used synonymous codons in the members of *Flavivirus*, *Hepacivirus*, and *Pestivirus* genus, respectively. Based on our findings, *Flavivirus* genus can be classified in four subgenera.

We have analyzed the codon usage, base and amino acid compositions of 13 species of Flaviviridae. Results reveal that there is not much codon usage bias in Flaviviridae genome. In the future studies, analysis of different factors, affecting codon usage variation, can increase our knowledge about processes involved in the Flaviviridae evolution and the selective forces that significantly influence codon usage bias.

Acknowledgment

The support of this research by University of Isfahan and Shiraz is highly acknowledged.

References

- 1- Bennetzen J. L. and Hall B. D. (1982) Codon selection in yeast. *The Journal of Biological Chemistry* 257: 3026-3031.
- 2- Gould E. A., de Lamballerie X., Zanotto P. M. and Holmes E. C. (2001) Evolution, epidemiology, and dispersal of flaviviruses revealed by molecular phylogenies. *Advances in Virus Research* 57: 71-103.
- 3- Jenkins G. M., Pagel M., Gould E. A., de A Zanotto P. M. and Holmes E. C. (2001) Evolution of base composition and codon usage bias in the genus *Flavivirus*. *Journal of Molecular Evolution* 52:383-390.
- 4- Grantham R., Gautier C., Gouy M. and Jacobzone M., Mercier R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Research* 9: 43-74.
- 5- Hou Z. C. and Yang N. (2002) Analysis of factors shaping *S. pneumoniae* codon usage. *Chinese Journal Genet* 29: 747-752.
- 6- Kerr A. R., Peden J. F. and Sharp P. M. (1997) Systematic base composition variation around the genome of *Mycoplasma genitalium*, but not *Mycoplasma pneumoniae*. *Molecular Microbiology* 25: 1177-1179.
- 7- Lobry J. R. and Gautier C. (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Research* 22: 3174-3180.
- 8- Musto H., Caccio` S., Rodriguez-Maseda H. and Bernardi G. (1997) Compositional constraints in the extremely GC-poor genome of *Plasmodium falciparum*. *Memórias do Instituto Oswaldo Cruz* 92: 835-841.
- 9- Naya H., Romero H., Carels N., Zavala. A. and Musto H. (2001) Translational selection shapes codon usage in the GC-rich genomes of *Chlamydomonas reinhardtii*. *FEBS Letters* 501: 127-130.
- 10- Powell J. R. and Moriyama E. N. (1997) Evolution of codon usage bias in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America* 94: 7784-7790.
- 11- Rodríguez-Maseda H. and Musto H. (1994) The compositional compartments of the nuclear genomes of *Trypanosoma brucei* and *T. cruzi*. *Gene* 151: 221-224.
- 12- Romero H., Zavala A. and Musto H. (2000) Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Research* 28: 2084-2090.
- 13- Shackelton L. A., Parrish C. R. and Holmes E. C. (2006) Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *Journal of Molecular Evolution* 62: 551-563.
- 14- Sharp P. M. and Li W. H. (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* 15: 1281-1295.
- 15- Sau K., Gupta S. K., Sau S. and Ghosh T. C. (2005) Synonymous codon usage bias in 16 *Staphylococcus aureus* phages: implication in phage therapy. *Virus Research* 113: 123-131.
- 16- Shen H. B. and Chou K. C. (2008) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Analytical Biochemistry* 373: 386-388.
- 17- Sueoka K. and Kawanishi Y. (2000) DNA G+C content of the third codon position and codon usage biases of human genes. *Gene* 261: 53-62.
- 18- Vicario S., Moriyama E. N. and Powell J. R. (2007) Codon usage in twelve species of *Drosophila*. *BMC Evolutionary Biology* 7: 226.
- 19- Van Regenmortel M. H. V., Fauquet C. M., Bishop D. H. L., Carstens E. B., Estes M. K., Lemon S. M., Maniloff J., Mayo M. A., McGeoch D. J., Pringle C. R. and Wickner R. B. (2000) *Virus Taxonomy*. Seventh Report of the International Committee on

- Taxonomy of Viruses, San Diego: Academic Press.
- 20- Xu X. Z., Liu Q. P., Fan L. J., Cui X. F. and Zhou X. P. (2008) Analysis of synonymous codon usage and evolution of begomoviruses. Journal of Zhejiang University. Science. B 9: 667-674.
- 21- Zhuo-Cheng H. and Ning Y. (2003) Factors affecting codon usage in *Yersinia pestis*. Acta Biochimica et Biophysica Sinica 35: 580-586.
- 22- Wright F. (1990) The effective number of codons used in a gene. Gene 87: 23-29.