Available Online at http://www.recentscientific.com

ISSN: 0976-3031

International Journal of Recent Scientific Research
Vol. 5, Issue, 5, pp.965-969, May, 2014

International Journal of Recent Scientific Research

RESEARCH ARTICLE

CODON USAGE BIAS IN H1N1 NEURAMINIDASE: SELECTION OR MUTATIONAL BIAS?

Himangshu Deka, Supriyo Chakraborty*

Department of Biotechnology, Assam University, Silchar-788011, Assam, India

ARTICLE INFO

Article History:

Received 15th, April, 2014
Received in revised form 27th, April, 2014
Accepted 17th, May, 2014
Published online 28th, May, 2014

Key words:

Codon usage bias, H1N1, Influenza A, Neuraminidase, Synonymous codon

ABSTRACT

Influenza A virus (IAV) has been a major concern worldwide as a cause of high mortality and morbidity. In the present study, the complete coding regions of viral neuraminidase (NA) gene of IAV subtype H1N1 reported from India were analyzed for the possible codon usage bias using statistical and bioinformatics tools. A total of 34 NA coding sequences were used in the study. The results show a low bias in the coding region of the NA gene sequences. The RSCU values suggest a very low preference of the codons having dinucleotide CpG whereas most of the codons showed a preferred use of the dinucleotides CpA and TpA. The results suggest that there exists a balance between mutational pressure and natural selection to shape the codon usage bias in the IAV subtype which helps the virus adapt to different host conditions.

INTRODUCTION

The exponential increase in the volume of sequence information during the early '90s facilitated for the first time the detailed statistical analyses of codon usage (Grantham, Gautier *et al*. 1980). It has been established that there exists a bias in the usage of synonymous codons in the biological system ranging from prokaryotes to complex organisms including the viruses. With the rapid availability of vast number of sequences after whole genome sequencing of large number of species, scientists are now trying to look the codon bias phenomena in holistic manner. Accordingly, on a global basis, investigators have focussed research interest in the context of codon bias phenomenon in specific genes as well as whole genome (Grantham, Gautier *et al*. 1980; Plotkin and Kudla 2011).

The major concern regarding the negative-stranded RNA virus, Influenza A virus, can be understood by the fact that roughly one-fifth of the human populations are infected by the virus every year, causing significant mortality and negative economic impacts on society worldwide. Among the major influenza pandemics two was caused by the H1N1 strain, one in the year 1918 and the latest in 2009 (Cox, Black *et al*. 1989; Dawood, Jain *et al*. 2009). The first outbreak of the H1N1 of this century originated in Mexico in 2009 which later spread to about 207 countries worldwide with a death toll of more than 7,800. Apart from these two several other outbreaks of H1N1 have been reported in 1950s and in 1970s (Goni, Iriarte *et al*. 2012).

While human immune system develops resistance against most pathogens upon exposure to them, IAV poses serious threat to the host immunity by presenting a moving antigenic target. This process, termed as antigenic drift, helps it to escape the specific immunity caused by earlier infections. Drift is the result of the selective fixation of mutations in the hemagglutinin (HA) and neuraminidase (NA) genes (Goni, Iriarte *et al*. 2012). The viral neuraminidase (NA) is frequently used as an antigenic determinant found on the surface of the Influenza virus. While, in some other variants of the influenza neuraminidase confers more virulence to the virus than others making it as a potential drug target for the prevention of the spread of influenza infection (Liu, Eichelberger *et al*. 1995).

Apart from genetic drift, which tends to be a slow evolutionary process, shift is another process by which IAV evolves. The genetic information is shared between the IAV strains triggering rapid evolutionary change in the virus. As happened in case of the 2009 pandemic, such rapid change may result in cross-species shift (Dawood, Jain *et al*. 2009).

Several workers have reported that the overall codon usage bias in RNA viruses is low and there is little variation in bias between genes (Jenkins and Holmes 2003; Gu, Zhou *et al*. 2004; Goni, Iriarte *et al*. 2012). The low codon usage bias in the RNA viruses is attributed to GC compositional properties and dinucleotide content in these viruses. Mutational bias has been projected as the main factor that drives the codon usage variation among the influenza A viruses which are phylogenetically conserved (Gu, Zhou *et al*. 2004).

The analysis of synonymous codon usage is used to investigate the interplay between the mutational pressure exerted by the pathogen on host and the selection pressure on the former by the latter (Jenkins and Holmes 2003). Over the years many authors have reported a number of tools which can be used to measure codon usage bias across genes and genomes. Among these measures, GC content, relative synonymous codon usages (RSCU), effective number of codons (ENC) are some most widely used parameters for codon bias study. RSCU measures the frequency of a particular codon compared to the expected frequency if all synonymous codons are used equally (Sharp and Li 1987; Novembre 2002). While ENC measures the deviation of the codon usage from equal usage of the synonymous codons in a gene or genome, it does not give the direction of bias (Plotkin and Dushoff 2003).

MATERIALS AND METHODS

Datasets

In this study, a total of 34 complete coding sequences of the neuraminidase (NA) gene of human-host derived influenza A virus subtype H1N1 reported from India were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/). The serial numbers (SN), accession numbers and other information are presented in **table 1**.

Parameters for codon usage bias study

To examine the synonymous codon usage in the genes RSCU values were calculated. RSCU is defined as the ratio of the observed frequency to the expected frequency if all the synonymous codons for

* Corresponding author: **Supriyo Chakraborty**
Department of Biotechnology

those amino acids are used equally (Sharp and Li 1987). If the RSCU value of a codon is more than 1.0 it is said to have a positive codon usage bias, while a value of less than 1.0 means a negative codon usage bias. When the RSCU value is close to 1.0, it means that this codon is chosen randomly and equally with other synonymous codons.

Table 1 Information of the complete coding sequences of the 34 NA genes

SN	Accession No	Gene Length
1	KF280657	1410
2	KF280665	1410
3	KF280673	1410
4	KF280681	1410
5	KF280689	1410
6	KF280697	1410
7	KF280705	1410
8	KF280713	1410
9	KF280721	1410
10	KF280729	1410
11	KF280737	1410
12	KF280745	1410
13	KF280753	1410
14	JX262202	1410
15	JX262201	1410
16	HM460506	1413
17	JF265672	1410
18	JF265671	1410
19	HM241726	1411
20	HM241719	1429
21	HM241712	1428
22	HM241705	1428
23	CY088710	1428
24	CY088703	1428
25	CY088696	1413
26	CY088689	1428
27	CY088682	1410
28	CY088675	1438
29	CY088668	1421
30	CY088661	1428
31	CY088654	1421
32	CY088647	1429
33	CY088640	1428
34	CY088633	1438

The effective number of codons (ENC) is estimated to quantify the synonymous codon usage across the target sequence which is given below:

$$ENC = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}$$

where, F_k (k = 2, 3, 4 or 6) is the average of the F_k values for k-fold degenerate amino acids. The F value denotes the probability that two randomly chosen codons for an amino acid with two codons are identical. The values of ENC range from 20 (when only one codon is used per amino acid) to 61 (when all synonymous codons are equally used for each amino acid) (Wright 1990; Novembre 2002).

GC_{3s} is the frequency of the nucleotides G+C at the synonymous 3rd positions of the codons excluding Met, Trp and the termination codons. Similarly $GC_{1s\ and}$ GC_{2s} represent G+C frequency at 1st and 2nd codon positions. GC_{3s} is a good indicator of the extent of base composition bias.

Gene expressivity was measured by codon adaptation index (CAI) as given by Sharp and Li (Sharp and Li 1987). CAI has been used as a simple and effective measure of the overall synonymous codon usage bias of a gene. CAI was originally proposed to provide a normalized estimate that can be used across genes and species, ranging from 0 to 1. The boundary values refer to the cases in which only the most frequent codons (CAI = 1) or only the least frequent codons (CAI = 0) are used within a gene. CAI is given by the following formula:

$$CAI = exp\frac{1}{L}\sum_{k=1}^{L} \ln w_{c(k)}$$

where, L is the number of codons in the gene and $w_{c(k)}$ is the ω value for the k-th codon in the gene.

Frequency of optimal codon (Fop) in a codon is used as an index to show the optimization level of synonymous codon choice in each gene to translation process (Ikemura 1982). Fop is defined as the ratio of total number of optimal codons in a gene to the total number of synonymous as well as non synonymous codons in that gene.
The codon usage bias measures namely RSCU, ENC, GCs and CAI for each coding sequence were estimated by using a Perl program developed by SC.

RESULT AND DISCUSSION

The nucleotide content and the overall GC content at the three codon positions reveal that most of the preferential codons use A at the synonymous third codon position. The overall percentage of A, T, G, C and overall GC content in the three codon positions are shown in table 2. Throughout the accessions A% is higher than the rest of the nucleotides with an average value of 30. As evident from the results, GC3% is higher than GC1% and GC2% in all the accessions with a mean of 46.0%. (Fig 1).

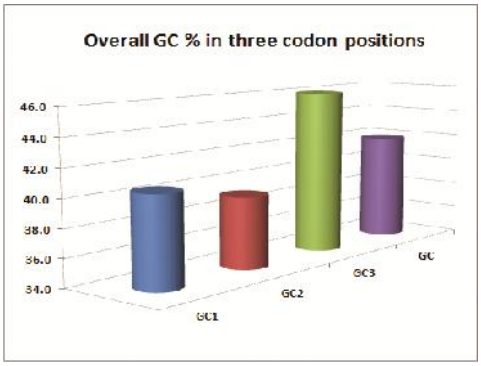

Fig 1 GC content in three codon positions: Percentage of GC content in the three codon positions (GC1, GC2 and GC3) and overall GC content in the coding sequences of the NA gene used in this study.

Previous studies have revealed that influenza A virus strains infecting human hosts since 1918 have been selected under strong pressure to reduce the frequency of CpG in its genome. The possible explanation for low CpG may be immunologic escape as unmethylated CpGs are recognized by the host's innate immune system as a pathogen signature (Greenbaum, Levine *et al.* 2008). Marked CpG deficiency has also been reported in several RNA viruses including H1N1 (Greenbaum, Levine *et al.* 2008; Wong, Smith *et al.* 2010). Thus, escape from the host antiviral response could act as a selective pressure contributing to codon usage in H1N1.

To peruse the possible effects of CpG under-represented on codon usage bias, the RSCU values were examined for the eight codons having dinucleotide combination of CpG (CCG, GCG, TCG, ACG, CGC, CGG, CGT, and CGA). Our analysis indicated that CGA and CGG were more preferred but the rest of the codons containing CpG are markedly suppressed. Similarly, out of six codons containing TpA (TTA, CTA, GTA, TAT, TAC, ATA) only two codons are preferred (i.e., TAT and ATA). However, codons containing CpA (TCA, CCA, ACA, GCA, CAA, CAG, CAT and CAC) show a remarkable preferentiality over others with five preferred codons out of eight. Codons containing the dinucleotide TpG (TTG, GTG, TGT, TGC, and CTG) also show a low preferentiality. However, it was interesting to note that most of the preferred codons contain dinucleotide CpA (Fig 2).

Table 2 Nucleotide composition of the 34 coding sequences of NA gene

SN	A%	T%	G%	C%	GC%	GC1%	GC2%	GC3%	ENC
1	31.8	26.1	23.6	18.5	42.1	40.6	39.4	46.4	59
2	32.1	26.0	23.2	18.7	41.9	41.1	38.7	46.0	59
3	32.1	26.0	23.3	18.6	41.9	40.4	39.1	46.3	58
4	32.1	26.0	23.4	18.6	42.0	40.9	39.1	46.0	59
5	32.1	26.0	23.3	18.7	42.0	40.9	39.1	46.0	59
6	32.1	26.1	23.3	18.5	41.8	40.4	38.7	46.4	58
7	32.1	26.0	23.3	18.6	41.9	41.3	38.9	45.5	59
8	31.9	26.0	23.5	18.7	42.1	41.3	39.1	46.0	59
9	31.8	26.1	23.5	18.7	42.1	40.9	38.9	46.6	59
10	32.0	26.0	23.5	18.6	42.1	40.6	39.1	46.4	59
11	32.1	26.0	23.3	18.6	41.9	40.4	39.1	46.1	58
12	31.8	26.2	23.6	18.4	42.0	40.0	39.8	46.2	58
13	32.5	26.0	22.8	18.7	41.5	40.4	38.3	45.7	58
14	32.0	26.1	23.3	18.6	41.9	40.6	39.1	46.1	59
15	32.0	26.1	23.3	18.6	41.9	40.6	39.1	46.1	59
16	32.7	26.3	23.2	17.8	41.0	38.0	38.6	46.3	57
17	32.1	26.1	23.3	18.6	41.8	40.6	38.9	46.0	59
18	32.0	26.0	23.3	18.7	42.0	40.8	39.1	46.0	59
19	31.9	26.1	23.4	18.6	42.0	40.8	39.3	46.0	59
20	32.1	26.2	23.1	18.7	41.8	40.3	38.9	46.2	58
21	31.9	26.3	23.2	18.6	41.7	39.9	39.5	45.7	58
22	31.9	26.3	23.2	18.6	41.8	40.3	39.5	45.7	58
23	31.9	26.1	23.2	18.8	41.7	40.0	39.9	45.6	59
24	31.9	26.3	23.2	18.6	41.8	40.3	39.5	45.5	58
25	31.8	26.2	23.4	18.6	42.0	40.6	39.5	46.0	59
26	31.9	26.2	23.2	18.7	41.9	40.5	39.5	45.7	59
27	31.8	25.9	23.5	18.8	42.3	41.3	39.6	46.1	59
28	31.8	26.4	23.1	18.7	41.8	40.4	39.2	45.7	59
29	31.8	26.2	23.4	18.6	42.0	40.5	39.3	46.1	59
30	31.9	26.2	23.2	18.6	41.8	40.3	39.5	45.8	58
31	32.0	26.0	23.3	18.6	41.9	40.7	39.5	45.7	59
32	32.1	26.1	23.1	18.7	41.7	40.3	39.0	46.0	58
33	31.9	26.2	23.2	18.7	41.8	40.3	39.5	45.7	58
34	31.8	26.4	23.1	18.7	41.8	40.4	39.2	45.7	58

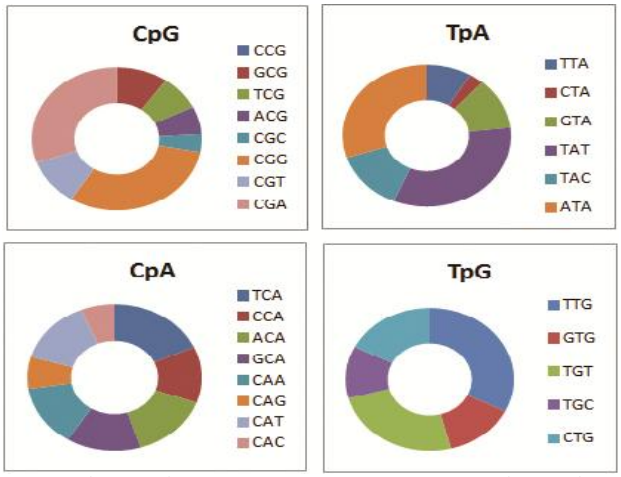

Fig 2 Preference of the dinucleotide in the codons. Most of the preferred codons use CpA while frequency of CpG is very less.

The general trend of ENC values is consistent throughout, (range 57.0-59.0) with an average of 58.6 and standard deviation of 0.0261. The high ENC values signify that the majority of the NA genes of H1N1 do not show a strong codon bias. This is in accordance with the previously published literature (Comeron and Aguade 1998; Jenkins and Holmes 2003; Zhou, Gu *et al*. 2005). The published data suggest that the reason for weak bias in different RNA viruses may be a strategy of these viruses to replicate efficiency in the vertebrate host cells with distinct codon choices (Sharp and Li 1987; Zhang, Wang *et al*. 2011).

Highly expressed genes tend to use limited number of codons and show a tendency of high biasness towards those codons.

The CAI value directly corresponds to the expression of the genes. It has been used to measure the extent of codon bias in a gene to examine the adaptation of its codons towards the codon usage of highly expressed genes (Sharp and Li 1987).

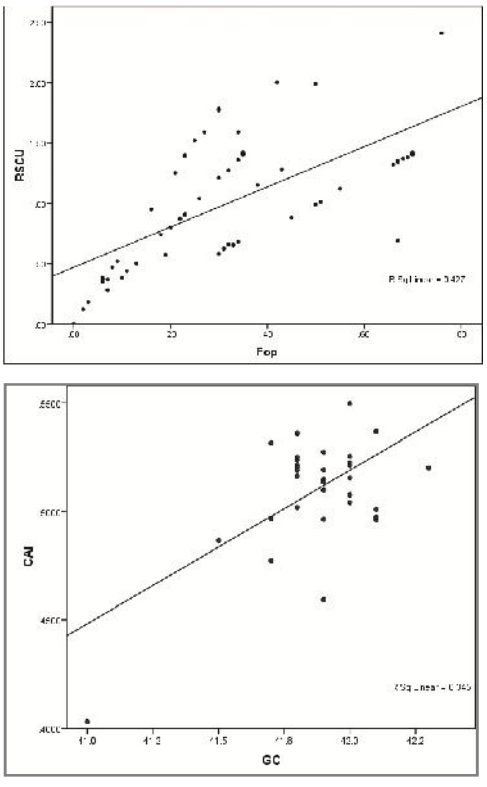

Fig 3 Correlation between a) RSCU and Fop and b) between GC and CAI values in the coding sequences of the NA gene

A value close to 1.0 indicates very high expression while lower values suggest a low expression and hence low codon bias (Grantham, Gautier *et al*. 1980; Sharp and Li 1987).

The CAI values in the present study are in the range of 0.4031-0.5496 with an average of 0.5100 and standard deviation of

ENC values along the Y-axis. If the GC3% is the only major factor playing role in the codon choice, the curve of the predicted values will lie above the ENC plots (Wright 1990).

The plot shows most of the points lying on inner side while a few points lying on outer side of the curve indicating that

Table 3 Synonymous codon usage pattern in the 34 coding sequences

AA	Codon	RSCU*	N*	Fop*	AA	Codon	RSCU*	N*	Fop*
Ala	GCA	**1.36**	**3**	**0.34**	Leu	TTA	0.37	3	0.06
	GCC	0.91	2	0.23		TTG	**1.77**	**14**	**0.30**
	GCG	0.44	1	0.11		CTT	1.52	12	0.25
	GCT	1.27	**3**	0.32		CTC	1.25	10	0.21
Arg	CGT	0.47	3	0.08		CTA	0.12	1	0.02
	CGC	0.18	1	0.03		CTG	0.95	7	0.16
	CGA	0.38	2	0.06	Lys	AAA	**1.34**	**19**	**0.67**
	CGG	1.39	8	0.23		AAG	0.66	9	0.33
	AGA	**2.00**	**12**	**0.42**	Phe	TTT	0.63	4	0.31
	AGG	1.59	9	0.34		TTC	**1.40**	**10**	**0.70**
Asn	AAT	**1.37**	**22**	**0.68**	Pro	CCT	0.38	1	0.10
	AAC	0.66	10	0.32		CCC	**2.41**	**6**	**0.76**
Asp	GAT	**1.12**	**10**	**0.55**		CCA	1.21	3	0.30
	GAC	0.88	8	0.45		CCG	0.00	0	0.00
Cys	TGT	**1.35**	**12**	**0.67**	Ser	TCT	0.37	2	0.07
	TGC	0.65	6	0.33		TCC	1.59	9	0.27
Gln	CAA	**1.32**	**24**	**0.66**		TCA	**1.78**	**10**	**0.30**
	CAG	0.68	12	0.34		TCG	0.35	2	0.06
Glu	GAA	0.99	**8**	0.50		AGT	1.40	8	0.23
	GAG	**1.01**	**8**	**0.51**		AGC	0.52	3	0.09
Gly	GGT	0.87	5	0.22	Thr	ACT	0.90	3	0.23
	GGC	1.04	6	0.26		ACC	1.41	**5**	0.35
	GGA	**1.40**	**8**	**0.35**		ACA	**1.42**	**5**	**0.35**
	GGG	0.69	4	0.67		ACG	0.28	1	0.07
His	CAT	**1.38**	**13**	**0.69**	Tyr	TAT	**1.42**	**8**	**0.70**
	CAC	0.62	6	0.31		TAC	0.58	3	0.30
Ile	ATT	1.15	8	**0.38**	Val	GTT	0.80	3	0.20
	ATC	0.57	4	0.19		GTC	**1.99**	**8**	**0.50**
	ATA	**1.28**	**9**	0.43		GTA	0.50	2	0.13
						GTG	0.74	3	0.18

*****RSCU, N and Fop values are mean values; AA means Amino acid and N stands for No of codons used.
The preferentially used codons for each amino acid are described in bold

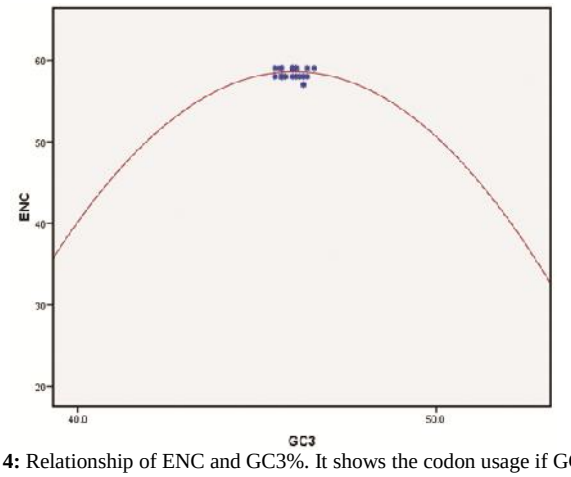

Fig 4: Relationship of ENC and GC3%. It shows the codon usage if GC compositional constraints accounts for codon usage bias alone

0.0261. This suggests the absence of a strong bias in the gene under study.

Correlation analysis was performed between GC content, ENC, CAI, RSCU and Fop values. There was a strong positive correlation between GC content and ENC ($r=0.700$, $p<0.01$) and also between GC and CAI ($r=0.587$, $p<0.01$) (Fig 3b). A strong positive correlation was observed between RSCU and Fop values ($r= 0.653$, $p<0.01$) (Fig 3a) and between RSCU and ENC ($r= 0.361$, $p<0.05$). No significant correlation was found between RSCU and CAI.

The ENC plot (Fig 4) was constructed to investigate the general pattern of synonymous codon usage. The plot was constructed by taking the GC3% values along the X-axis and

mutational pressure is not the sole force acting in the codon usage bias in the NA gene. There possibly exists a balance between mutational bias and natural selection to shape the codon usage which allows the virus to re-adapt its codon usage to different host environments over time (Zhou, Gu *et al*. 2005; Goni, Iriarte *et al*. 2012).

CONCLUSION

Natural selection and mutational pressure are two major factors which have been reported to affect codon usage bias in various organisms (Sharp and Li 1987). Earlier studies have revealed that mutational pressure, rather than natural selection, is the main factor playing crucial role in shaping the codon usage in most RNA viruses. Apart from mutation pressure in determining patterns of codon usage bias in RNA viruses, the analysis has revealed that the virus is under host immune selection pressure.

Vector-borne RNA viruses are said to have a lower codon usage bias than other RNA viruses (Jenkins and Holmes 2003). One possible explanation that can be attributed here is that a low bias is advantageous to viruses replicating in two different cell types with potentially distinct codon preferences. The replication cycle of IAV is dependent on host machinery and hence the viral replication is affected by the codon usage in the host as well as in the viral genomes. As in case of other RNA viruses, mutation rate of IAV is very high and the effects of codon usage bias too small for natural selection to operate efficiently (Brown 1997). RNA secondary structure may also influence the codon choice in synonymous sites (Simmonds and Smith 1999). The viral neuraminidase presents one of the

most important target sites for human immune system (Plotkin and Dushoff 2003). Hence, detailed information about the synonymous codon usage profile may aid in the development of vaccines against the virus.

Acknowledgement

The authors are grateful to Assam University, Silchar-788011, Assam and India for providing the research facility.

DISCLOSURE

The authors do not have any competing interest. The authors did not avail any financial assistance from any source in undertaking the present study.

References

Brown, A. J. (1997). "Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population." Proc Natl Acad Sci U S A 94(5): 1862-1865.

Comeron, J. M. and M. Aguade (1998). "An evaluation of measures of synonymous codon usage bias." J Mol Evol 47(3): 268-274.

Cox, N. J., R. A. Black, *et al*. (1989). "Pathways of evolution of influenza A (H1N1) viruses from 1977 to 1986 as determined by oligonucleotide mapping and sequencing studies." J Gen Virol 70 (Pt 2): 299-313.

Dawood, F. S., S. Jain, *et al*. (2009). "Emergence of a novel swine-origin influenza A (H1N1) virus in humans." N Engl J Med 360(25): 2605-2615.

Goni, N., A. Iriarte, *et al*. (2012). "Pandemic influenza A virus codon usage revisited: biases, adaptation and implications for vaccine strain development." Virol J 9: 263.

Grantham, R., C. Gautier, *et al*. (1980). "Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type." Nucleic Acids Res 8(9): 1893-1912.

Greenbaum, B. D., A. J. Levine, *et al*. (2008). "Patterns of evolution and host gene mimicry in influenza and other RNA viruses." PLoS Pathog 4(6): e1000079.

Gu, W., T. Zhou, *et al*. (2004). "The relationship between synonymous codon usage and protein structure in Escherichia coli and Homo sapiens." Biosystems **73**(2): 89-97.

Ikemura, T. (1982). "Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs." J Mol Biol 158(4): 573-597.

Jenkins, G. M. and E. C. Holmes (2003). "The extent of codon usage bias in human RNA viruses and its evolutionary origin." Virus Res 92(1): 1-7.

Liu, C., M. C. Eichelberger, *et al*. (1995). "Influenza type A virus neuraminidase does not play a role in viral entry, replication, assembly, or budding." J Virol 69(2): 1099-1106.

Novembre, J. A. (2002). "Accounting for background nucleotide composition when measuring codon usage bias." Mol Biol Evol **19**(8): 1390-1394.

Plotkin, J. B. and G. Kudla (2011). "Synonymous but not the same: the causes and consequences of codon bias." Nat Rev Genet 12(1): 32-42.

Plotkin, J. B. and J. Dushoff (2003). "Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus." Proc Natl Acad Sci U S A 100(12): 7152-7157.

Sharp, P. M. and W. H. Li (1987). "The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications." Nucleic Acids Res 15(3): 1281-1295.

Simmonds, P. and D. B. Smith (1999). "Structural constraints on RNA virus evolution." J Virol 73(7): 5787-5794.

Wong, E. H., D. K. Smith, *et al*. (2010). "Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus." BMC Evol Biol 10: 253.

Wright, F. (1990). "The 'effective number of codons' used in a gene." Gene 87(1): 23-29.

Zhang, J., M. Wang, *et al*. (2011). "Analysis of codon usage and nucleotide composition bias in polioviruses." Virol J **8**: 146.

Zhou, T., W. Gu, *et al*. (2005). "Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses." Biosystems 81(1): 77-86.

****