Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Virus Research 109 (2005) 33–37
Virus Research
www.elsevier.com/locate/virusres

Composition bias and genome polarity of RNA viruses

Prasert Auewarakul*

Department of Microbiology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand

Received 17 August 2004; received in revised form 13 September 2004; accepted 6 October 2004
Available online 18 November 2004

Abstract

I have observed a relationship between GC content in coding sequences of RNA viruses and their genome polarity. Positive-stranded RNA viruses have significantly higher GC contents than negative-stranded RNA viruses. Coding sequences of all negative-stranded RNA viruses are biased toward high A in coding strands (high T in genomes), while two distinct patterns were observed among positive-stranded RNA genomes. This finding suggests that RNA viruses with different genome polarity are under different mutational pressure, which may be a consequence of the difference in the strategies of viral genome expression and replication. The GC content directly affects the viral codon adaptation index using highly expressed human genes as the reference set, which may theoretically predict the efficiency of viral gene expression in human cells.

Keywords: RNA genome; GC content; Genome polarity

1. Main text

Genome composition of living organisms can vary widely. This is considered to be the result of the directional mutational bias toward GC or AT (Lobry and Sueoka, 2002; Sueoka, 1988; Sueoka, 1993). This directional mutational bias could theoretically be due to a bias in the copying error of viral RNA polymerase, selection pressure, or editing by host RNA-editing enzymes. Certain types of hypermutation have been described in a number of viruses (Cattaneo et al., 1988; Vartanian et al., 1994, 2002), and may also contribute to the viral genome composition. Genome composition bias in viruses has not been systematically analyzed. A global view of the pattern of viral genome composition bias may give us an insight into the complex evolution history of viruses and viral-host interactions.

GC content of genome has been shown to be a major contributing factor to the codon usage bias, which could affect expression efficiency (Aota and Ikemura, 1986; Chen et al., 2004; Francino and Ochman, 1999; Ikemura and Wada, 1991; Kanaya et al., 2001). It is interesting to see how GC content interacts with genome polarity and codon usage bias in RNA viruses. Genome composition and codon usage bias are particularly interesting in RNA viruses because the same RNA may be used as mRNA, genome, or anti-genome. The replication of RNA genome is also very different from DNA replication of the host using different polymerase enzymes and taking place in different environments, which may contribute to the mutational bias that drives the genome composition. RNA viruses with positive- and negative-stranded genome are very different in their strategies of genome expression and replication, which may contribute to mutational bias and selection pressure.

To do the analyses, I retrieved compositions of coding sequences of 50 viruses from a codon usage database available at Kazusa Research Institute, Japan (http://www.kazusa.or.jp/codon/cgi-bin/). The viruses were chosen to cover most viral families that cause diseases in human. When distinct separation between human and animal strains can be made, only human strains were included in the analyses, for example human influenza virus

* Corresponding author. Fax: +66 662 4184148.
E-mail address: sipaw@mahidol.ac.th.

0168-1702/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.virusres.2004.10.004

Table 1
List of the analyzed viruses

Name	Polarity	CAI	GC (%)	G (%)	C (%)	A (%)	T (%)
Astrovirus	+	0.362	46.18	23.18	22.99	29.67	24.14
Coronavirus	+	0.302	38.42	21.50	16.91	26.80	34.77
CoxsackievirusA9	+	0.398	46.97	24.47	22.49	29.18	23.84
Dengue virus type 1	+	0.353	46.38	25.88	20.49	31.96	21.65
Dengue virus type 2	+	0.361	45.8	25.21	20.58	33.21	20.98
Dengue virus type 3	+	0.359	46.47	25.91	20.55	32.12	21.40
Dengue virus type 4	+	0.366	46.91	26.31	20.59	31.03	22.06
Enterovirus 71	+	0.399	47.99	24.37	23.61	27.55	24.46
Eastern equine encephalitis	+	0.412	50.33	24.34	25.98	27.69	21.97
Hepatitis Avirus	+	0.298	37.15	21.74	15.40	30.08	32.76
Hepatitis C virus	+	0.477	57.88	28.13	29.74	20.54	21.57
Hepatitis E virus	+	0.445	57.29	26.19	31.09	17.32	25.38
Hepatitis G virus	+	0.46	58.83	32.18	26.64	17.81	23.36
Japanese encephalitis virus	+	0.418	51.48	28.42	23.05	27.46	21.05
Norwalk virus	+	0.37	48.54	23.62	24.91	26.50	24.95
O'Nyong Nyong virus	+	0.383	48.56	24.42	24.13	30.89	20.54
Poliovirus type 3	+	0.392	45.96	23.28	22.68	30.18	23.85
Rhinovirus type 89	+	0.291	38.29	19.44	18.84	32.14	29.56
Ross river virus	+	0.439	52.18	26.60	25.57	27.37	20.44
Rubella virus	+	0.612	69.59	30.87	38.71	14.90	15.50
SARS coronavirus	+	0.325	41.02	21.08	19.93	28.25	30.72
Sindbis virus	+	0.418	51.05	25.18	25.87	27.98	20.96
Venezuelan encephalitis virus	+	0.423	50.12	25.49	24.62	28.08	21.79
West Nile virus	+	0.426	51.2	28.79	22.40	27.23	21.56
Yellow fever virus	+	0.416	49.73	28.58	21.13	27.06	23.21
HIV-1	Retro	0.328	43.28	24.93	18.34	34.66	22.05
HIV-2	Retro	0.355	45.89	25.19	20.69	33.34	20.76
HTLV-1	Retro	0.41	52.68	18.16	34.51	23.06	24.25
HTLV-2	Retro	0.428	53.62	17.75	35.86	24.40	21.96
Borna virus	–	0.401	50.65	25.06	25.58	25.06	24.21
Bunyamwera virus	–	0.353	35.97	19.25	16.71	35.31	28.71
Crimean-Congo virus	–	0.369	43.59	22.44	21.14	31.55	24.85
Ebola virus	–	0.344	44.36	21.54	22.81	30.62	25.02
Hantaan virus	–	0.318	40.44	22.59	17.84	31.82	27.74
Hendra virus	–	0.343	42.38	22.97	19.40	32.53	25.08
Influenza A virus (H3N2)	–	0.356	43.57	24.31	19.25	32.43	23.98
Influenza B virus	–	0.325	41.13	22.47	18.65	35.23	23.63
Influenza C virus	–	0.301	38.6	20.58	18.02	35.78	25.60
La Crosse virus	–	0.316	37.64	20.39	17.24	34.60	27.74
Marburg virus	–	0.311	40.71	19.66	21.04	31.94	27.34
Measles virus	–	0.383	47.19	24.34	22.84	28.45	24.35
Metapneumovirus	–	0.299	39.09	20.97	18.11	36.94	23.96
Mokola virus	–	0.383	45.28	25.34	19.94	30.15	24.56
Mumps virus	–	0.32	41.98	19.24	22.73	29.78	28.23
Nipah virus	–	0.329	40.36	21.50	18.85	33.21	26.43
Parainfluenzavirus 1	–	0.296	38.53	20.10	18.42	36.39	25.06
Parainfluenzavirus 2	–	0.306	39.9	18.43	21.46	31.71	28.39
Parainfluenzavirus 3	–	0.288	36.52	18.94	17.56	37.99	25.48
Rabies virus	–	0.388	46.05	24.35	21.69	28.67	25.27
Respiratory syncytial virus	–	0.296	35.32	15.17	20.14	38.98	25.70
Rift Valley fever virus	–	0.385	45.61	25.27	20.33	28.12	26.26
Sendai virus	–	0.375	46.5	24.90	21.59	29.40	24.09
Sin Nombre virus	–	0.299	39.13	22.29	16.84	31.31	29.54
Vesicular stomatitis virus	–	0.34	41.76	22.27	19.48	31.82	26.42

A (H3N2). The names of the viruses and their codon composition are shown in Table 1 . There is a significant difference between GC contents of positive-stranded viruses versus negative-stranded viruses ($p < 0.01$, *t*-test) (Fig. 1). The positive-stranded viruses have a mean GC content of 49.8% in their coding sequences, while that of the negative-stranded viruses is 40.4%. I excluded retroviruses from positive-stranded viruses in these analyses because their replication strategies are very different. If the strategies of genome replication and expression could affect mutational pressure or exert a selection pressure on codon usage, it is likely to be different between retroviruses and other positive-stranded RNA viruses. For retroviruses, two distinct patterns were observed: HIV has lower GC content, while HTLV has high GC con-

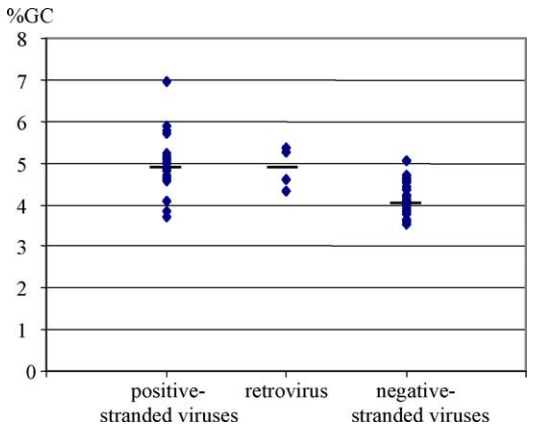

Fig. 1. A dot plot shows genome GC contents of positive-stranded RNA viruses on the left and those of retroviruses in the middle and negative-RNA viruses on the right.

tent (Fig. 1). This is in agreement with a previous observation, but the reason for this difference is unclear (Berkhout et al., 2002).

Codon usage bias of many human viruses does not match the pattern for efficient expression in human and has been shown to be driven mainly by GC contents of their genomes (Jenkins and Holmes, 2003). Expression of viral genes can be restricted by codon usage bias (Haas et al., 1996), and codon optimization can enhance expression of viral genes and has been used in development of DNA vaccines (Andre et al., 1998). To study the codon bias in relation to predicted translational efficiency in human cells, I calculated codon adaptation index (CAI) using highly expressed human genes as the reference set (Haas et al., 1996). This highly expressed codon set has been used successfully for codon optimization of viral genes for efficient expression in human cells (Andre et al., 1998; Haas et al., 1996). The CAI was designed for predicting the level of expression of a gene and for assessing the adaptation of viral genes to their hosts. A higher CAI value indicates a better codon adaptation. Genes with well-adapted codons for efficient translation generally have CAIs of > 0.6 (Sharp and Li, 1987). The CAI was calculated on a server of evolving code group at the University of Maryland (http://www.evolvingcode.net/codon/CAI_Calculator.php). In this set of RNA viruses, GC contents correlated with CAIs with a Pearson correlation coefficient of 0.959 ($p < 0.01$) (Fig. 2a). CAIs varied widely among viruses ranging from 0.288 for parainfluenza virus to 0.612 for rubella virus with an average of 0.369. This result confirmed that codon bias of RNA viruses is driven mainly by GC content, and consequently the positive-stranded viruses have higher CAI than the negative-stranded viruses (0.403 versus 0.325, $p < 0.001$, *t*-test). Since codons contain different number of GC, amino acid content can be biased by GC content. To determine the influence of GC content on amino acid choice, I counted the number of amino acids Glycine, Alanine, Arginine, and Proline (GARP), of which codons are GC rich. The GARP contents in this set of viruses show a Pearson correlation coefficient of 0.959 ($p < 0.01$) with GC content (Fig. 2b). This indicates that amino acid contents in the viral proteins are determined mainly by GC contents of their genomes.

I further analyzed coding nucleotide-count of these viruses. Most positive-stranded viruses, HIVs, and All negative-stranded viruses have high A and low C (in the positive-strands), although the positive-stranded viruses

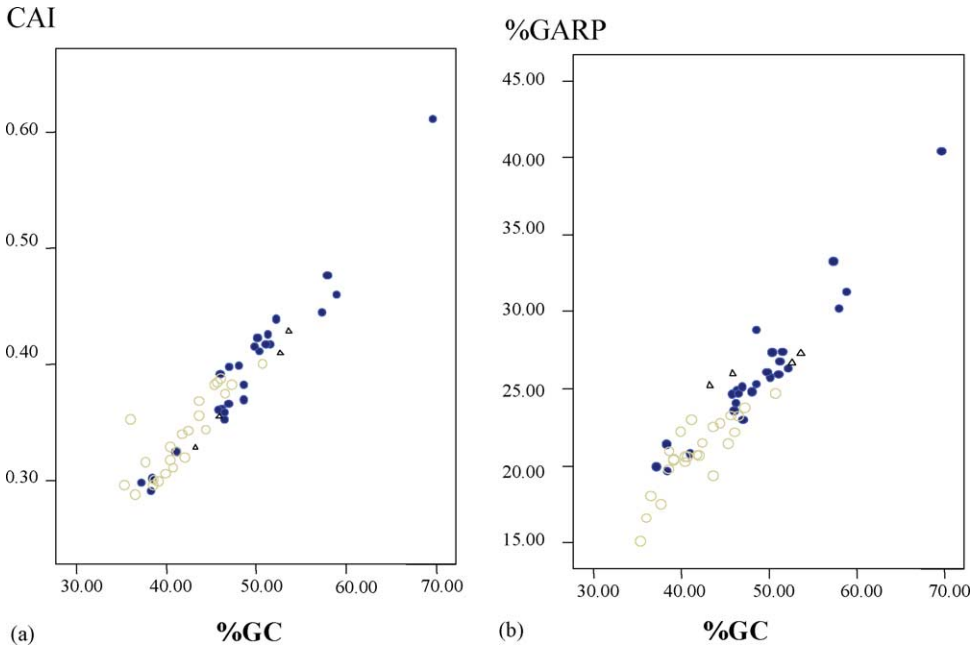

Fig. 2. Dot plots show the relationships between genome GC contents and CAIs (a), and between GC contents and %GARP (b) in all analyzed viruses. Solid circles represent positive-stranded RNA viruses, open circles represent negative-stranded viruses, and open triangles are retroviruses.

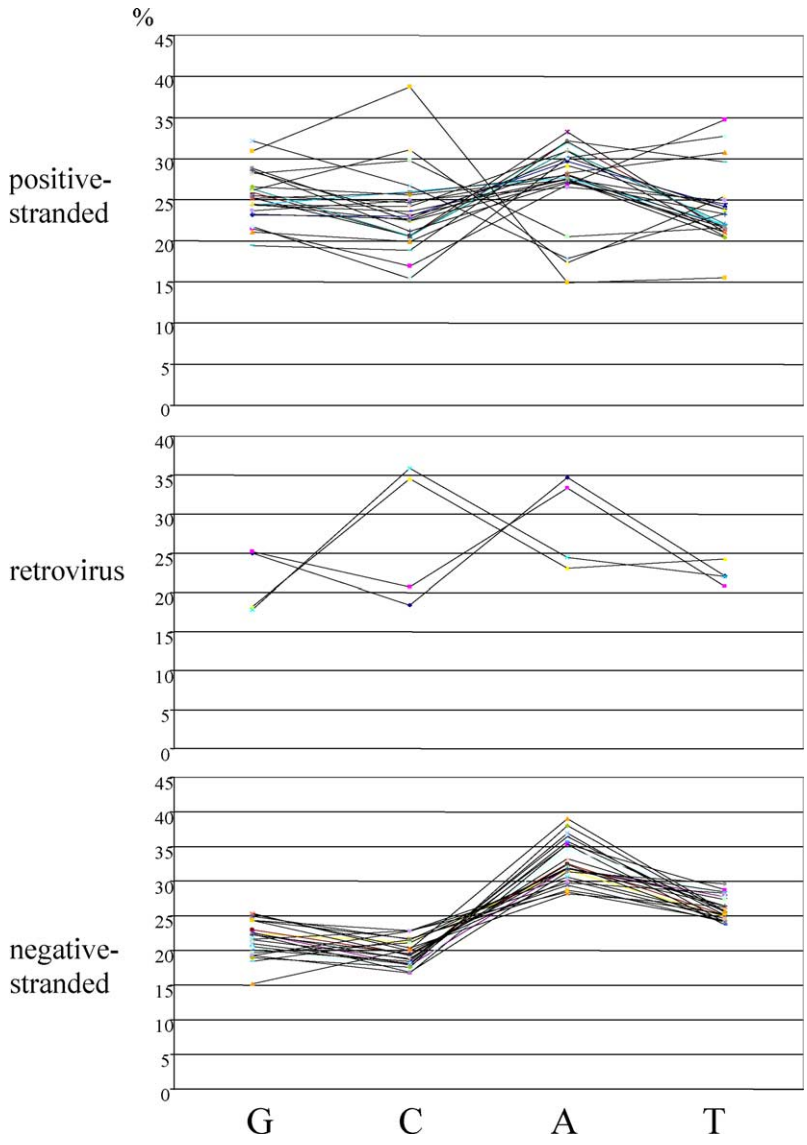

Fig. 3. Nucleotide frequencies of coding sequences in positive-stranded viruses (upper), retroviruses (middle), and negative-stranded viruses (lower).

show relatively lower level of bias. Some of the positive-stranded viruses and HTLVs, on the other hand, have low A and high C (Fig. 3). The reason for these two opposite pattern of biases is not clear. These patterns of nucleotide biase were similar when first, second, and third positions of codon were analyzed separately (data not shown). This suggested that selection pressure on codon preference is not likely to be the cause of the nucleotide bias. Because a similar pattern (high A and low C) was observed in both positive- and negative-stranded viruses on the same plus strand, i.e. genome of positive-stranded viruses and anti-genome of negative-stranded viruses, the mechanism underlying the bias may be similar and act in a strand-specific manner. Because copying of both strands uses the same viral RNA polymerase and takes place in similar intracellular environment, intrinsic copying error of the enzyme is unlikely to cause the strand-specific nucleotice bias.

Recently, a mechanism responsible for G to A hypermutation in HIV-1 by a host innate defense has been discovered (Mangeat et al., 2003; Shindo et al., 2003). Other types of RNA-editing, some of which target double-stranded RNA, have been also reported in some RNA viruses (Galinski et al., 1992; Polson et al., 1996). If host responses are also responsible for mutational bias in other RNA viruses, it is possible that they are less effective for positive-stranded RNA genomes as they might be recognized as self mRNA. It is also possible that the minus strand RNA may be the main target of host RNA-editing mechanism. This would explain the strand-specific pattern of nucleotide bias. It might also explain the nucleotide bias difference between the viruses with different genome polarity, because positive-stranded viruses produce only limited amount of minus strand anti-genome, which may be well protected in their replication complex. Negative-stranded viruses, on the other hand, must produce

numerous amount of minus strand RNA. While the explanation awaits further exploration, my analysis gives an initial clue to an interaction between strategies of genome replication (genome polarity) and mutational bias (GC content) of RNA viruses.

Acknowledgements

I would like to acknowledge supports from the Ellison Foundation and the Thailand Research Fund.

References

Andre, S., Seed, B., Eberle, J., Schraut, W., Bultmann, A., Haas, J., 1998. Increased immune response elicited by DNA vaccination with a synthetic gp120 sequence with optimized codon usage. J. Virol. 72, 1497–1503.

Aota, S., Ikemura, T., 1986. Diversity in G + C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res. 14, 6345–6355.

Berkhout, B., Grigoriev, A., Bakker, M., Lukashov, V.V., 2002. Codon and amino acid usage in retroviral genomes is consistent with virus-specific nucleotide pressure. AIDS Res. Hum. Retroviruses 18, 133–141.

Cattaneo, R., Schmid, A., Eschle, D., Baczko, K., ter Meulen, V., Billeter, M.A., 1988. Biased hypermutation and other genetic changes in defective measles viruses in human brain infections. Cell 55, 255–265.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L., McAdams, H.H., 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci. USA 101, 3480–3485.

Francino, M.P., Ochman, H., 1999. Isochores result from mutation not selection. Nature 400, 30–31.

Galinski, M.S., Troy, R.M., Banerjee, A.K., 1992. RNA editing in the phosphoprotein gene of the human parainfluenza virus type 3. Virology 186, 543–550.

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

Ikemura, T., Wada, K., 1991. Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data. Nucleic Acids Res. 19, 4333–4339.

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

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., Ikemura, T., 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J. Mol. E vol.53, 290–298.

Lobry, J.R., Sueoka, N., 2002. Asymmetric directional mutation pressures in bacteria. Genome Biol. 3, RESEARCH0058.

Mangeat, B., Turelli, P., Caron, G., Friedli, M., Perrin, L., Trono, D., 2003. Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. Nature 424, 99–103.

Polson, A.G., Bass, B.L., Casey, J.L., 1996. RNA editing of hepatitis delta virus antigenome by dsRNA-adenosine deaminase. Nature 380, 454–456.

Sharp, P.M., Li, W.H., 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295.

Shindo, K., Takaori-Kondo, A., Kobayashi, M., Abudu, A., Fukunaga, K., Uchiyama, T., 2003. The enzymatic activity of CEM15/Apobec-3G is essential for the regulation of the infectivity of HIV-1 virion but not a sole determinant of its antiviral activity. J. Biol. Chem. 278, 44412–44416.

Sueoka, N., 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA 85, 2653–2657.

Sueoka, N., 1993. Directional mutation pressure, mutator mutations, and dynamics of molecular evolution. J. Mol. E vol.37, 137–153.

Vartanian, J.P., Henry, M., Wain-Hobson, S., 2002. Sustained G → A hypermutation during reverse transcription of an entire human immunodeficiency virus type 1 strain Vau group O genome. J. Gen. Virol. 83, 801–805.

Vartanian, J.P., Meyerhans, A., Sala, M., Wain-Hobson, S., 1994. G → A hypermutation of the human immunodeficiency virus type 1 genome: evidence for dCTP pool imbalance during reverse transcription. Proc. Natl. Acad. Sci. USA 91, 3092–3096.