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.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales

Wanjun Gu, Tong Zhou, Jianmin Ma, Xiao Sun, Zuhong Lu*

Key Laboratory of Molecular and Biomolecular Electronics, Southeast University, Ministry of Education, Nanjing, Jiangsu 210096, China

Received 25 October 2003; received in revised form 9 January 2004; accepted 9 January 2004

Abstract

In this study, we calculated the codon usage bias in severe acute respiratory syndrome Coronavirus (SARSCoV) and performed a comparative analysis of synonymous codon usage patterns in SARSCoV and 10 other evolutionary related viruses in the Nidovirales. Although there is a significant variation in codon usage bias among different SARSCoV genes, codon usage bias in SARSCoV is a little slight, which is mainly determined by the base compositions on the third codon position. By comparing synonymous codon usage patterns in different viruses, we observed that synonymous codon usage pattern in these virus genes was virus specific and phylogenetically conserved, but it was not host specific. Phylogenetic analysis based on codon usage pattern suggested that SARSCoV was diverged far from all three known groups of Coronavirus. Compositional constraints could explain most of the variation of synonymous codon usage among these virus genes, while gene function is also correlated to synonymous codon usages to a certain extent. However, translational selection and gene length have no effect on the variations of synonymous codon usage in these virus genes.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Synonymous codon usage; Mutational bias; Selection pressure; Base composition; SARS; Coronavirus

1. Introduction

Synonymous codons are not used equally both within and between genomes (Grantham et al., 1980; Martin et al., 1989; Lloyd and Sharp, 1992). Compositional constraints and natural selection are thought to be the two main factors accounting for codon usage variation among genes in different organisms (Karlin and Mrazek, 1996; Sharp et al., 1986; Lesnik et al., 2000). The diverse patterns of codon usage in mammals may arise from compositional constraints of the genomes (Karlin and Mrazek, 1996; Francino and Ochman, 1999; Majumdar et al., 1999; Ghosh et al., 2000). In contrast, in some unicellular organisms, such as Escherichia coli and Saccharomyces cerevisiae, high expressed genes have a strong selective preference for codons with a high concentration of the corresponding acceptor tRNA molecule, whereas low expressed genes displayed a more uniform pattern of codon usage (Gouy and Gautier, 1982; Grantham et al., 1981; Ikemura, 1981, 1985; Sharp et al., 1986; Lesnik et al., 2000). Moreover, mutational pressure rather than translational selection is the most important determinant of the codon bias in some human RNA viruses (Levin and Whitome, 2000; Jenkins et al., 2001; Jenkins and Holmes, 2003). Furthermore, replicational and transcriptional selection is responsible for the codon usage variation among the genes of Borrelia burgdorferi (McInerney, 1998). In some other researches, codon usage was also found to be related to gene function (Chiapello et al., 1998; Epstein et al., 2000; Ma et al., 2002), protein secondary structure (Chiusano et al., 1999, 2000; Oresic and Shalloway, 1998; Xie and Ding, 1998; Gupta et al., 2000), cellular location of gene products (Chiapello et al., 1999) and gene length (Moriyama and Powell, 1998).

Severe acute respiratory syndrome (SARS) is a respiratory disease that was recently reported in Asia, North America and Europe (Chan-Yeung and Yu, 2003; Drazen,
Although genome sequence of severe acute respiratory syndrome Coronavirus (SARS-CoV) has been published and many studies have been performed on SARS-CoV in recent months (Paul et al., 2003; Qin et al., 2003; Marra et al., 2003; Snijder et al., 2003), little genomic analysis might give some clues to the features of SARS-CoV genome and some evolutionary information of this virus. Here, we analyzed the codon usage data of this virus and other coronaviruses in virus taxonomy (Marra et al., 2003). The complete genome and coding sequences (six viruses in the genus Coronavirus) were also parsed from GenBank (Version 134.0) (Table 2). Although genome sequence of SARS-CoV might give some clues to the features of SARS-CoV genome and some evolutionary information of this virus. Here, we analyzed the codon usage data of this virus and other coronaviruses in virus taxonomy (Marra et al., 2003). The complete genome and coding sequences (six viruses in the genus Coronavirus) were also parsed from GenBank (Version 134.0) (Table 2).

Table 1

| Gene product                | ENC   | GC3S (%) | L* |
|-----------------------------|-------|----------|----|
| Putative orf1ab polyprotein | 48.47 | 32.20    | -0.60 |
| Orf1a polyprotein           | 48.24 | 31.10    | -0.57 |
| Putative spike glycoprotein  | 47.66 | 34.50    | -0.37 |
| Putative uncharacterized protein | 42.80 | 45.10    | 1.34 |
| Putative small envelope protein E | 59.06 | 38.70    | 0.34 |
| Putative protein M           | 40.14 | 42.50    | 0.51 |
| Putative uncharacterized protein | 42.19 | 28.80    | -1.06 |
| Putative uncharacterized protein | 43.05 | 30.60    | -0.85 |
| Putative nucleocapsid protein | 54.16 | 37.60    | 0.49 |
| Putative uncharacterized protein | 46.62 | 58.10    | 1.87 |

* L* represents the length of identified ORF

Table 2

| Organism a | Accession number | GC3S (%) | f1′ | f2′ |
|------------|-----------------|----------|-----|-----|
| Coronavirus |                 |          |     |     |
| HCoV 229E | NC_002645       | 39.09    | -0.84 | -0.16 |
| PIVDV      | NC_001436       | 37.32    | -0.04 | 0.42 |
| TGEV       | NC_002106       | 27.02    | -0.99 | -0.08 |
| BCoV       | NC_000045       | 29.43    | -0.75 | -0.48 |
| MHV        | NC_001446       | 38.30    | -0.16 | 0.27 |
| ABV        | NC_001451       | 26.09    | -0.90 | -1.30 |
| SARS-CoV   | NC_002714       | 37.23    | 0.05  | 0.96 |
| Arterivirus |                 |          |     |     |
| EAV        | NC_002532       | 47.28    | 0.80  | 0.47 |
| LDEV       | NC_002534       | 45.18    | 0.53  | 0.43 |
| PRRSV      | NC_001961       | 53.76    | 1.31  | 0.55 |
| SHFV       | NC_003992       | 46.43    | 1.09  | -0.14 |

* Organism abbreviation: HCoV 229E, human Coronavirus 229E; PIVDV, porcine epidemic diarrhea virus; TGEV, transmissible gastroenteritis virus; BCoV, bovine Coronavirus; MHV, murine hepatitis virus; ABV, avian infectious bronchitis virus; SARS-CoV, SARS Coronavirus; EAV, equine arteritis virus; LDEV, lactate dehydrogenase elevating virus; PRRSV, porcine respiratory and reproductive syndrome virus; SHFV, simian hemorrhagic fever virus.

2. Materials and methods

2.1. Materials

SARS-CoV is a large, enveloped, positive-stranded RNA virus, which belongs to family Coronaviridae, genus Coronavirus in virus taxonomy (Marra et al., 2003). The complete genome and coding sequences of SARS-CoV TOR2 isolation were obtained from GenBank (Version 134.0). To keep the statistical significance of codon usage bias (Comeron and Aguade, 1998), ENC value which is the best overall estimator of absolute synonymous codon usage bias (Wright, 1990), the effective number of codons of a gene (ENC) was used to quantify the codon usage bias of a gene (Wright, 1990), which is the best overall estimator of absolute synonymous codon usage bias (Cameron and Aguda, 1998). ENC value ranges from 20 (when only one codon is used per amino acid) to 61 (when all synonymous codons are equally used for each amino acid).

2.2. Methods

2.2.1. Synonymous codon usage measures (RSCU)

Relative synonymous codon usage values of each codon in a gene were used to examine the synonymous codon usage without the confounding influence of amino acid composition (Sharp and Li, 1986). NGS, the frequency of base N at synonymous third codon positions, was also used to calculate the extent of base composition bias. Additionally, the effective number of codons of a gene (ENC) was used to quantify the codon usage bias of a gene (Wright, 1990), which is the best overall estimator of absolute synonymous codon usage bias (Cameron and Aguda, 1998). ENC value ranges from 20 (when only one codon is used per amino acid) to 61 (when all synonymous codons are equally used for each amino acid).

2.2.2. Correspondence analysis (CA)

Correspondence analysis was used to investigate the major trend in codon usage variation among genes. Each gene is represented as a 59 dimensional vector, and each dimension corresponds to the RSCU value of one sense codon (excluding AUG, UGG and three stop codons). The second step provides the visualization of these Euclidean distances among all genes, and the resulting distance table will be used to compute the coordinates of the genes in a multidimensional space. The second step provides the visualization of these Euclidean distances through positioning genes by successive orthogonal projections of the cloud of points. Essentially, this process consists in finding the linear transformations \( f_1′, f_2′, \ldots, f_{59}′ \) of the original variables \( f_1, f_2, \ldots, f_{59} \). The \( f′ \) variables are calculated and ordered according to the values of relative variance. \( f_1′ \) is the maximum value; \( f_2′ \) is the next value and is by construction not correlated with \( f_1′ \). The same applies to \( f_2′, f_3′ \) and so on, until \( f_{59}′ \). So, genes with similar codon usage are neighbors on the components of projection.
2.2.3. Statistical methods

Linear regression analysis was used to find the correlation between codon usage bias and nucleotide composition. One tailed t-test was used to compare the variation of codon usage between different gene groups (Ewens and Grant, 2001). As a null hypothesis, it is assumed that mean values of codon usage indices in different gene groups is statistically the same. Under the null assumption, t-statistic could be calculated. Then, P-value is derived and it is taken as significance when P-value is below 0.05.

A C++ program was developed to calculate the codon usage indices for each gene. CA and other statistical analysis were performed with statistical software SPSS 11.0.

3. Results

3.1. Synonymous codon usage in SARSCoV

The details of coding genes in SARSCoV and the overall RSCU values of 61 sense codons in SARSCoV were, respectively, shown in Tables 1 and 3. All preferentially used codons in SARSCoV are all A-ended or U-ended codons (Table 3). SARSCoV is a GC poor genome with GC content of 37.52%. Due to compositional constraints, it is expected that A-ended and/or U-ended codons should be preferentially used in this genome. To study the codon usage bias in SARSCoV genome is a little slight. However, there is a marked variation in codon usage pattern among different SARSCoV genes (S.D. = 6.41). Because all ENC values of SARSCoV genes are much higher (ENC > 40), codon usage bias in SARSCoV genome is a little slight. However, there is a marked variation in codon usage pattern among different SARSCoV genes, which range from 28.3 to 58.1% with a mean of 37.23 and S.D. of 8.78%.

3.2. Synonymous codon usage in different viruses is virus specific, but not host specific

CA was implemented for all identified ORFs from each of the 11 virus genomes as a single dataset, which consists of 103 coding sequences. CA detected one major trend in the first axis which accounted for 15.40% of the total variation, and none of the other axes individually accounted for more than 7.60% of the total variation. A plot of the first axis and the second axis of each gene was shown in Fig. 1. Although this graph is a little complex with some overlap among genes from different genomes, it is clear that genes from a particular genome tend to cluster together. The separation of one virus genome from other virus genomes is determined to be significant on both axes (t-test, P-value < 10^-15 on the first axis and P-value < 10^-3 on the second axis). So, similar to codon usage in mammals and bacteria, synonymous codon usage in these viruses is also virus specific.

To show whether there is a correlation between virus codon usage and its host, these 103 virus genes were divided into several groups according to the virus host. For example, because both SARSCoV genes and human Coronavirus 229E infect human, genes in these two viruses were incorporated as a group. Next, t-test was also used to test whether the separation of different viral genes which infect different hosts is significant. The P-value is 0.57 on the first axis and is 0.08 on the second axis, which suggested that codon usage in different virus genes was not host specific.

3.3. Phylogenetic analysis of these viruses based on codon usage pattern

In Fig. 1, all virus genes in the genus Coronavirus were plotted in red. At the same time, all viral genes in the genus Arterivirus were plotted in blue. Coronavirus genes are mainly located on the left side of the plot, while a majority
of Arterivirus genes are located on the right side. The separation of Coronavirus genes and Arterivirus genes on the first axis is statistically significant (t-test, P-value <10^{-15}). Hence, synonymous codon usage appears to be conserved between phylogenetically related viruses.

Also, SARS-CoV genes were widely extended in the first axis (Fig. 1). Six of eleven SARS-CoV genes were located in the cluster of Coronavirus genes, while the other five SARS-CoV genes were located in the cluster of Arterivirus genes. Therefore, SARS-CoV might have been diverged far from all three known Coronavirus groups. Comparing with all other viruses in the genus Coronavirus, it might be more evolutionary related to the genus Arterivirus.

3.4. Mutational bias is the main factor determines the codon usage variation among different viruses

Linear regression analysis was implemented to find whether there is some correlation between synonymous codon usage bias and nucleotide compositions. The $R^2$ value and significance level of these regression analyses was listed in Table 4. The first axis value of each gene in CA is closely correlated with all the base compositions on the third codon position, while the second axis of each gene is correlated with some base compositions on the third codon position to a certain extent. Therefore, compositional constraint mainly determines the variation of synonymous codon usage among these virus genes.

Furthermore, we plotted the first axis values in CA and GC3S values of each gene (Fig. 2). The GC3S mean value of genes in coronaviruses ranges from 26.09 to 37.32, and it ranges from 45.18 to 53.76 in arteriviruses (Table 2). Although codon usage bias appears to be conserved between evolutionary related viruses (Section 3.3), the patterns of codon usage in different virus genes also appear to be a direct function of the GC content on the third codon position of these genes.

3.5. Gene function also drives the codon usage variation among different viruses

The plot of ENC and GC3S is another effective way to explore codon usage variation among genes (Wright, 1990). ENC values of each virus gene were plotted against its

| Base composition | $f_1^*$ | $f_2^*$ |
|------------------|--------|--------|
| A3S              | 0.791  | 0.085  |
| T3S              | 0.239  | 0.444  |
| G3S              | 0.484  | 0.000  |
| C3S              | 0.720  | 0.000  |
| GC3S             | 0.936  | 0.018  |

NS in superscript represent non-significant.

$^*$ Value in this table is the $R^2$ value of each linear regression analysis.

$^1$ $f_1$ and $f_2$, respectively, represent the values of the first and the second axis of each gene in CA.

$^{	ext{NS}}$ P-value <0.0001.

Table 4
Summary of linear regression analysis between the first two axes in CA and the nucleotide contents on the third codon position in all selected virus genes*
corresponding GC3S (Fig. 3). The solid line represents the curve if codon usage is only determined by GC content on the third codon position. A large proportion of points lie near to the solid line on the left region of this distribution. It also suggests that mutational bias is the main factor determines the codon usage variation among these genes. However, there are also some points lying below the expected curve. Hence, other than mutational bias, there might be some additional factors drive the codon usage variation among these genes.

To show whether translational selection or gene function were correlated with the observed variation in codon bias, all virus genes were grouped into several classes according to gene function. Because most of these viruses contain genes coding for RNA polymerase, envelop protein and structural glycoprotein, these three gene groups were selected to find whether there is some correlation between codon usage and gene function. One tailed $t$-test was then performed on ENC values of these genes with the hypothesis that there is no correlation between codon usage bias and gene function. Some associations have been found. Average codon usage bias is higher in RNA polymerase gene group than in envelop gene group ($t$-test, $P$-value = 0.031), and it is higher in polymerase gene group than in structural glycoprotein gene group ($t$-test, $P$-value = 0.002). But, there is no association between codon usage in structural glycoprotein gene group and envelop protein gene group ($t$-test, $P$-value = 0.74). Because the structural glycoprotein and envelop protein are all structural proteins in these viruses and RNA polymerase is a nonstructural protein, it is clear that codon usage in structural genes is significantly diverged from that in nonstructural genes. On the other hand, structural genes are generally highly expressed than nonstructural genes. So, if translational selection was also contributed to codon usage bias in these genes, codon usage bias in structural genes should be higher than in RNA polymerase genes. However, RNA polymerase genes (ENC = 49.25) were found to have greater codon usage bias than structural genes (ENC = 54.60 for envelop gene and ENC = 55.33 for structural glycoprotein). Hence, codon usage bias in these virus genes is not related to gene expression level. Furthermore, we also performed a linear regression analysis on ENC value and gene length of each gene. But, there was no significant correlation between codon usage and gene length in these virus genes ($P$-value > 0.05). So, gene function, rather than translational selection
2000; Jenkins et al., 2001; Jenkins and Holmes, 2003). A little variation in bias between genes (Levin and Whittome, 1999), it is understandable that mutation pressure and natural selection. Further, such information might be helpful to understand the pathogenesis and the origin of SARS-CoV.

4. Discussion

Our analysis revealed that synonymous codon usage bias in SARS-CoV was less biased, which was mainly determined by the base compositions on the third codon position. Comparative analysis of codon usage bias in the order Nidovirales also suggested that codon usage in these viruses was virus specific and mutational bias was the main factor drives the codon usage variation among these viruses. Gene function was also related to codon usage bias in these viruses to some extent. But, translational selection and gene length might have no effect on the codon usage pattern in these viruses. Some published results has shown that the overall extent of codon usage bias in RNA viruses is low and there is little variation in bias between genes (Levin and Whitome, 2000; Jenkins et al., 2001; Jenkins and Holmes, 2003). Although SARS-CoV is a newly detected RNA virus infecting human, the synonymous codon usage pattern in SARS-CoV we described here is also in accordance with these published codon usage pattern of human RNA viruses (Jenkins and Holmes, 2003). Because mutation rates in RNA viruses are much higher than those in DNA viruses (Drake and Holland, 1999), it is understandable that mutation pressure is the main determinant of codon usage bias in SARS-CoV. Our analysis also revealed that there was no host specific codon usage pattern in these viruses. So, host genome might have no obvious effect on the evolution of these viruses.

Some phylogenetic analysis of SARS-CoV (Qin et al., 2003; Marra et al., 2003) has shown that SARS-CoV does not closely resemble any of the three previously known groups in genus Coronavirus. But Snijder et al. (2003) has proposed that SARS-CoV is most closely related to group 2 Coronavirus. Based on different codon usage patterns in different coronaviruses, we revealed that codon usage patterns of each virus was phylogenetically distinct and SARS-CoV might have been diverged far from all three known Coronavirus groups, which is in accordance with the results Qin et al. (2003) and Marra et al. (2003) proposed.

Codon usage patterns and the phylogenetic results we proposed here are useful to understand the processes governing the evolution of SARS-CoV, especially the roles played by mutation pressure and natural selection. Further, such information might be helpful to understand the pathogenesis and the origin of SARS-CoV.

Acknowledgements

This research is a part of Projects 60121101 and 60223002 supported by National Natural Science Foundation of China.

References

Chan-Yen, M., Yu, W.C., 2003. Outbreak of severe acute respiratory syndrome in Hong Kong Special Administrative Region: case report. Brit. Med. J. 326, 850–852.

Chapelle, H., Lisacek, F., Caboche, M., Henaut, A., 1998. Codon usage and gene function are related in sequences of Arabidopsis thaliana. Gene 209, GC1–GC8.

Chapelle, H., Olivier, E., Lander-Devauchelle, C., Nitschke, P., Riether, J.J., 1999. Codon usage as a tool to predict the cellular location of eukaryotic ribosomal proteins and aminoacyl-tRNA synthetases. Nucl. Acids Res. 27, 2848–2851.

Chinazu, M., D’Onofrio, G., Alvarez-Valin, F., Jabbari, K., Colomba, G., Bernardi, G., 1999. Correlations of nucleic substitution rates and base composition of mammalian coding sequences with protein structure. Gene 236, 23–31.

Chinazu, M., Alvarez-Valin, F., Di Giuli, M., D’Onofrio, G., Ammirato, G., Colomba, G., Bernardi, G., 2000. Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. Gene 261, 63–69.

Coggan, A., Wolfe, K.H., 2000. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. Yeast 16, 1131–1145.

Comeran, J.M., Aquade, M., 1998. An evaluation of measures of synonymous codon usage bias. J. Mol. Evol. 47, 268–274.

Drake, J.W., Holland, J.J., 1999. Mutation rates among RNA viruses. Proc. Natl. Acad. Sci. U.S.A. 96, 13910–13913.

Duzan, J.M., 2003. Case clusters of the severe acute respiratory syndrome. New Engl. J. Med. 348, r6–r7.

Epstein, B.J., Lin, K., Tan, T.W., 2000. A functional significance for codon third bases. Gene 245, 291–298.

Ewens, W.J., Grant, G.R., 2001. Statistical Methods in Bioinformatics. Springer, New York.

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

Ghosh, T.C., Gupta, S.K., Mujumdar, S., 2000. Studies on codon usage in Entamoeba histolytica. Int. J. Parasitol. 30, 715–722.

Gooi, M., Gautier, C., 1982. Codon usage in bacteria: correlation with gene expressivity. Nucl. Acids Res. 10, 7055–7074.

Grahnau, R., Gantier, C., Gris, M., Mercier, R., Parse, A., 1980. Codon catalog usage and the genome hypothesis. Nucl. Acids Res. 8, e69–e62.

Grahnau, R., Gantier, C., Gooi, M., Jacobme, M., Mercier, R., 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucl. Acids Res. 9, e63–e74.

Gupta, S.K., Majumdar, S., Bhattacharya, T.K., Ghosh, T.C., 2000. Studies on the relationships between the synonymous codon usage and proteins secondary structural units. Biochem. Biophys. Res. Commun. 269, 692–696.

Ikenura, T., 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes; a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J. Mol. Biol. 151, 389–409.

Ikenura, T., 1985. Codon usage and RNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2: 13–34.

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

Kafur, S., Mrozuk, J., 1996. What drives codon choices in human genes? J. Mol. Biol. 262, 459–472.

Lesnik, T., Sotomovic, J., Deana, A., Ehrlich, R., Reiss, C., 2000. Ribosome traffic in E. coli and regulation of gene expression. J. Theor. Biol. 202, 175–185.
Levin, D.B., Whittome, B., 2000. Codon usage in nucleopolyhedroviruses. J. Gen. Virol. 81, 2133–2135.

Lloyd, A.T., Sharp, P.M., 1992. Evolution of codon usage patterns: the extent and nature of divergence between Candida albicans and Saccharomyces cerevisiae. Nucleic Acids Res. 20, 5289–5295.

Ma, J.M., Zhou, T., Gu, W.J., Sun, X., Liu, Z.H., 2002. Cluster analysis of the codon use frequency of MHC genes from different species. BioSystems 63, 199–207.

Majumdar, S., Gupta, S.K., Sundararaj, V.S., Gosh, T.C., 1999. Compositional correlation studies among the three different codon positions in 12 bacterial genomes. Biochem. Biophys. Res. Commun. 266, 66–71.

Mutas, G., Duret, L., 2001. Synonymous codon usage, accuracy of translation, and gene length in Caenorhabditis elegans. J. Mol. Evol. 52, 275–280.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate analysis. Academic press, New York.

Marra, M.A., Jones, S.J., Astell, C.R., 2003. The genome sequence of the SARS-associated Coronavirus. Science 300, 1399–1404.

Martin, A., Bertranpetit, J., Oliver, J.L., Melina, J.R., 1989. Variation in G + C content and codon choice: differences among synonymous codon groups in vertebrate genes. Nucleic Acids Res. 17, 6181–6189.

McInerney, J.O., 1998. Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proc. Natl. Acad. Sci. U.S.A. 95, 10698–10703.

Moriyama, E.N., Powell, J.R., 1998. Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. Nucleic Acids Res. 26, 3188–3193.

Oresic, M., Shalloway, D., 1998. Specific correlations between relative synonymous codon usage and protein secondary structure. J. Mol. Biol. 281, 31–48.

Paul, A.R., Steven, O.M., Stephan, S.M., 2003. Characterization of a novel Coronavirus associated with severe acute respiratory syndrome. Science 300, 1394–1399.

Qin, E.D., Zhu, Q.Y., Yu, M., 2003. A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01). Chin. Sci. Bull. 48, 941–948.

Sharp, P.M., Li, W.H., 1986. Codon usage in regulatory genes in Escherichia coli does not reflect selection for ‘rare’ codons. Nucleic Acids Res. 14, 7737–7740.

Sharp, P.M., Tushby, T.M., Mosourski, K.R., 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14, 5125–5143.

Snijder, E.J., Bredenbeek, P.J., Dobbe, J.C., 2003. Unique and conserved features of genome and proteome of SARS-Coronavirus, an early split-off from the Coronavirus group 2 lineage. J. Mol. Biol. 331, 991–1004.

Wright, F., 1990. The ‘effective number of codons’ used in a gene. Gene 87, 23–29.

Xie, T., Ding, D.F., 1998. The relationship between synonymous codon usage and protein structure. FEBS Lett. 438, 93–96.