Volume 266 Issue 1 7 September 2010 ISSN 0022-5193

ELSEVIER

Journal of Theoretical Biology

Journal of Theoretical Biology 266 (2010) 88–98

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

The level of cytosine is usually much higher than the level of guanine in two-fold degenerated sites from third codon positions of genes from Simplex- and Varicelloviruses with G+C higher than 50%

Vladislav Victorovich Khrustalev*, Eugene Victorovich Barkovsky

Department of General Chemistry, Belarussian State Medical University, Communisticheskaya 7-24, Dzerzinskogo 83, Minsk 220029, Belarus

ARTICLE INFO

Article history:
Received 4 February 2010
Received in revised form
5 May 2010
Accepted 15 June 2010
Available online 19 June 2010

Keywords:
Mutational pressure
Amino acid replacements
8-oxo-G
Alphaherpesviruses
Two-fold degenerated sites

ABSTRACT

We studied usage of cytosine and guanine in 914 genes from completely sequenced genomes of five Simplex- and seven Varicelloviruses. In genes with total GC-content higher than 50% usage of cytosine is usually higher than usage of guanine (an average difference for genes with G+C higher than 70% reaches 4.0%). This difference is caused mostly by the elevated usage of cytosine in two-fold degenerated sites situated in third codon positions relatively to the usage of guanine in two-fold degenerated sites situated in third codon positions (an average difference for genes with G+C higher than 70% is equal to 28.2%). The usage of amino acids that are encoded by codons containing cytosine in two-fold degenerated sites situated in third codon positions (AA2TC) is much higher than the usage of amino acids encoded by codons containing guanine in two-fold degenerated sites situated in third codon positions (AA2AG). The usage of AA2AG declines much more steeply with the growth of GC-content than the usage of AA2TC. This effect is the consequence of the nature of genetic code and of the negative selection. In GC-rich genes the usage of cytosine in four-fold degenerated sites is only a little (but significantly) higher than the usage of guanine (in genes with G+C higher than 70% an average difference is equal to 4.3%). This difference may be caused by transcription-associated mutational pressure.

1. Introduction

Although second Chargaff's rule (G=C and A=T) strictly obeys in doublestranded DNA, there are usually significant deviations from this rule in a single strand of DNA (Lobry and Sueoka, 2002). It was shown that replication-associated mutational pressure causes deviations from second Chargaff's rule in large viral genomes, including the genome of Herpes simplex virus type 1 (Grigoriev, 1999). The difference between guanine and cytosine levels in genes from leading strand of viral DNA is usually not the same as the difference between guanine and cytosine levels in genes from lagging strand (Grigoriev, 1999). Since there are from two to three origins of replication in genomes of Simplex- or Varicelloviruses (Strang and Stow, 2005), nucleotide content of a coding region should depend on its location in the genome.

Certain types of nucleotide mutations occur in singlestranded DNA more frequently than in doublestranded DNA (Geacintov and Shafirovich, 2003). On the other hand, certain glycosylases involved in repair also show a preference to singlestranded DNA (Bellamy and Baldwin, 2001; Dou et al., 2003). During transcription, a region of DNA identical to the mRNA (a coding region) exists in a singlestranded form, while the region of DNA complementary to mRNA serves as a matrix for the synthesis of this mRNA. During the period of transcription a coding region of DNA may accumulate certain types of nucleotide mutations. The resulting effect of all the processes introducing mutations into the coding regions of DNA during transcription is called "transcription-associated mutational pressure" (Chen and Chen, 2007).

There is also a group of processes responsible for transcription-coupled repair (Hanawalt and Spivak, 2008). This is a subpathway of nucleotide excision repair that removes lesions from the template DNA strands (from the strands which are complementary to coding regions) of actively transcribed genes (Hanawalt and Spivak, 2008).

Expression of genes from Herpes simplex virus type 1 (HSV1) has been extensively studied (Honess and Roizman, 1974; Stingley et al., 2000). According to the time of their expression during lytic infection genes of HSV1 can be roughly divided into three groups: immediate-early genes, early genes and late genes. Immediate-early genes are coding for regulatory proteins. The division of HSV1 genes into early and late ones is not always equal to the well-known classification of viral proteins into structural

* Corresponding author. Tel.: +80172845957.
E-mail address: vvkhrustalev@mail.ru (V.V. Khrustalev).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.06.023

and nonstructural. For example, glycoproteins L, B, G, I and E are encoded by early genes, while glycoproteins M, H, N, J and D are encoded by late genes (Honess and Roizman, 1974; Stingley et al., 2000); catalytic subunit of DNA-polymerase that complexes with polymerase accessory protein is encoded by early gene, while polymerase accessory protein is encoded by late gene (see this web page for further details: http://darwin.bio.uci.edu/~faculty/wagner/table.html). Time of expression and the level of expression of a gene should affect its nucleotide content.

Replication-associated mutational pressure and transcription-associated mutational pressure together with transcription-coupled repair lead to the unequal distribution of nucleotide usage along the genome: the bias in genes from leading strands is different from the bias in lagging strands (Grigoriev, 1999; Lobry and Sueoka, 2002); the bias in coding regions of DNA is different from the bias in their complementary regions (Chen and Chen, 2007). However, mutational pressure usually has a general (symmetric) direction for the whole viral genome (Khrustalev and Barkovsky, 2009b).

It has already been stated that the level of purine usage (A+G) shows an inverse correlation with GC-content of viral genes (Cristillo et al., 2001; Forsdyke, 2006). Interesting consequences of "purine-loading" and "pyrimidine-loading" on the formation of secondary structure of viral mRNAs and on virus–cell interactions have been suggested (Cristillo et al., 2001). However, a proper analysis of the cause of this phenomenon has not been performed yet. The aim of this work was to find out why the level of cytosine is higher than the level of guanine in coding regions from GC-rich Alphaherpesviruses. To answer this question we compared cytosine and guanine usage in three codon positions and in four- and two-fold degenerated sites of 914 genes from completely sequenced genomes of Simplex- and Varicelloviruses.

Finally we came to the conclusion that the level of the "asymmetry" of nucleotide distribution highly depends on the general direction of the mutational pressure. Negative selection allows fixation of certain types of amino acid substitutions more frequently than fixation of other types of amino acid substitutions. So, relatively symmetric mutational pressure controlled by "asymmetric" negative selection may produce differences in nucleotide usage inside coding regions. As we have shown in this work, cytosine usage in two-fold degenerated sites is much higher than guanine usage mostly because of "asymmetric" negative selection. Asymmetry in negative selection of amino acid substitutions caused by GC-pressure (as well as by AT-pressure) is associated with features of universal genetic code.

2. Materials and methods

In this study we analyzed all open reading frames from completely sequenced genomes of five Simplexviruses and seven Varicelloviruses. The total number of studied genes is equal to 914. One genome of Simplex- or Varicellovirus contains about 75 genes, the most of which have homologues in all other Simplex- or Varicelloviruses (Khrustalev and Barkovsky, 2009a). Names and GenBank accession numbers for complete genome records are listed below.

Simplexviruses: Macacine herpesvirus 1 (MaHV1) [NC_004812], Cercopithecine herpesvirus 2 (CeHV2) [NC_006560], Papiine herpesvirus 2 (PaHV2) [NC_007653], Human herpesvirus 1 (HSV1) [NC_001806] and Human herpesvirus 2 (HSV2) [NC_001798].

Varicelloviruses: Human herpesvirus 3 (VZV) [NC_001348], Bovine herpesvirus 5 (BoHV5) [NC_005261], Equid herpesvirus 1 (EqHV1) [NC_001491], Equid herpesvirus 4 (EqHV4) [NC_001844], Equid herpesvirus 9 (EqHV9) [AP010838], Cercopithecine herpesvirus 9 (CeHV9) [NC_002686] and Felid herpesvirus 1 (FeHV1) [NC_013590].

With the help of our "VVK in group" algorithm we calculated all the indexes representing nucleotide usage for each gene. These indexes are:

- G+C (total GC-content);
- G, 1G, 2G and 3G (total level of guanine and level of guanine in first, second and third codon positions, respectively);
- C, 1C, 2C and 3C (total level of cytosine and level of cytosine in first, second and third codon positions, respectively);
- "(1G+2G)/2" (average level of guanine in first and second codon positions);
- "(1C+2C)/2" (average level of cytosine in first and second codon positions);
- G4f and G2f3p (level of guanine in four-fold degenerated sites and level of guanine in two-fold degenerated sites situated in third codon positions, respectively);
- C4f and C2f3p (level of cytosine in four-fold degenerated sites and level of cytosine in two-fold degenerated sites situated in third codon positions, respectively);
- AA3AG (the usage of amino acids which are encoded by codons containing adenine or guanine in two-fold degenerated sites situated in third codon positions); and
- AA3TC (the usage of amino acids which are encoded by codons containing thymine or cytosine in two-fold degenerated sites situated in third codon positions).

We also calculated amino acid usage in proteins coded by studied genes.

All these indexes are calculated by "VVK in group" MS Excel spreadsheet right after the pasting of nucleotide sequences into the designated cells on its "sequences" list. The maximal number of nucleotide sequences for a single round of calculations is fifty. The algorithm is available via our web page www.barkovsky.hotmail.ru.

Slopes of the dependence between each amino acid usage and GC-content and coefficients of correlation (R) have been calculated by MS Excel.

Paired differences test has been applied to compare indexes describing guanine usage with indexes describing cytosine usage. The data obtained for all 914 genes have been sorted with the respect of their G+C level. We made five groups of genes: genes with G+C lower than 40%; genes with G+C from 40% to 50%; genes with G+C from 50% to 60%; genes with G+C from 60% to 70%; and genes with G+C higher than 70%. First, the difference between each two indexes has been calculated for every gene. Then an average difference for all genes from a given group has been calculated. Statistical significance of an average difference for each group of genes has been checked by t-test.

Additionally, we calculated frequencies of CpG and GpC dinucleotides for each coding genome. Frequencies of CpG and GpC dinucleotides situated in first and second (CpG 1-2 and GpC 1-2), in second and third (CpG 2-3 and GpC 2-3) and in third and first codon positions (CpG 3-1 and GpC 3-1) have been calculated separately and then compared to each other in paired differences test.

3. Results

3.1. Comparison between guanine and cytosine usage in three codon positions

As one can see in Fig. 1a, the growth of cytosine makes a greater contribution into the increase of total GC-content of viral genes than the growth of guanine. Indeed, the level of cytosine is significantly higher than the level of guanine in genes with G+C from 50% to 60%, although an average difference between them is rather small (1.2%). In genes with G+C from 60% to 70% this significant difference between C and G is higher (3.4%). The highest level of the difference between C and G (4.0%) is characteristic to the genes with GC-content higher than 70%. *The higher is the total GC-content of studied genes, the higher is the difference between C and G.*

To find out the cause of the excess of cytosine in coding regions with high GC-content we decided to test whether this excess is due to the bias in cytosine and guanine usage in third codon positions (3C and 3G, respectively). Fig. 1b shows that the main cause of the difference between C and G in genes with high GC-content is really in the unequal usage of 3C and 3G. Actually, an average difference between 3C and 3G is equal to 4.2% in genes with G+C from 50% to 60%. In genes with G+C from 60% to 70% this significant difference is equal to 7.7%, while in genes with G+C higher than 70% this average difference reaches the level of 9.4%.

There is a significant difference between 1G and 1C for all the genes studied. The level of 1G is always higher than the level of 1C. However, the level of 2G is always lower than the level of 2C. This phenomenon has also been observed in archaeal (Khrustalev and Barkovsky, 2007) and bacterial genes (Hu et al., 2007). In our opinion, this characteristic pattern of guanine and cytosine distribution between first and second codon positions is the consequence of common predecessor's effect (Khrustalev and Barkovsky, 2007). Levels of guanine and cytosine usage in first and second codon positions show approximately constant and equal slopes of the dependence on G+C. However, it can be clearly seen in Fig. 1c that 1G is growing more steeply than 1C in genes with G+C higher than 75%. On the other hand, 2C is growing more steeply than 2G in genes with G+C higher than 75% (see Fig. 1d). This effect is caused mostly by the steep increase of alanine usage under the influence of GC-pressure (Khrustalev and Barkovsky, 2009a). This amino acid is encoded by GCX codons.

As one can see in Fig. 2a, average levels of guanine and cytosine usage in first and second codon positions are very close to each other. In genes with G+C from 40% to 60% there is no significant difference between the levels of "(1G+2G)/2" and "(1C+2C)/2". In genes with G+C from 60% to 70% the level of "(1C+2C)/2" is just a little higher than the level of "(1G+2G)/2" (average difference is equal to 1.3%). In genes with G+C higher than 70% this minimal difference (1.4%) is also significant. However, in general, the level of guanine in first and second codon positions is practically the same as the level of cytosine (see Fig. 2a).

3.2. *Comparison between guanine and cytosine usage in four-fold and two-fold degenerated sites situated in third codon positions*

Observed excess of 3C may be due to elevated usage of C in four-fold degenerated sites or due to elevated usage of C in two-fold degenerated sites situated in third codon positions.

Four-fold degenerated sites are situated in third codon positions of 32 from 64 codons. All the nucleotide substitutions in four-fold degenerated sites are synonymous (they do not cause amino acid replacement in the encoded protein). Four-fold degenerated sites should accumulate only neutral nucleotide substitutions, so the existence of a bias between cytosine and

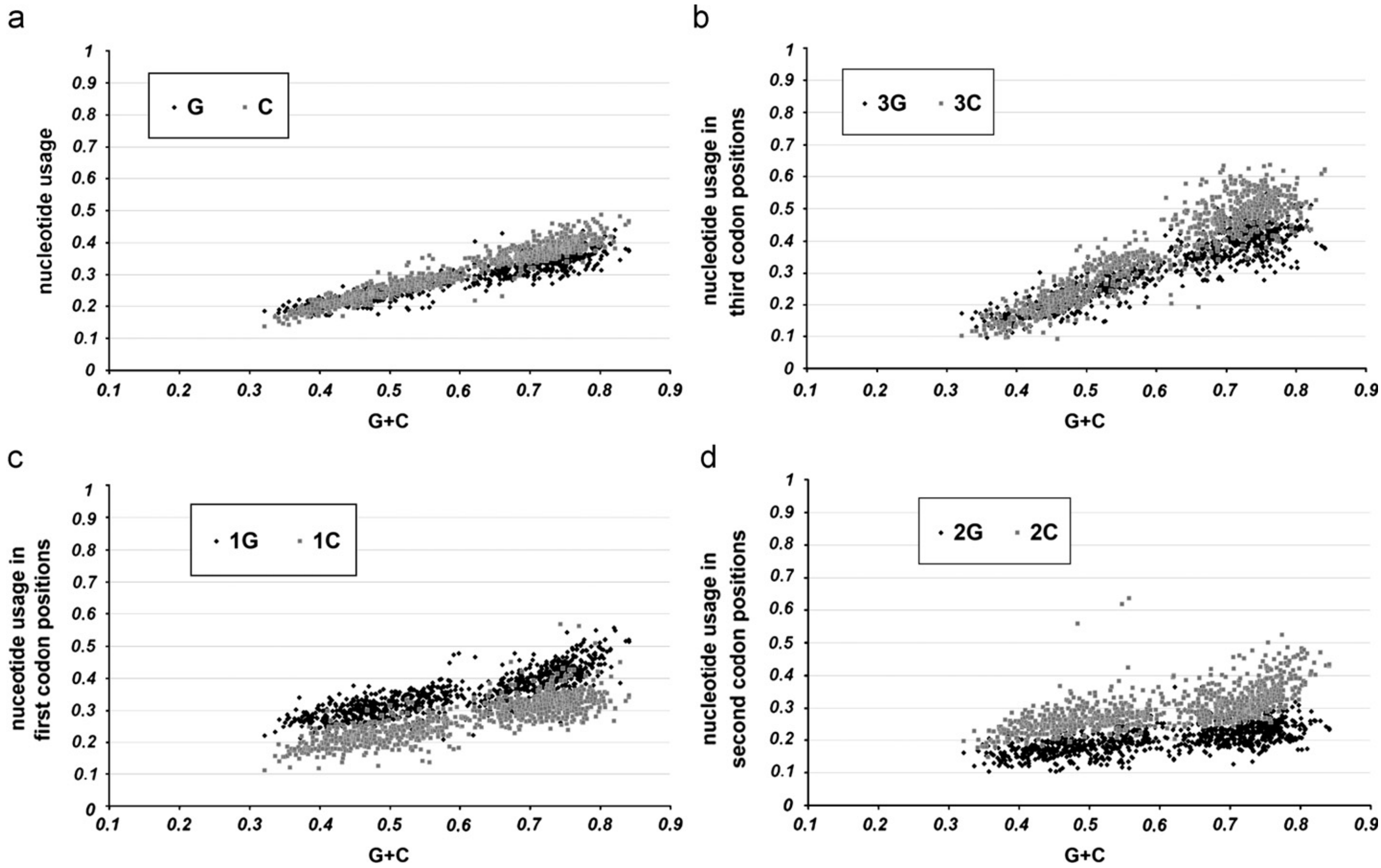

Fig. 1. Dependence between G+C and (a) total content of guanine (G) and cytosine (C); (b) level of guanine and cytosine in third codon positions (3G and 3C, respectively); (c) level of guanine and cytosine in first codon positions (1G and 1C, respectively) and (d) level of guanine and cytosine in second codon positions (2G and 2C, respectively) in coding regions of Simplex- and Varicelloviruses.

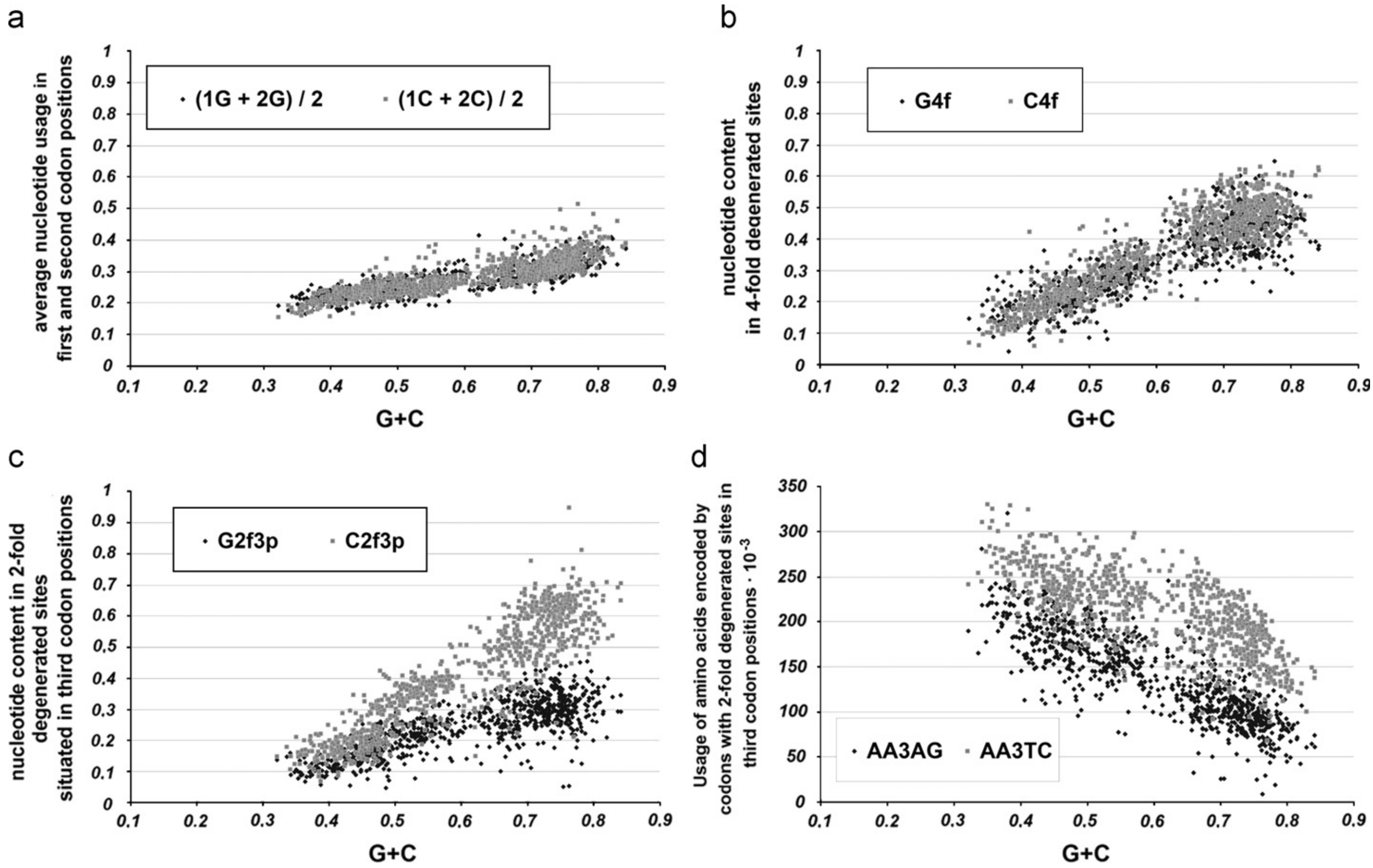

Fig. 2. Dependence between G+C and (a) average level of guanine and cytosine in first and second codon positions ("(1G+2G)/2" and "(1C+2C)/2", respectively); (b) level of guanine and cytosine in 4-fold degenerated sites (G4f and C4f, respectively); (c) level of guanine and cytosine in two-fold degenerated sites situated in third codon positions (G2f3p and C2f3p, respectively) and (d) the usage of amino acids which are encoded by codons containing adenine or guanine in two-fold degenerated sites situated in third codon positions (AA2AG) and the usage of amino acids which are encoded by codons containing thymine or cytosine in two-fold degenerated sites situated in third codon positions (AA2TC) in coding regions of Simplex- and Varicelloviruses.

guanine usage in four-fold degenerated sites is the evidence of asymmetric mutational pressure (Lobry and Sueoka, 2002).

The level of cytosine situated in these four-fold degenerated sites (C4f) is some higher than the level of guanine (G4f) in genes with G+C from 50% to 60% (average difference is equal to 2.2%). In genes with G+C from 60% to 70% an average difference between C4f and G4f is statistically significant and equal to 4.0%. In genes with G+C higher than 70% this significant difference is equal to 4.3%.

Fig. 2b shows that there are wide variations of G4f and C4f in genes with high GC-content. Although the level of C4f is significantly higher than the level of G4f, there are some genes with G4f higher than C4f (see Fig. 2b). Once again, an average difference between C4f and G4f is not too high, especially in comparison with the difference between C2f3p and G2f3p (see Fig. 2c).

Two-fold degenerated sites are situated in third codon positions of 26 from 64 codons (including two terminal codons). There can be three possible nucleotide substitutions in each site: one transition and two transversions. One possible transition is synonymous, while two possible transversions are nonsynonymous in two-fold degenerated sites situated in third codon positions. The difference between C2f3p and G2f3p is really great and undoubtedly significant (see Fig. 2c).

An average difference between C2f3p and G2f3p for genes with G+C from 40% to 50% is equal to 6.0%; in genes with G+C from 50% to 60% it is equal to 12.7%; in genes with G+C from 60% to 70% this difference is equal to 20.0%; the highest level of the difference demonstrated in Fig. 2c (28.2%) is characteristic for genes with G+C higher than 70%. These data make us sure that *the main contribution into the difference between total cytosine and guanine usage in GC-rich genes of Simplex- and Varicelloviruses is made by the difference in their usage in two-fold degenerated sites situated in third codon positions.*

There are two types of two-fold degenerated sites situated in third codon positions. There can be either thymine or cytosine in 14 from 26 codons containing two-fold degenerated sites situated in third codon positions (AA2TC). There can be either adenine or guanine in 12 from 26 codons containing two-fold degenerated sites situated in third codon positions (AA2AG). It means that the amount of codons containing two-fold degenerated sites situated in third codon positions in which cytosine can be found is different from the amount of codons containing two-fold degenerated sites situated in third codon positions in which guanine can be found.

The excess of C2f3p can be explained by two hypotheses. According to the first hypothesis C2f3p is higher than G2f3p because of the asymmetric mutational pressure. According to the second hypothesis C2f3p is higher than G2f3p because of the negative selection on amino acid substitutions leading to the excess of AA2TC against AA2AG. The first hypothesis works if C2f3p is significantly higher than G2f3p, while AA2TC is equal to or lower than AA2AG. The second hypotheses works if the ratio between C2f3p and G2f3p is equal to or lower than the ratio between AA2TC and AA2AG. If the ratio between C2f3p and G2f3p is significantly higher than the ratio between AA2TC and AA2AG, while AA2TC is higher than AA2AG, both processes (asymmetric mutational pressure and "asymmetric" negative selection) should make contribution into the excess of C2f3p.

As one can see in Fig. 2d, the level of codons containing two-fold degenerated sites situated in third codon positions in which guanine or adenine can be found (AA2AG) is always lower than the level of codons containing two-fold degenerated sites situated in third codon positions in which cytosine or thymine can be found (AA2TC). An average difference between AA2TC and AA2AG for genes with G+C lower than 40% is equal to 56.5×10^{-3}; for genes with G+C from 40% to 50% it is equal to 60.8×10^{-3}; for genes with G+C from 50% to 60% it is equal to 65.9×10^{-3}; while for genes with G+C from 60% to 70% this difference is equal to 84.0×10^{-3} and for genes with G+C higher than 70% it is equal to 81.7×10^{-3}. Fig. 2d shows that *the level of AA2AG is decreasing much more steeply with the growth of GC-content than the level of AA2TC.*

Calculations described below show that one cannot completely decline the first hypothesis for genes with G+C lower than 60%. The ratio between C2f3p and G2f3p for genes with G+C from 40% to 50% is close to but significantly higher than the ratio between AA2TC and AA2AG (1.37 versus 1.34), as well as for genes with G+C from 50% to 60% (1.56 versus 1.40). For genes with G+C from 60% to 70% and for genes with G+C higher than 70% the difference between these ratios is not significant (1.71 versus 1.70 and 1.91 versus 1.88, respectively).

These data approve that *the main cause of the great difference between C2f3p and G2f3p is the great difference between AA2TC and AA2AG*. There are much more two-fold degenerated sites in third codon positions in which cytosine may be situated than two-fold degenerated sites in third codon positions which may contain guanine.

However, the ratio between C2f3p and G2f3p is a little higher than the ration between AA2TC and AA2AG. It means that the "weak" asymmetric mutational pressure makes its contribution into the excess of C2f3p, as well as into the excess of C4f.

3.3. *Comparison between the usage of amino acids encoded by codons with adenine or guanine in two-fold degenerated sites situated in third codon positions and the usage of amino acids encoded by codons with thymine or cytosine in two-fold degenerated sites situated in third codon positions*

To find out why the level of AA2TC declines less steeply under the influence of GC-pressure than the level of AA2AG we continued our comparisons focusing on the usage of each amino acid encoded by codons with two-fold degenerated sites in third codon positions.

Asparagine is encoded by two codons (AAT/C) containing thymine or cytosine in third positions. These codons contain no guanine or cytosine in first and second codon positions. That is why the total usage of asparagine in proteomes should decline under the influence of GC-pressure and grow under the influence of AT-pressure (Singer and Hickey, 2000; Khrustalev and Barkovsky, 2009a), as it can be seen in Fig. 3a. Lysine is also encoded by two GC-poor codons (AAA/G), but they contain adenine or guanine in third codon positions. As one can see in Fig. 3b, the usage of lysine in viral proteins is showing dependence on G+C similar to that for the usage of asparagine (see Fig. 3a). Coefficients of correlation for these dependences are higher than 0.7. This is the evidence that there is a strong inversed linear dependence between the usage of asparagine in proteins and G+C in genes, as well as between the usage of lysine and G+C.

Aspartic acid is encoded by two codons (GAT/C) with thymine or cytosine in third positions, while glutamic acid is encoded by two codons (GAA/G) with adenine or guanine in third codon positions. As one can see in Fig. 3c and d, usages of both glutamic and aspartic acid show no correlation on G+C (Khrustalev and Barkovsky, 2009a). Usage of aspartic acid in viral proteomes is close to that of the glutamic acid.

Usages of Asn, Asp, Lys and Glu do not cause the difference in AA2TC and AA2AG (see Fig. 3).

Although phenylalanine is encoded by GC-poor codons (TTT/C), the correlation between its usage and G+C is weak (see Fig. 4a). The correlation between the usage of GC-poor duplet of leucine (TTA/G) and G+C is strong (see Fig. 4b). There is also a quartet of codons coding for leucine (CTX) in universal genetic code. Transitions of T to C direction in first positions of the duplet of leucine are synonymous. So, the level of Leu2 usage decreases under the influence of GC-pressure mostly due to synonymous Leu2 to Leu4 mutations. Mutations of phenylalanine caused by GC-pressure are not fixed as frequently as Leu2 to Leu4 mutations.

Tyrosine is encoded by two GC-poor codons (TAT/C) and its usage shows only a weak correlation on G+C (see Fig. 4c), just like the usage of phenylalanine does. The coefficient of correlation between the usage of the duplet of arginine (Arg2) and G+C is higher than 0.5. This duplet (AGA/G) can be substituted by quartet (Arg4: CGX) of codons coding for arginine by the way of A to C transversion. So, the usage of Arg2 decreases under the influence of GC-pressure mostly due to synonymous Arg2 to Arg4 mutations (Khrustalev and Barkovsky, 2007).

Fig. 4 shows that the main cause of the steep decrease of AA2AG usage under the influence of mutational GC-pressure is in the frequent fixation of synonymous mutations from Leu2 to Leu4 and from Arg2 to Arg4. Furthermore, Fig. 4 shows that the level of AA2TC is not decreasing as steeply as the level of AA2AG under the influence of GC-pressure because of the radicalism of mutations decreasing levels of Phe and Tyr usage. Negative selection eliminates a lot of mutations decreasing Phe and Tyr usage; however, levels of their usage became a little lower in proteins encoded by GC-rich genes (see Fig. 4a and c). Once again, usage of Phe and Tyr is not as low as usage of Leu2 and Arg2 in proteins encoded by GC-rich genes (see Fig. 4).

In Fig. 5a one can see that the usage of glutamine, which is encoded by two codons containing thymine or cytosine in third position (CAA/G) declines only in proteins encoded by genes with G+C higher than 75% (Khrustalev and Barkovsky, 2009a). The same tendency (see Fig. 5b) is characteristic to the usage of the duplet of codons coding for serine (Ser2: AGT/C) (Khrustalev and Barkovsky, 2009a). Ser2 cannot be substituted by the quartet of codons coding for the same amino acid (Ser4: TCX) by the way of a single nucleotide mutation.

Histidine and cysteine are encoded by codons containing thymine or cytosine in third positions. So, the number of amino acids encoded by codons with T or C in third codon positions is higher than the number of amino acids encoded by codons with A or G in third codon positions. Levels of histidine and cysteine usage show no correlation on G+C (see Fig. 5c and d), although their levels of usage decrease in proteins encoded by genes with extremely high G+C (Khrustalev and Barkovsky, 2009a).

The final conclusion is the following. The level of AA2AG is decreasing much more steeply than the level of AA2TC under the influence of mutational GC-pressure because of the neutrality of mutations leading to Leu2 and Arg2 disappearance (this is due to the feature of universal genetic code, which contains three six-fold degenerated series of codons). On the other hand, the radicalism of phenylalanine and tyrosine replacements is the factor making the difference between AA2AG and AA2TC much higher, especially, with the growth of GC-content.

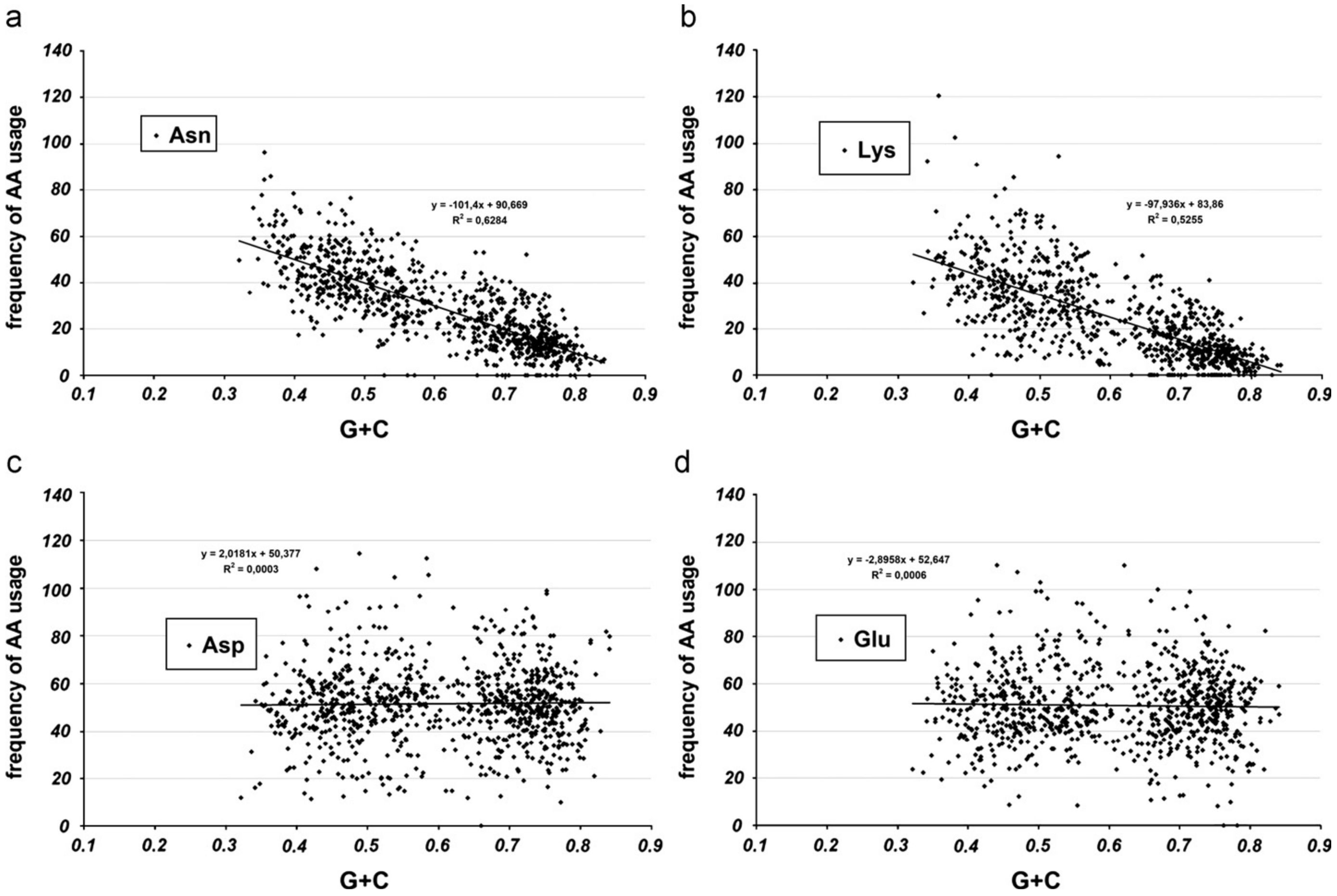

Fig. 3. Dependence between G+C in coding regions and (a) level of asparagine usage; (b) level of lysine usage; (c) level of aspartic acid usage and (d) level of glutamic acid usage in proteins of Simplex- and Varicelloviruses.

3.4. *Comparison between usage of amino acids encoded by single codons with guanine in third position and usage of isoleucine*

To study complete set of codons we have to analyze the usage of a single codon coding for methionine (ATG) and a single codon coding for tryptophan (TGG). They both contain guanine in third position. As you can see in Fig. 6a, the usage of methionine shows only weak inversed correlation on G+C. There is no correlation between the usage of Trp and G+C (see Fig. 6b) (Khrustalev and Barkovsky, 2009a).

Isoleucine is encoded by three codons (ATT, ATA and ATC). One of these codons contains cytosine in third position. The usage of Ile is some higher than the sum of Met and Trp levels of usage in proteins coded by genes with G+C from 50% to 60% (an average difference is equal to 9.4×10^{-3}), while there is no significant difference between Ile and Met+Trp in proteins coded by genes with G+C from 60% to 70%. In proteins coded by genes with G+C higher than 70% the usage of Ile is even lower (an average difference is equal to 7.1×10^{-3}) than the sum of Met and Trp. These data show that the usage of ATC codon does not make a contribution into the difference between 3C and 3G in genes with high GC-content.

3.5. *Levels of CpG and GpC usage in coding regions from Simplex- and Varicelloviruses*

Distribution of cytosine and guanine usage between codon positions described in this work may have an impact on the frequencies of CpG and GpC dinucelotides usage. On the other hand, frequencies of CpG and GpC usage may have an impact on the distribution of cytosine and guanine usage between codon positions. To clarify these relationships we calculated frequencies of CpG and GpC dinucleotides usage in each of the coding genomes studied (see Table 1).

Cytosine residues included in CpG dinucleotides are targets for DNA-methyltransferases. Methylated cytosine (5-methyl-C) forms thymine in case of its deamination, while nonmethylated cytosine forms uracil in case of its deamination. It was suggested that CpG dinucleotides are hot spots for mutation and their usage should decrease because of frequent C to T and G to A transitions. This hypothesis was created in times when nothing was known about thymine-DNA-glycosylases. These reparative enzymes excise thymine from T:G mispairs which can occur by the way of 5-methyl-C to T mutation (Gros et al., 2002). So, CpG dinucleotides are not very much “hotter” spots for mutation than GpC dinucleotides in case if thymine-DNA-glycosylases are functional.

The genome of HSV1 is never methylated during latency (Low et al., 1969). HSV1 establishes latency in neural tissue. The level of CpG was shown to be higher than the level of GpC in HSV1 genome (Honess et al., 1989). So, it was hypothesized that the absence of DNA methylation is responsible for higher amount of CpG dinucleotides in HSV1 genome (Honess et al., 1989).

The genome of Epstein-Barr virus (Gammaherpesvirus) is methylated during latency (Honess et al., 1989). This virus is able to establish latency only in lymphoid tissue. The level of CpG was shown to be lower than the level of GpC in DNA of Epstein-Barr virus (Honess et al., 1989).

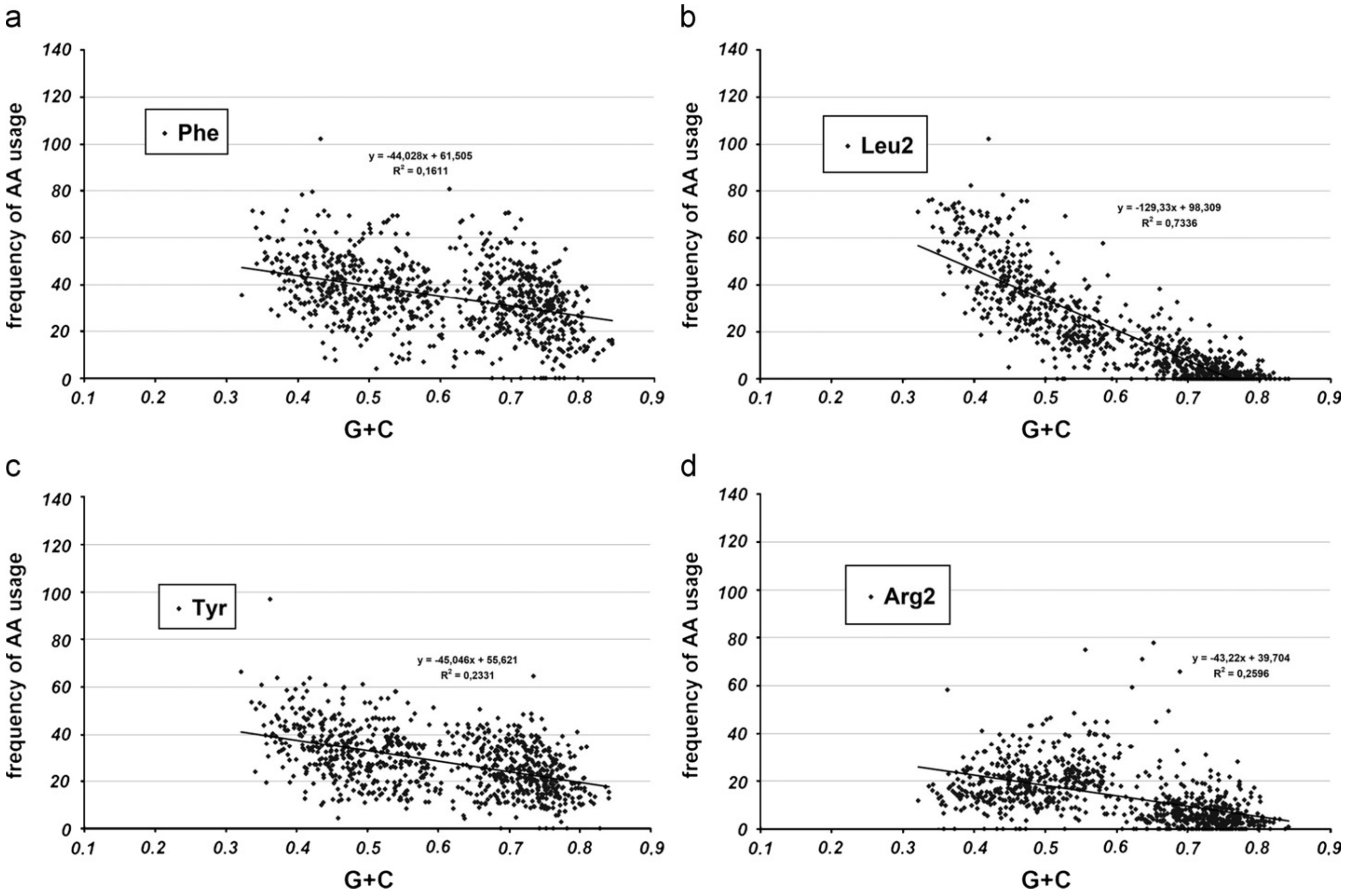

Fig. 4. Dependence between G+C in coding regions and (a) level of phenylalanine usage; (b) level of the duplet of leucine usage (Leu2); (c) level of tyrosine usage and (d) level of the duplet of arginine usage (Arg2) in proteins of Simplex- and Varicelloviruses.

The following hypothesis has been formulated (Honess et al., 1989). Genomes of herpesviruses are methylated during their latency in lymphoid tissue, this leads to the decrease of CpG dinucleotides usage in their DNA (Honess et al., 1989). Genomes of herpesviruses are not methylated during their latency in neural tissue, this leads to the excess of CpG dinucleotides in their DNA (Honess et al., 1989).

Since genetic code is organized in trinucleotides, but not in dinucleotides, each dinucleotide can occupy different codon positions. Table 1 shows that the frequency of CpG dinucleotides situated in first and second codon positions (CpG 1-2) is significantly lower than the frequency of GpC dinucleotides situated in first and second codon positions (GpC 1-2) in each of the viral coding genomes studied. Alanine is encoded by codons containing GpC dinucleotide in their first and second positions. CpG 1-2 is actually a frequency of Arg4 usage. Arginine is encoded by the quartet of codons containing CpG dinucleotide in their first and second positions (Arg4) and by the duplet of codons (Arg2: AGA and AGG). In all the viral proteomes studied the usage of Ala is higher than the usage of Arg4, including the proteome of HSV1, the genome of which is not methylated in vivo.

We found out that in all the twelve viral genomes the level of CpG 3-1 is much higher than the level of GpC 3-1 (see Table 1). To understand the cause of this inequality one should come back to Fig. 1. Indeed, the level of 1G is much higher than the level of 1C, and the level of 3C is usually higher than the level of 3G.

The level of CpG 2-3 is higher than the level of GpC 2-3 in all the viral genomes, except those of three Varicelloviruses infecting horses (EqHV1, EqHV4 and EqHV9). In genomes of EqHV1 and EqHV9 the difference between CpG 2-3 and GpC 2-3 is not significant (see Table 1), while in genome of EqHV4 the level of CpG 2-3 is significantly lower than the level of GpC 2-3. In general, level of 2C is higher than the level of 2G (see Fig. 1d), but the level of 3G is usually lower than the level of 3C (see Fig. 1b).

EqHV1, EqHV4 and EqHV9 are able to establish latency in both neural and lymphoid tissues (Takács et al., 2001). So, it was suggested that the genome of EqHV4 is methylated during latency in lymphoid tissue, and that this circumstance leads to the decrease of CpG dinucleotides (Takács et al., 2001). Here we show that the total level of CpG dinucleotides which can mutate synonymously (CpG 2-3 plus CpG 3-1 is equal to 4.6 ± 0.2%) is higher (yet, insignificantly) than the total level of GpC dinucleotides which can mutate synonymously (GpC 2-3 plus GpC 3-1 is equal to 4.4 ± 0.1%) in coding regions from EqHV4 genome. However, the level of CpG dinucleotides which cannot mutate synonymously (CpG 1-2) is much lower than the level of GpC dinucleotides which cannot mutate synonymously (GpC 1-2) in EqHV4 genome. For proteomes of EqHV1, EqHV4 and EqHV9 the difference between Ala and Arg4 is higher than that for proteome of HSV1. This should be the consequence of the increase in Arg2 codons usage in EqHV1, EqHV4 and EqHV9 genes (see Table 1) due to higher GC to AT trasversions (but not transitions) rates causing Arg4 to Arg2 synonymous mutations.

Interestingly, all guanine residues from CpG 2-3 dinucleotides are situated in four-fold degenerated sites, while only a half of cytosine residues from GpC 2-3 dinucleotides are situated in that kind of sites. It means that the level of CpG 2-3 may be changed not only by possible synonymous transitions caused by

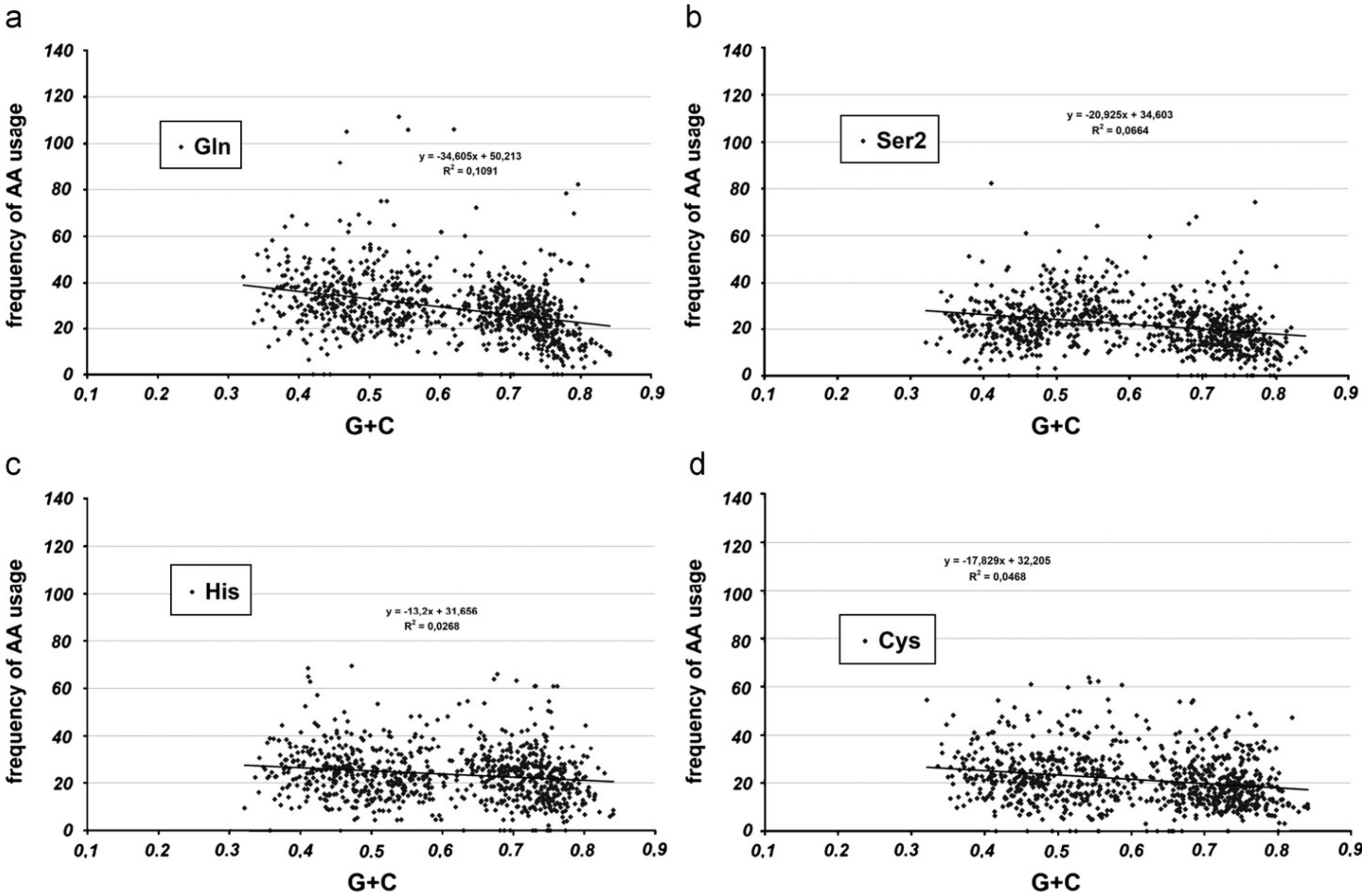

Fig. 5. Dependence between G+C in coding regions and (a) level of glutamine usage; (b) level of the duplet of serine usage (Ser2); (c) level of histidine usage and (d) level of cysteine usage in proteins of Simplex- and Varicelloviruses.

5-methyl-C to T mutations on opposite DNA strand but also by synonymous transversions.

As we have shown previously (Khrustalev and Barkovsky, 2009b), EqHV4 genome is the only one among Simplex- and Varicelloviruses in which the level of GC-content in four-fold degenerated sites is lower than the level of GC-content in two-fold degenerated sites. So, high rates of GC to AT transversions may be responsible for the decrease in CpG 2-3 dinucleotide usage in EqHV4 rather than DNA-methylation.

In the coding genome of BoHV5 the level of CpG is also significantly lower than the level of GpC (see Table 1). However, this virus is neurotropic and has never been shown to establish latency in lymphoid tissue (Perez et al., 2002). The main cause of the excess of GpC dinucleotides in BoHV5 is the elevated usage of alanine (encoded by codons with GpC 1-2) in its proteome. As one can see in Table 1, the level of CpG 1-2 (coding for Arg4) is practically the same in BoHV5 and CeHV1, while the level of GpC 1-2 (coding for Ala) is much higher in coding regions of BoHV5. Once again, we have to state that CpG dinucelotides able to mutate synonymously are used much more frequently in BoHV5 genome than "synonymous" GpC dinucleotides (see Table 1).

In general, decreased usage of CpG in BoHV5 as well as in EqHV1, EqHV4 and EqHV9 coding genomes cannot be caused by DNA-methylation alone, since the level of CpG dinucleotides which can mutate synonymously by the way of transition in BoHV5, EqHV1 and EqHV9 are significantly higher than the level of "synonymous" GpC dinucleotides, the difference between levels of "synonymous" CpG and "synonymous" GpC dinucleotides in EqHV4 is not significant.

Analogous calculations in bacterial genomes showed that the difference between GpC and CpG is not a good marker of DNA methylation (Wang et al., 2004). One should remember that Arg4 and Ala usage have an impact on CpG and GpC usage (Khrustalev and Barkovsky, 2007). Mutational processes different from 5-methyl-C to T transitions may influence CpG and GpC frequencies. So, *the fact of DNA methylation or the absence of DNA methylation should be confirmed in vivo, and not by the way of CpG and GpC comparisons, especially in coding regions.*

3.6. Comparison between early and late genes from HSV1 genome

Theoretically, differences in time of expression and in the level of expression should cause differences in the intense (and sometimes even in the direction) of transcription-associated mutational pressure for differentially expressed genes. However, this effect may be "hidden" by other mutational processes. For example, ICPO and ICP4 immediate-early genes are situated in long inverted repeats (LTR) of HSV1 genome, while other immediate-early genes are situated in unique long (UL) or unique short (US) regions of the genome. As we have shown previously, GC-content in genes from LTR is much higher than GC-content in UL and US because the probability of AT to GC transversion is higher in inverted repeats than in unique regions (Khrustalev and Barkovsky, 2009b).

Time of gene expression is known for HSV1 genes (Honess and Roizman, 1974), while for other Simplex- and Varicelloviruses the pattern of gene expression may be different. Indeed, genes

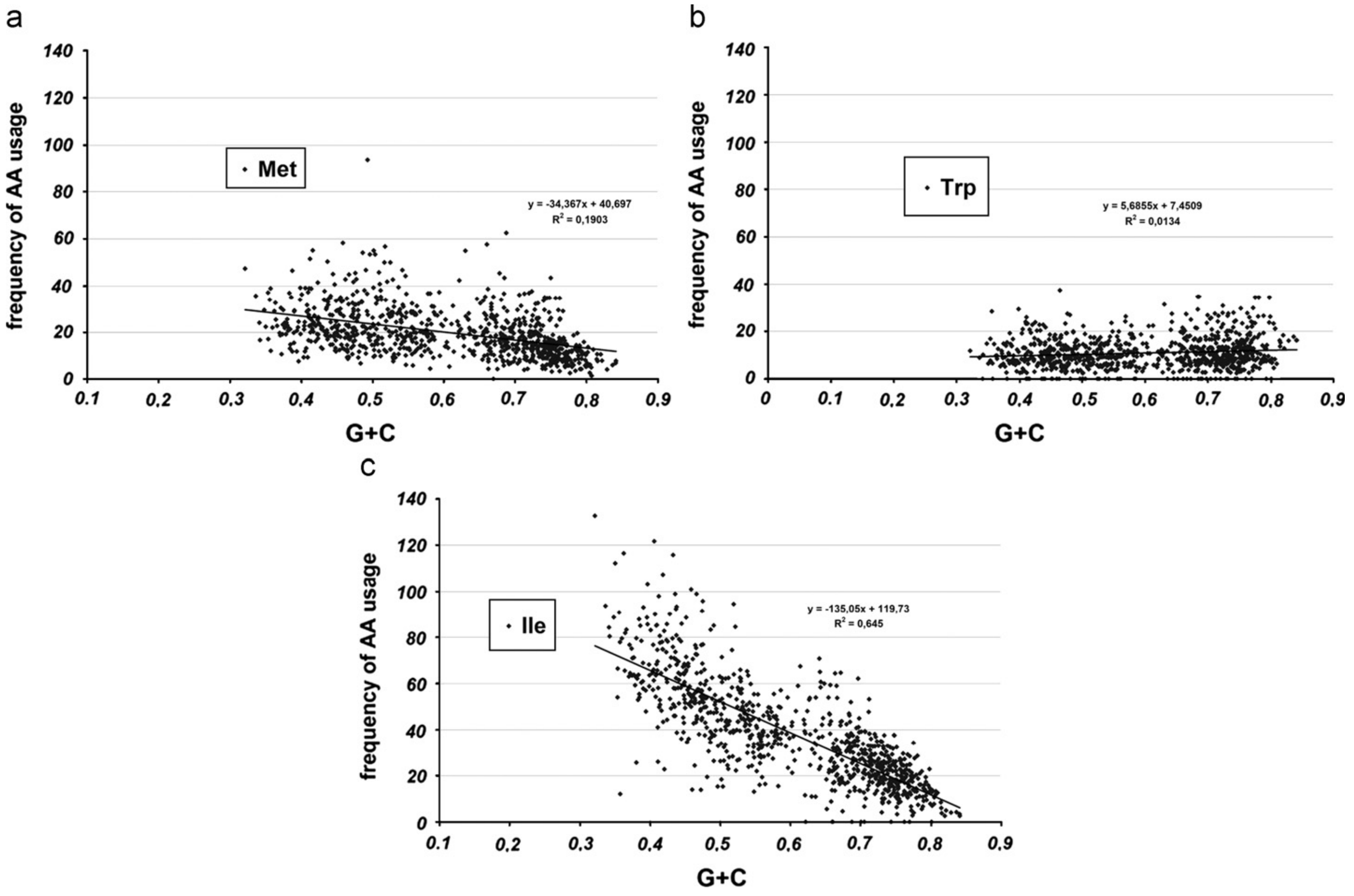

Fig. 6. Dependence between G+C in coding regions and (a) level of methionine usage; (b) level of tryptophan usage and (c) level of isoleucine usage in proteins of Simplex- and Varicelloviruses.

Table 1
Distribution of CpG and GpC dinucleotides between three codon positions of coding regions from Simplex- and Varicelloviruses. Insignificant difference between frequencies of CpG and GpC in certain codon positions is shown by *italic* font.

Virus	G+C	GpC 1-2 (Ala)	CpG 1-2 (Arg4)	AGA/AGG 1-2-3 (Arg2)	GpC 2-3	CpG 2-3	GpC 3-1	CpG 3-1	GpC	CpG
CeHV16	75.3 ± 0.4	5.2 ± 0.2	3.2 ± 0.1	0.2 ± 0.1	4.3 ± 0.1	5.0 ± 0.1	4.8 ± 0.1	7.8 ± 0.1	14.3 ± 0.2	15.9 ± 0.2
CeHV2	75.1 ± 0.4	5.2 ± 0.2	3.3 ± 0.1	0.2 ± 0.1	4.3 ± 0.1	5.0 ± 0.1	4.8 ± 0.1	7.6 ± 0.2	14.3 ± 0.2	15.9 ± 0.2
CeHV1	73.5 ± 0.4	5.0 ± 0.2	3.0 ± 0.1	0.2 ± 0.1	4.1 ± 0.1	4.8 ± 0.1	4.6 ± 0.1	7.4 ± 0.2	13.7 ± 0.2	15.2 ± 0.2
BoHV5	75.0 ± 0.4	6.2 ± 0.2	3.1 ± 0.1	0.1 ± 0.1	5.0 ± 0.1	5.6 ± 0.1	6.0 ± 0.1	8.0 ± 0.2	17.2 ± 0.2	16.7 ± 0.2
HSV2	69.2 ± 0.5	4.4 ± 0.1	2.8 ± 0.1	0.3 ± 0.1	3.5 ± 0.1	4.2 ± 0.1	3.9 ± 0.1	6.0 ± 0.1	11.7 ± 0.2	13.0 ± 0.2
HSV1	67.6 ± 0.5	4.0 ± 0.1	2.7 ± 0.1	0.3 ± 0.1	3.2 ± 0.1	3.7 ± 0.1	3.5 ± 0.1	5.4 ± 0.1	10.8 ± 0.2	11.8 ± 0.2
EqHV1	56.9 ± 0.6	3.3 ± 0.1	1.7 ± 0.1	0.7 ± 0.1	*2.7 ± 0.1*	*2.6 ± 0.1*	2.7 ± 0.1	4.0 ± 0.1	8.7 ± 0.2	8.2 ± 0.2
EqHV9	56.4 ± 0.6	3.3 ± 0.1	1.7 ± 0.1	0.7 ± 0.1	*2.6 ± 0.1*	*2.5 ± 0.1*	2.7 ± 0.1	3.9 ± 0.1	8.6 ± 0.2	8.0 ± 0.2
EqHV4	50.6 ± 0.6	3.0 ± 0.1	1.5 ± 0.1	0.7 ± 0.1	2.2 ± 0.1	1.8 ± 0.1	2.2 ± 0.1	2.7 ± 0.1	7.4 ± 0.2	6.1 ± 0.2
VZV	46.5 ± 0.6	2.4 ± 0.1	1.7 ± 0.1	0.8 ± 0.1	1.3 ± 0.1	2.2 ± 0.1	1.4 ± 0.1	2.4 ± 0.1	5.2 ± 0.1	6.3 ± 0.2
FeHV1	45.3 ± 0.5	2.1 ± 0.1	1.5 ± 0.1	0.8 ± 0.1	1.0 ± 0.1	1.5 ± 0.1	1.1 ± 0.1	2.0 ± 0.1	4.2 ± 0.1	5.0 ± 0.2
CeHV9	40.5 ± 0.7	2.2 ± 0.1	1.3 ± 0.1	0.6 ± 0.1	1.0 ± 0.1	1.5 ± 0.1	1.2 ± 0.1	1.8 ± 0.1	4.4 ± 0.1	4.7 ± 0.2

expressed during latency of Varicello-zoster virus are not the same as genes expressed during latency of HSV1 and even as genes expressed during Simian varicella virus (CeHV9) latency (Ou et al., 2007). So, we decided not to consider homologues of early and late HSV1 genes to be early and late genes in other viruses.

Comparison between indexes describing nucleotide usage (see Section 2) calculated for 29 early and 34 late genes of HSV1 (by *t*-test) showed that differences between them are insignificant. This result is not the evidence that mutational pressure is equal for early and late genes. Actually, indexes counted for genes inside each group are too variable. So, not only the time of expression but the level of expression should affect the nucleotide content of a gene. Levels of expression vary for different early and different late genes widely (see the animation on the following web page describing expression of some HSV1 early and late genes: http://darwin.bio.uci.edu/~faculty/wagner/rnatime.html) (Stingley et al., 2000). As we have mentioned before, location of a gene highly affects its nucleotide usage because of the influence of replication-associated mutational pressure. One should also remember that frameshifting, tandem repeats, insertions and deletions may cause biases in nucleotide usage.

4. Discussion

4.1. Features of genetic code and negative selection make a major contribution into the elevated cytosine usage in two-fold degenerated sites situated in third codon positions

Codons containing guanine or cytosine in third positions are used to code for amino acids in genes from GC-rich genomes (Singer and Hickey, 2000; Khrustalev and Barkovsky, 2009b). Mutational GC-pressure surely causes both synonymous and nonsynonymous AT to GC mutations. Mutations in third codon positions are fixed much frequently than mutations in first and second positions. However, the frequency of the fixation of relatively neutral amino acid replacements caused by nonsynonymous AT to GC mutations is much higher than that for relatively radical amino acid replacements (Khrustalev and Barkovsky, 2009a).

According to our data (see Fig. 4a and c), replacements of phenylalanine and tyrosine are much more radical than replacements of isoleucine, lysine and asparagine. There is still high usage of GC-poor codons coding for phenylalanine and tyrosine in genes with high GC-content. These codons contain cytosine but not guanine in their third codon positions.

The usage of Leu2 and Arg2 codons with guanine in third codon position is close to zero in GC-rich genes (see Fig. 4b and d). These codons can be synonymously substituted by GC-rich codons from Leu4 and Arg4 series. Interestingly, the usage of isoleucine shows even higher slope of the dependence on GC-content than codons from Leu2 duplet. It seems like replacements of isoleucine are fixed with the rates closer to that for synonymous mutation. So, the main cause of the increase of the difference between C2f3p and G2f3p with the growth of G+C is in the radicalism of tyrosine and phenylalanine replacements.

Indeed, isoleucine can be substituted to amino acids with close physico-chemical features by the way of AT to GC mutations (A to G transition in first codon position causes Ile to Val replacement; A to C transversion in first codon position causes Ile to Leu replacement). Isoleucine, leucine and valine have hydrophobic aliphatic side chains.

Asparagine can be substituted to aspartic acid by the way of A to G transition in first codon position. Lysine can be substituted by arginine (Arg2) by the way of A to G transition in second codon positions. Lysine and arginine are amino acids with positively charged long side chains. So, they could substitute each other in most of the cases without significant changes in the function of the protein.

Phenylalanine and tyrosine both have aromatic side chains. They cannot be substituted by any other amino acid with similar physico-chemical features of its side chain. It means that substitutions of phenylalanine and tyrosine usually cause significant changes in structure and function of the protein. The most of these changes are "negative" for the viral fitness, so they are usually eliminated by natural selection.

Described features of genetic code produce "asymmetric" negative selection eliminating the most of radical amino acid changes and so leading to the "pyrimidine-loading" (actually, cytosine-loading) of viral coding districts and mRNAs under the influence of "symmetric" mutational GC-pressure. The number of two-fold degenerated sites in which cytosine can be situated is much higher in GC-rich genes than the number of two-fold degenerated sites in which guanine can be found.

4.2. Mechanisms of transcription-associated mutational "C-pressure" making minor but significant contribution into the elevated cytosine usage

There is statistically significant but relatively small (2.2–4.3%) difference between the level of cytosine and level of guanine in four-fold degenerated sites of GC-rich genes studied. The most of these genes belong to GC-rich genomes of all Simplexviruses included in this study, as well as to the GC-rich genome of Bovine herpesvirus 5 (Khrustalev and Barkovsky, 2009b). This difference cannot be attributed to the replication-associated mutational pressure. The round of theta-replication of Simplex- and Varicelloviruses begins after the circularization of their genomes (Strang and Stow, 2005). It means that the length of leading strands is equal to the length of lagging strands. So, there should be transcription-associated mutational pressure along with replication-associated pressure in these viral genomes.

Uracil-DNA-glycosylase is encoded by Herpes simplex virus type 1, as well as by all other Simplex- and Varicelloviruses. This enzyme was shown to excise uracil both from mispairs and from singlestranded DNA (Bellamy and Baldwin, 2001). However, uracil-DNA-glycosylase from Herpes simplex virus type 1 preferably excises uracil from singlestranded DNA (Bellamy and Baldwin, 2001).

Deamination of cytosine residues in DNA leads to the appearance of uracil. If this mutation occurs in a coding region of DNA, uracil residue will be excised by uracil-DNA-glycosylase during transcription with a high probability. The probability of C to U mutation reparation should be some lower in case if this mutation occurs in a part of DNA strand complementary to a coding region. Being unprepared this mutation will be inherited by a coding region as G to A transition. So, *features of viral uracil-DNA-glycosylase should lead to the higher frequency of G to A transitions relatively to the frequency of C to T transitions in coding regions of DNA.*

As we have shown in our previous work, the leading mechanism of mutational GC-pressure in Simplexviruses and in Bovine herpesvirus 5 is the elevated rates of AT to GC transversions (Khrustalev and Barkovsky, 2009b). These transversions occur due to incorporation of oxidized guanine (8-oxo-G) into the growing strands of DNA against adenine (Gros et al., 2002). Reparation of the 8-oxo-G:A mispair usually leads to A to C mutation. MutY glycosylase excises adenine from this mispair. Then cytosine is incorporated against 8-oxo-G. Finally, MutM glycosylase exices 8-oxo-G from this newly formed mispair (8-oxo-G:C) and guanine is incorporated against cytosine (Gros et al., 2002).

There are enzymes (Neil-1 and Neil-2) able to excise 8-oxo-G from singlestranded DNA (Dou et al., 2003). These enzymes may prevent promutagenic repair of A:8-oxo-G mispair described above. If 8-oxo-G is situated in a coding region, it may be excised by Neil-1 or Neil-2 during transcription. In this case thymine will occur in the place of abasic site. *The probability to be excised by Neil-1 and Neil-2 should be higher for 8-oxo-G incorporated into a coding region than for 8-oxo-G incorporated into a strand of DNA complementary to a coding region.*

According to our hypothesis, G to A transitions and A to C transversions should be more frequent than C to T transitions and T to G transversions, respectively, in coding regions of GC-rich Simplex- and Varicelloviruses. These transcription-associated mutational processes should make the level of C some higher than the level of G in four-fold degenerated sites of coding regions.

References

Bellamy, S.R.W., Baldwin, G.S., 2001. A kinetic analysis of substrate recognition by uracil-DNA glycosylase from herpes simplex virus type 1. Nucl. Acids Res. 29, 3857–3863.

Chen, C., Chen, C.W., 2007. Quantitative analysis of mutation and selection pressures on base composition skews in bacterial chromosomes. BMC Genom. 8, 286.

Cristillo, A.D., Mortimer, J.R., Barrette, I.H., Lillicrap, T.P., Forsdyke, D.R., 2001. Double-stranded RNA as a not-self alarm signal: to evade, most viruses purine-load their RNAs, but some (HTLV-1, Epstein-Barr) pyrimidine-load. J. Theor. Biol. 208, 475–489. doi:10.1006/jtbi.2000.2233.

Dou, H., Mitra, S., Hazra, T.K., 2003. Repair of oxidized bases in DNA bubble structures by human DNA glycosylases Neil1 and Neil2. J. Biol. Chem. 278 (50), 49679–49684.

Forsdyke, D.R., 2006. Evolutionary Bioinformatics. Springer, New York 273pp.

Geacintov, J.A., Shafirovich, N.E., 2003. DNA lesions derived from the site selective oxidation of Guanine by carbonate radical anions. Chem. Res. Toxicol. 16, 1528–1538.

Grigoriev, A., 1999. Strand-specific compositional asymmetries in double-stranded DNA viruses. Virus Res. 60 (1), 1–19. doi:10.1016/S0168-1702(98)00139-7.

Gros, L., Saparbaev, M.K., Laval, J., 2002. Enzymology of the repair of free radicals-induced DNA damage. Oncogene 21, 8905–8925.

Hanawalt, P.C., Spivak, G., 2008. Transcription-coupled DNA repair: two decades of progress and surprises. Nat. Rev. Mol. Cell. Biol. 9 (12), 958–970.

Honess, R.W., Gompels, U.A., Barrell, B.G., Craxton, M., Cameron, K.R., Staden, R., Chang, Y.N., Hayward, G.S., 1989. Deviations from expected frequencies of CpG dinucleotides in herpesvirus DNAs may be diagnostic of differences in the states of their latent genomes. J Gen. Virol. 70 (4), 837–855.

Honess, R.W., Roizman, B., 1974. Regulation of herpesvirus macromolecular synthesis. I. Cascade regulation of the synthesis of three groups of viral proteins. J Virol. 14 (1), 8–19.

Hu, J., Zhao, X., Zhang, Z., Yu, J., 2007. Compositional dynamics of guanine and cytosine content in prokaryotic genomes. Res. Microbiol. 158 (4), 363–370. doi:10.1016/j.resmic.2007.02.007.

Khrustalev, V.V., Barkovsky, E.V., 2007. Levels of CpG and GpC dinucleotides in coding districts of archaeal genomes. Computational Phylogenetics and Molecular Systematics "CPMS' 2007", pp. 354–357.

Khrustalev, V.V., Barkovsky, E.V., 2009a. Main pathways of proteome simplification in alphaherpesviruses under the influence of the strong mutational GC-pressure. J. Proteom. Bioinformat. 2 (2), 88–96.

Khrustalev, V.V., Barkovsky, E.V., 2009b. Mutational pressure is a cause of inter- and intragenomic differences in GC-content of simplex and varicello viruses. Comput. Biol. Chem. 33 (4), 295–302. doi:10.1016/j.compbiolchem.2009.06.005.

Lobry, J.R., Sueoka, N., 2002. Asymmetric directional mutation pressures in bacteria. Genome Biol. 3, 0058.

Low, M., Hay, J., Keir, H.M., 1969. DNA of herpes simplex virus is not a substrate for methylation in vivo. J. Mol. Biol. 46 (1), 205–207. doi:10.1016/0022-2836(69)90068-0.

Ou, Y., Davis, K.A., Traina-Dorge, V., Gray, W.L., 2007. Simian varicella virus expresses a latency-associated transcript that is antisense to open reading frame 61 (ICP0) mRNA in neural ganglia of latently infected monkeys. J. Virol. 81 (15), 8149–8156.

Perez, S.E., Bretschneider, G., Leunda, M.R., Osorio, F.A., Flores, E.F., Odeón, A.C., 2002. Primary infection, latency, and reactivation of Bovine herpesvirus type 5 in the Bovine nervous system. Vet. Pathol. 39 (4), 437–444.

Singer, G.A.C., Hickey, D.A., 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol. Biol. Evol. 17, 1581–1588.

Strang, B.L., Stow, N.D., 2005. Circularization of the herpes simplex virus type 1 genome upon lytic infection. J. Virol. 79 (19), 12487–12494.

Stingley, S.W., Garcia Ramirez, J.J., Aguilar, S.A., Simmen, K., Sandri-Goldin, R.M., Ghazal, P., Wagner, E.K., 2000. Global analysis of Herpes simplex virus type 1 transcription using an oligonucleotide-based DNA microarray. J. Virol. 74 (21), 9916–9927.

Takács, M., Segesdi, J., Balog, K., Mezei, M., Tóth, G., Mináróvits, J., 2001. Relative deficiency in CpG dinucleotides is a widespread but not unique feature of Gammaherpesvirinae genomes. Acta Microbiol. Immunol. Hung. 48 (3–4), 349–357.

Wang., Y., Rocha, E.P., Leung, F.C., Danchin, A., 2004. Cytosine methylation is not the major factor inducing CpG dinucleotide deficiency in bacterial genomes. J. Mol. Evol. 58 (6), 692–700.