Unusual nucleotide content of Rubella virus genome as a consequence of biased RNA-editing: comparison with Alphaviruses

Vladislav Victorovich Khrustalev*
and Eugene Victorovich Barkovsky

Department of General Chemistry,
Belarussian State Medical University,
83 Dzerzinskogo Prospect, Minsk 220000, Belarus
E-mail: vvkhrustalev@mail.ru
E-mail: barkovsky@hotmail.ru
*Corresponding author

Abstract: The usage of cytosine in third codon positions of 22 complete Rubella virus genomes (52.4%) is significantly higher than the usage of guanine (28.9%), adenine (6.9%) and uracil (11.8%). The percentage of U ↔ C transitions (55%) between 22 Rubella virus genomes is two times higher than the percentage of A ↔ G transitions (23%). Predicted microRNA from ORF1 of Rubella virus may target human APOBEC1 mRNA, blocking APOBEC1-editing of viral RNA-minus and RNA-plus strands (preventing G → A and C → U transitions, respectively), while their ADAR-editing (causing U → C and A → G transitions, respectively) occurs frequently.

Keywords: rubella virus; alphavirus; togaviridae; ADAR; APOBEC1; mutational pressure; microRNA; GC-content; inosine; 8-oxo-G; bioinformatics; RNA-editing; nucleotide content.

Reference to this paper should be made as follows: Khrustalev, V.V. and Barkovsky, E.V. (2011) 'Unusual nucleotide content of Rubella virus genome as a consequence of biased RNA-editing: comparison with Alphaviruses', *Int. J. Bioinformatics Research and Applications*, Vol. 7, No. 1, pp.82–100.

Biographical notes: Vladislav Victorovich Khrustalev is an Assistant Professor, pursuing his PhD in the Department of General Chemistry at the Belarussian State Medical University. He received his Belarussian Academy of Science Award twice (in 2006 and in 2008). His research interests are in the areas of biochemistry, computational biology, immunology, virology, microbiology, genomics, proteomics and bioinformatics. All his scientific projects are connected with directional mutational pressure theory. He is a member of the American Chemical Society.

Eugene Victorovich Barkovsky is the Head of the Department of General Chemistry at the Belarussian State Medical University (since 1988). He is the recipient of a 2000 Belarussian State Award for Science and Engineering. His research interests are in the areas of biochemistry, molecular biology, molecular evolution, proteomics, immunology, computational biology and bioinformatics. He has written over 300 research articles, 6 monographs and 30 books. He is a member of the American Chemical Society. Both authors contributed equally to this work.

1 Introduction

Rubella virus is only one species of the Rubivirus genus. The structure of the Rubella virus is very similar to the structure of Alphaviruses. That is why Alphaviruses and Rubella virus have been grouped together in the Togaviridae family (Strauss and Strauss, 1994).

It is known that GC-content of the genome of Rubella virus (which is presented by the single-stranded RNA-plus strand) is extremely high (Katow and Matsuno, 1980), unlike the GC-content of most of the Alphaviruses. However, the term 'GC-content' is not entirely suitable for any single-stranded viral genome, because the level of C inside it is rarely equal to the level of G. Indeed, in the completely sequenced genome of the Rubella virus reference strain the level of C (38.8%) is higher than the level of G (30.8%).

Mutational pressure is an imbalance in rates in the occurrence of different types of nucleotide mutations (Sueoka, 1988). So, the kind of mutation occurring more frequently than others will be fixed more frequently by the genetic drift (if it is neutral) or by natural selection (if it is beneficial) (Sueoka, 1988).

Most of the nucleotide substitutions in third codon positions are synonymous. That is why levels of nucleotide usage in third codon positions are indicators of the direction of mutational pressure (Khrustalev and Barkovsky, 2009). Substitutions in second codon positions and the most of those in first codon positions are non-synonymous. Many variants of non-synonymous mutations are negative for the structure and function of proteins and, thus, for the fitness of the virus (Khrustalev and Barkovsky, 2010b). That is why negative selection eliminates many more nucleotide mutations occurring in first and second codon positions relative to those occurring in third codon positions. The higher is the usage of a given nucleotide in third codon positions, the more intensive should be the process of substitutions elevating the usage of this nucleotide, and the less intensive should be the process of decreasing the usage of this nucleotide (Sueoka, 1988).

The main cause of the directional mutational pressure in genomes of RNA viruses should be the process of RNA-editing. The process of RNA-editing is widespread in nature (Deichman et al., 2005). RNA-editing machinery may edit not only a cellular but also a viral RNA (Liuharles and Samuel, 1996). RNA-editing may theoretically cause a so-called error-prone catastrophe in viral quasi-species, but on the other hand it can increase the rates of viral evolution. As we have shown in this work, the direction and intensity of RNA-editing depends much on the features of the viral life cycle.

Molecular process causing elevated rates of A to G and U to C transitions in viral RNA is the deamination of adenine residues by enzymes from the ADAR family (Liuharles et al., 1997; Maas et al., 2003). The product of adenine deamination is inosine (hypoxantine). Inosine (I) preferably forms hydrogen bounds with cytosine (Maas et al., 2003). If ADAR-editing occurs mostly in RNA-plus strands, the rates of A to G transitions should increase, if this process occurs mostly in RNA-minus strand, the rates of U to C transitions should increase.

Enzymes from APOBEC3 family are able to deaminate cytosine in single-stranded DNA (for example, in HIV and parvoviruses (Narvaiza et al., 2009)). Enzymes from the APOBEC1 family are able to deaminate cytosine in single-stranded RNA (Petit et al., 2009).

Normally, human APOBEC1 is expressed only in cells from the intestinal epithelium (Fujino et al., 1998; Lee et al., 1998) together with several co-factors to introduce site-specific C to U mutation into Apoliporotein B mRNA. Site-specificity of that mRNA editing is due to the activity of the mentioned co-factors (Deichman et al., 2005). Mouse APOBEC1 is thought to possess a different type of promoter (relatively to human APOBEC1) allowing its expression not only in intestinal epithelium (Fujino et al., 1998). Expression of human APOBEC1 is elevated in different gastrointestinal cancers (Lee et al., 1998) as well as in neurofibromatosis (Mukhopadhyay et al., 2002).

Enzymes from both ADAR and APOBEC families are thought to be involved in the cellular antiviral defence. Their expression is stimulated by viral infection via interferon signalling pathways. Alpha interferon production is stimulated by the elevated concentration of dsRNA in the cell (Liuharles and Samuel, 1996). One of the effects of alpha interferon production is the induction of ADAR expression. ADAR enzymes are able to bind only double-stranded and not single-stranded RNA.

Expression of APOBEC3 DNA-editing enzymes has been found in cells infected not only by viruses with DNA stage in their life cycles (such as HIV1 (reviewed by Khrustalev, 2009)), but also by the Influenza virus with no DNA stage (Pauli et al., 2009). Recently, expression of human APOBEC1 has been found in cirrhotic livers infected by Hepatitis B and Hepatitis C viruses (Vartanian et al., 2010).

Many viruses may somehow escape ADAR and/or APOBEC editing. The results of our work led us to the conclusion that the Rubella virus suppresses APOBEC1 editing but does not suppress ADAR editing. Genomes of Alphaviruses are edited by both ADAR and APOBEC1 RNA deaminases.

Alphaviruses can replicate in numerous cell types of at least three distinct species (in human, in other mammalian species or birds and in blood-sucking insects transmitting these viruses) (Strauss and Strauss, 1994). Some of them are even able to develop latency states in human neurons (Strauss and Strauss, 1994). So, their genomes should be edited in different ways by different homologous RNA-editing enzymes. Actually, levels of 3C, 3G, 3A and 3U are somewhere around 25% in open reading frames of the most of Alphaviruses. The Rubella virus infects a single host – Homo sapiens (Strauss and Strauss, 1994). So, the genome of the Rubella virus should be the subject of relatively stable mutational pressure, as we have shown in the present work.

2 Materials and methods

Sixteen reference GenBank records describing completely sequenced genomes of Togaviruses, as well as 21 records describing non-reference Rubella virus completely sequenced genomes have been used as a material for this work.

The accession number of the Rubella virus reference genome is NC_001545. Accession numbers of other 21 complete genome records for this virus are the following: FJ211587; FJ211588; AB222608; AB222609; AB047329; AB047330; DQ388279 – DQ388281; DQ085338 – DQ085343; L78917; AF435865; AF435866; AY258322; AY258323; AF188704. Accession numbers for the reference genomes of Alphaviruses can be found in Table 1. The record describing the reference Human mRNA coding for APOBEC1 [NM_001644] has also been used in this study.

Table 1 Results of the statistical test applied to confirm that the difference between nucleotide content in third codon positions is characteristic for the most of ORF1 and ORF2 parts

Virus	*GenBank accession number*	*ORF1*	*ORF2*
Rubella virus	NC_001545	3C > 3G; 3A < 3U	3C > 3G; 3A < 3U
Salmon pancreas disease virus	NC_003930	3C > 3G; 3A = 3U	3C > 3G; 3A = 3U
Sleeping disease virus	NC_003433	3C > 3G; 3A = 3U	3C > 3G; 3A = 3U
Eastern equine encephalitis virus	NC_003899	3C > 3G; 3A = 3U	3C > 3G; 3A = 3U
Barmah Forest virus	NC_001786	3C > 3G; 3A > 3U	3C > 3G; 3A > 3U
Sindbis virus	NC_001547	3C = 3G; 3A > 3U	3C > 3G; 3A > 3U
Getah virus	NC_006558	3C = 3G; 3A > 3U	3C > 3G; 3A > 3U
Aura virus	NC_003900	3C = 3G; 3A > 3U	3C > 3G; 3A > 3U
O'nyong-nyong virus	NC_001512	3C = 3G; 3A > 3U	3C > 3G; 3A > 3U
Chikungunya virus	NC_004162	3C = 3G; 3A > 3U	3C = 3G; 3A > 3U
Semliki forest virus	NC_003215	3C = 3G; 3A > 3U	3C = 3G; 3A > 3U
Ross River virus	NC_001544	3C = 3G; 3A > 3U	3C = 3G; 3A = 3U
Western equine encephalomyelitis virus	NC_003908	3C = 3G; 3A = 3U	3C = 3G; 3A > 3U
Highlands J virus	NC_012561	3C = 3G; 3A = 3U	3C = 3G; 3A > 3U
Venezuelan equine encephalitis virus	NC_001449	3C = 3G; 3A = 3U	3C = 3G; 3A > 3U
Mayaro virus	NC_003417	3C = 3G; 3A = 3U	3C = 3G; 3A = 3U

For the alignment of nucleotide sequences MEGA4 software has been used (Tamura et al., 2007). To build the NJ-dendrogram (see Results section) we aligned reference sequences coding for RNA-dependent-RNA-polymerase (from the first open reading frame) of 16 Togaviruses with the sequence coding for HIV1 reverse transcriptase [NC_001802] (it was used as an 'outgroup' since distant relations between these viral enzymes have been described (Iyer et al., 2003)). Both synonymous (d*S*) and non-synonymous (d*N*) evolutionary distances have been calculated by the modified Nej-Gojobori method (complete deletion) included in the recent version of MEGA software (Tamura et al., 2007). The modified Nej-Gojobori method allows the correction of d*S* and d*N* by the ratio between transitions and transversions (*R*). This ratio ($R \approx 1.0$) has been calculated by the Maximum Likelihood method.

For the calculations of nucleotide content in two open reading frames (ORF1 and ORF2) from reference genomes of Togaviruses, as well as in those from twenty-two Rubella virus complete genomes we used the 'VVK Consensus' algorithm (Khrustalev, 2009). This algorithm included in the MS Excel spreadsheet is able to calculate nucleotide content in 100 or less inserted sequences. We applied a paired differences test to confirm that the level of cytosine in third codon positions (3C) of ORF1 is higher than

3C of ORF2. To perform this test we calculated paired differences between 3C in ORF1 and 3C in ORF2 for each genome. Then those paired differences were treated in *t*-test.

For the calculations of nucleotide content distribution along the length of two open reading frames in each reference genome of Togaviruses we used another original algorithm – 'VVK in length' (Khrustalev, 2009). This algorithm cuts inserted sequence in parts of the same length and calculates the nucleotide content in each of these parts. We cut nucleotide sequences of the first open reading frame (ORF1) into 300 nucleotide parts. The second open reading frame (ORF2) is approximately two times shorter than the first one. That is why we cut nucleotide sequences of ORF2 into 150 nucleotide parts. To test whether the level of cytosine in third codon positions is significantly higher than the level of guanine in third codon positions along the length of ORF1 and ORF2 we performed the paired differences test mentioned above (we calculated differences between 3C and 3G for each part of ORF1 and ORF2 and treat them in *t*-test). We also applied this method to test the significance of the difference between 3A and 3U along the length of ORF1 and ORF2.

The 'VVK Consensus' algorithm is able to calculate the percentage of nucleotide mutations in the alignment. It builds a consensus sequence for all the sequences inserted and counts the percentage of each type of nucleotide mutation from the consensus sequence. So, we calculated this percentage for the alignment of 22 complete genomes of the Rubella virus. There are two ways of nucleotide mutation counting in the alignment. The first way (counting 'per site') is to calculate numbers of sites containing each kind of nucleotide mutation. The second way (counting 'per nucleotide') is to calculate numbers of mutated nucleotides in every sequence. The first method has been built on the hypothesis that identical nucleotide mutations in a given site of the alignment are the consequences of a single mutation in their common predecessor. The second method has been built on the hypothesis that identical nucleotide mutations in a given site of the alignment occurred independently.

The real direction of nucleotide mutation cannot be estimated just by counting from consensus sequence. That is why we calculated the sums of nucleotide mutations in opposite directions. The preferred direction for all the nucleotide mutation types has been determined with the help of the mutational pressure theory (Sueoka, 1988).

For the mapping of putative microRNAs in the reference genome of the Rubella virus we used the 'mirEval' tool (http://tagc.univ-mrs.fr/mireval/) (Xue et al., 2005). The outstanding feature of this tool is the Support Vector Machine (SVM) allowing accurate discrimination between real and pseudo pre-microRNAs. This SVM classifier built on human data can correctly identify up to 90% of the pre-microRNAs from other species, including plants and viruses, without utilising any comparative genomics information (Xue et al., 2005).

For the prediction of the secondary structure of these putative pre-microRNAs we used 'CentroidFold' software (http://www.ncrna.org/centroidfold/) (Hamada et al., 2009). This software, unlike the most of its analogues, can almost successfully predict the secondary structure of a typical tRNA (Hamada et al., 2009), so we decided that it should be the best method to visualise the secondary structure of pre-microRNAs predicted by 'mirEval'.

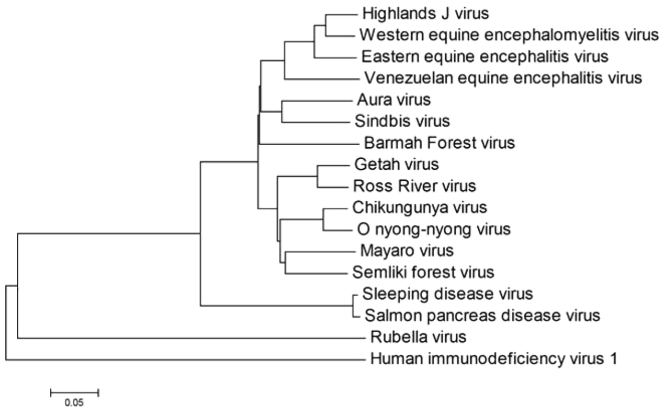
Our original 'VVK Consensus' and 'VVK in length' algorithms can be downloaded from www.barkovsky.hotmail.ru

3 Results

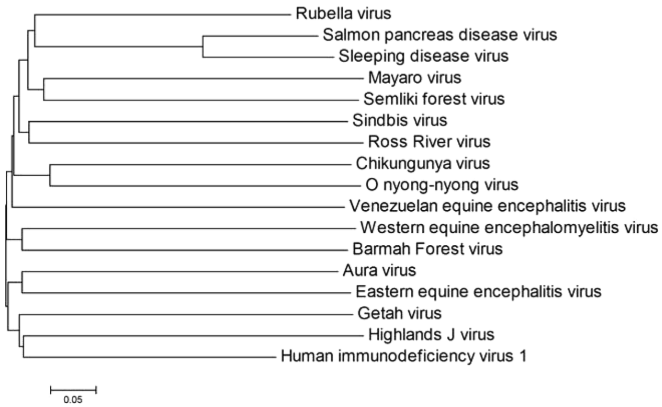
3.1 Directional mutational pressure has an impact on both synonymous and non-synonymous distances between coding regions from Togaviruses

In Figure 1(a) the phylogenetic tree of Togaviridae family is shown. This NJ-tree has been built by us on the basis of non-synonymous evolutionary distances (modified Nej-Gojobori method) between conserved regions of ORF1 coding for RNA-depended-RNA-polymerase.

Figure 1 Neighbour-joining phylogenetic trees built on the basis of non-synonymous (a) and synonymous (b) evolutionary distances (modified Nej-Gojobori method) between regions coding for RNA-dependent-RNA-polymerase of 16 viruses from Togaviridae family. Region coding for HIV1 reverse transcriptase has been used as an 'outgroup'



(a)



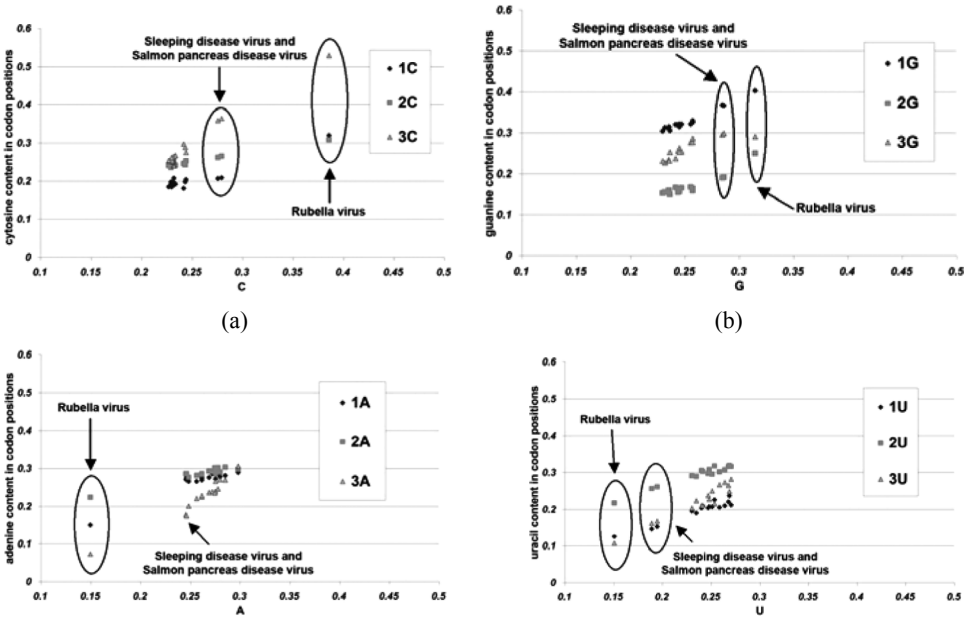
(b)

It is clearly seen in Figure 1(a) that the region coding for HIV1 reverse transcriptase is an 'outgroup', while Sleeping disease and Salmon pancreas disease viruses seem to be 'closer relatives' of the Rubella virus among Alphaviruses. However, the Rubella virus infects humans, and most of the Alphaviruses infect humans and other mammals or birds,

while Sleeping disease and Salmon pancreas disease viruses infect fish (Rainbow trout and Salmon, respectively). Why do aquatic viruses (Sleeping disease and Salmon pancreas disease viruses) demonstrate more similarity to the Rubella virus (strictly human pathogen) than other Alphaviruses (transmitted from mammals or birds to human by blood-sucking insects), such as the Sindbis virus?

The answer to this reasonable question can be found in Figures 2 and 3. In Figure 2 the nucleotide content of ORF1 from Togaviruses is shown. The most outstanding point in Figure 2(a) belongs to ORF1 from the Rubella virus. Indeed, the level of cytosine usage in Rubella virus ORF1 is the highest among its relatives. The level of cytosine in the third codon positions (3C) of the Rubella virus ORF1 is much higher than the level of cytosine in the first (1C) and second (2C) codon positions. The level of guanine (see Figure 2(b)) in the Rubella virus ORF1 is also the highest among its relatives, while G and 3G are much lower than C and 3C for ORF1 of the Rubella virus. The levels of adenine and uracil in Rubella virus ORF1 are much lower than in open reading frames of its relatives (see Figure 2(c) and (d), respectively).

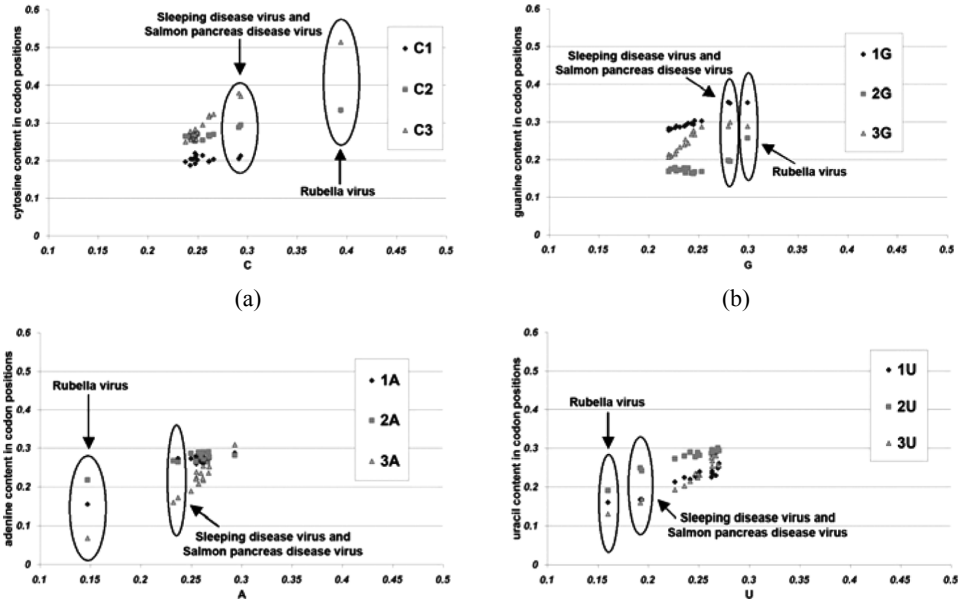
Figure 2 Dependences between (a) total cytosine content and cytosine content in three codon positions; (b) total guanine content and guanine content in three codon positions; (c) total adenine content and adenine content in three codon positions; (d) total uracil content and uracil content in three codon positions of the first open reading frame (ORF1) from 16 Togaviruses



Two other points with elevated cytosine content in Figure 2(a) belong to ORF1 from Sleeping disease and Salmon pancreas disease viruses. The cytosine content in ORF1 from these two viruses is higher than that in ORF1 from the rest of Alphaviruses but lower than that in Rubella virus ORF1. The same situation can be found in Figure 2(b). The level of uracil usage in ORF1 of Sleeping disease and Salmon pancreas disease

viruses is lower than that in ORF1 from other Alphaviruses, but higher than that in Rubella virus ORF1 (see Figure 2(c)).

Figure 3 Dependences between (a) total cytosine content and cytosine content in three codon positions; (b) total guanine content and guanine content in three codon positions; (c) total adenine content and adenine content in three codon positions; (d) total uracil content and uracil content in three codon positions of the second open reading frame (ORF2) from 16 Togaviruses



Nucleotide content in ORF2 (see Figure 3) of Togaviruses follows the same tendency as nucleotide content in ORF1: Rubella virus ORF2 has the highest cytosine and guanine content and the lowest adenine and uracil content; the nucleotide content of ORF2 from Sleeping disease and Salmon pancreas disease viruses is close to that of Rubella virus ORF2.

Looking at Figures 2 and 3 we can state that there is a 'C-pressure' in the genome of the Rubella virus as well as in the genomes of Sleeping disease and Salmon pancreas disease viruses, but the strength of mutational bias is higher in the Rubella virus. However, the level of cytosine in third codon positions is higher than 23.95% for all open reading frames from Alphaviruses. It means that the same factor producing C-pressure should act on the genomes from other Alphaviruses too.

In Figure 1(b) one can see that coding regions relatively enriched with cytosine from several other viruses (such as those from Semliki forest virus with 3C = 0.34 and Mayaro virus with 3C = 0.28) can be found near the branch containing coding regions from Rubella, Sleeping disease and Salmon pancreas disease viruses.

Interestingly, an 'outgroup' with extremely high 3A (from HIV1) can be found in the opposite part of the d*S* tree (see Figure 1(b)) together with several coding regions which have relatively elevated 3A (such as those from Highlands J virus with 3A = 0.27 and Getah virus with 3A = 0.26). Mutational pressure in the A-direction (A-pressure) has

been found in the HIV virus (Berkhout and van Hemert, 1994). The cause of this mutational bias lies in the APOBEC3-editing of HIV DNA minus strands (reviewed in works written by Narvaiza et al. (2009) and Khrustalev (2009).

The same direction of mutational pressure in distinct viruses may lead not only to a decrease in synonymous distances between their coding regions (just like in case with HIV1 and Highlands J virus), but also to a decrease in non-synonymous distances (just like in the case of the Rubella virus and aquatic Alphaviruses).

3.2 Universal rules of the nucleotide usage distribution between first and second codon positions (1G > 2G and 1U < 2U) obey in ORF1 and ORF2 from Togaviruses

One can see that guanine and uracil are distributed quite non-random between codon positions of Togaviruses open reading frames (see Figures 2 and 3). The level of 1G is always much higher than 2G, while the level of 1U is always lower than 2U. These rules of nucleotide distribution between codon positions also obey in genes from Alphaherpesviruses (Khrustalev and Barkovsky, 2010a) as well as in bacterial and archaeal genes (Khrustalev and Barkovsky, 2010b). The fact that these universal rules are obeyed by genes from single-stranded RNA-plus viruses can be interpreted as an evidence of the common ancient origin of the most of the protein-coding genes. However, these rules can simply be destroyed by frame-shifting and insertion of tandem repeats. We once described this situation in the first part of the third exon of ICP0 gene from Simplexviruses (Khrustalev and Barkovsky, 2008).

In ORF2 (see Figure 3) the difference between 1G and 2G, as well as between 2U and 1U is some lower than in ORF1 (see Figure 2). Levels of 2G and 1U are higher, while levels of 1G and 2U are lower in ORF2 than in ORF1. ORF1 encodes non-structural proteins, including RNA-depended-RNA-polymerase and RNA-helicase, which should be of a greater importance for viral survival than structural proteins (capsid and glycoproteins) encoded by ORF2. Probably, ORF2 contains more insertions, local frame-shiftings and tandem repeats than ORF1. More silent natural selection allowed fixation of these genetic events disturbing ancient rules of nucleotide distribution in ORF2.

3.3 Relatively homogenous distribution of nucleotide content in third codon positions along the length of ORF1 and ORF2 from Togaviruses; the difference in 3C between ORF1 and ORF2

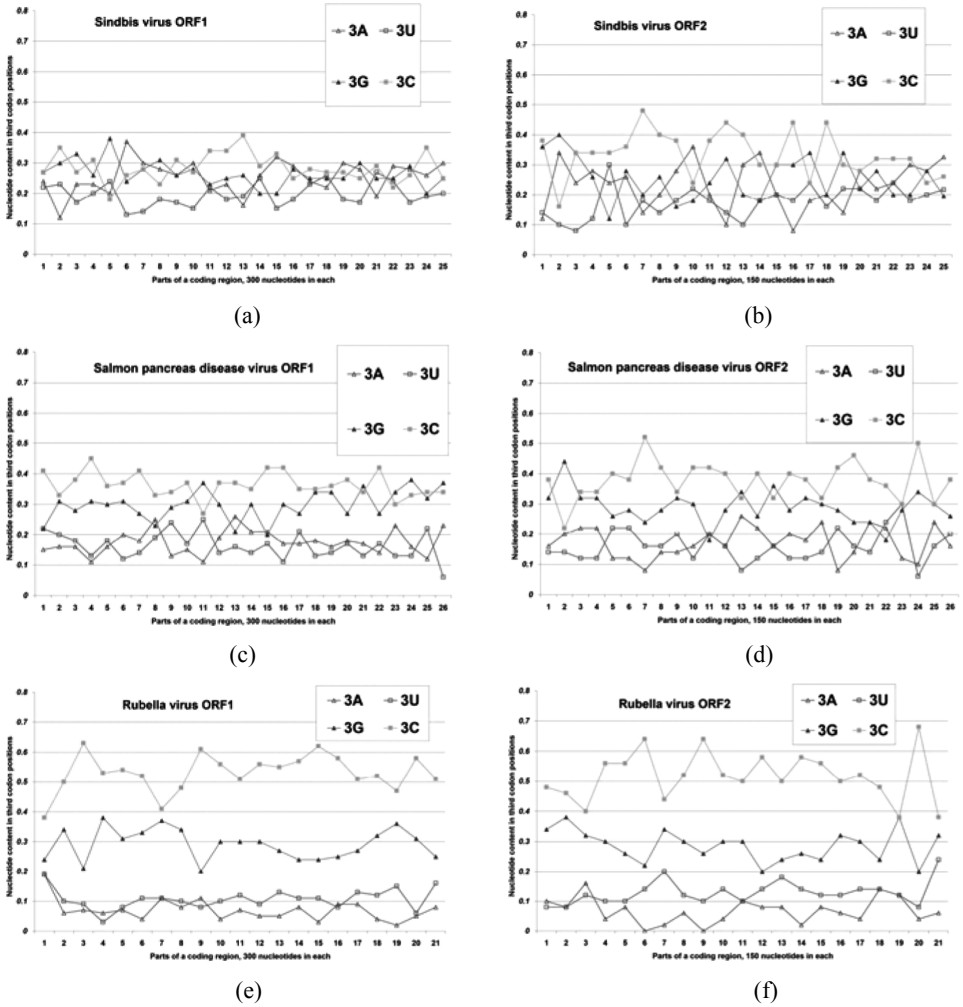
In Figure 4(a) and (b) one can see the distribution of nucleotide content in third codon positions of Sindbis virus ORF1 and ORF2. It is clear that 3C level in Sindbis virus ORF2 is higher than 3C in ORF2.

To be edited by APOBEC1, RNA should not contain elements of a secondary structure (it should not be double-stranded) (Deichman et al., 2005). Matrix RNA is usually released from multiple hairpins before the translation by translation initiation factors with RNA-helicase activity. This single-stranded viral RNA should be the substrate for APOBEC1 editing.

Indeed, genomic RNA-plus strand of Togaviruses is used for the translation of only non-structural polypeptide (Tzeng and Frey, 2005). Structural polypeptide is translated

from sub-genomic RNA containing only ORF2 which is synthesised on RNA-minus template. These facts let us make the suggestion that ORF2 from genomic RNA may not be totally released from hairpins by helicases during the translation. So, the probability to be edited by APOBEC1 should be higher for the translated region of genomic RNA (for ORF1) than for the untranslated region of genomic RNA (for ORF2). That is why the level of cytosine is significantly ($P < 0.05$) lower in third codon positions of ORF1 than in third codon positions of ORF2 in studied Alphaviruses (an average difference between 3C in ORF2 and 3C in ORF1 is $1.61 \pm 0.35\%$), and especially in Sindbis virus, in which this difference is the highest one (4.22%).

Figure 4 Nucleotide content in third codon positions along the length of two open reading frames from Sindbis virus ((a), (b)), Salmon pancreas disease virus ((c), (d)) and Rubella virus ((e), (f))



(a) (b)

(c) (d)

(e) (f)

There are two exceptions from this rule: there is practically no difference between 3C in ORF1 and 3C in ORF2 of O'nyong-nyong virus, while 3C in ORF1 is higher than that in

ORF2 of Rubella virus. This difference is significant for 22 Rubella virus genomes studied (average difference is equal to $1.95 \pm 0.17\%$, $P < 0.001$). This effect may also be caused by the characteristic feature of the viral life cycle. In our opinion, APOBEC1 editing of Rubella virus RNA is suppressed (see Section 3.7). In the absence of intensive APOBEC1 editing the level of 3C should not be lower in ORF1 relative to ORF2. ORF2 is transcribed from the Rubella virus RNA-minus template not only as a part of full-length genomic RNA, but also as a sub-genomic RNA (Lee and Bowden, 2000). Lower 3C in ORF2 may be due to the faster transcription/replication rates: ADAR should bind the following part of dsRNA intermediate not so frequently as the part of dsRNA intermediate containing ORF1. This effect should be characteristic for other Togaviruses too, while in the presence of APOBEC1 activity it becomes 'hidden' by the extensive APOBEC1 editing of ORF1.

In Figure 4(c)–(f) the level of cytosine is higher than the level of any other nucleotide in third codon positions, including guanine ($P < 0.05$). In Rubella virus 3C has reached unbelievably high levels: it varies around 52% along the whole length of ORF1 and ORF2.

3.4 Hypothesis of the unequal rates of ADAR-editing for Rubella virus RNA-minus and RNA-plus strands

Why are A to I mutations accumulated mostly in replicative strand of the Rubella virus? This happens because this strand serves as a temple for numerous viral genomes (Strauss and Strauss, 1994). All the mutations that occurred in the RNA-minus strand during all the previous rounds of replication will be inherited by the newly synthesised RNA-plus strand. The later is the period of time for this RNA-plus strand synthesis, the more will be the number of mutations in the template for its synthesis. So, replicative RNA-minus strand can serve as a kind of 'accumulator' for mutated nucleotides. That is why the level of cytosine (which is complementary to guanine and inosine), is so elevated in the Rubella virus RNA-plus strand, especially in third codon positions.

The level of 3G is also elevated in ORFs from the Rubella virus. So, its RNA-plus strands should undergo ADAR-editing (resulting in A to G mutations), while the number of A to I mutations accumulated by the RNA-minus replicative strands (resulting in U to C mutations in RNA-plus strands) is higher than that for RNA-plus strands. RNA-plus strands can be edited by ADAR during the phase of double-stranded replicational intermediate existence or in case they form secondary structures.

As one can see in Table 1, in five out of sixteen reference genomes 3C is significantly higher than 3G along the length of ORF1. For other reference genomes we cannot state that there is a difference between 3C and 3G usage in parts of ORF1 300 nucleotides in length each. In five viruses 3C is significantly higher than 3G only in ORF2, but not in ORF1 (for the explanation of this fact see the section above).

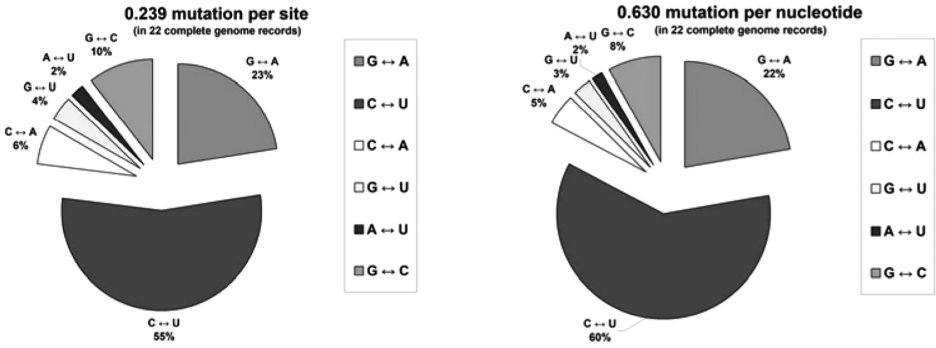
In seven viruses (see Table 1) the level of 3A is significantly higher than the level of 3U in both ORFs; in four viruses the level of 3A is significantly higher than the level of 3U in one of the ORFs. This situation may be due to the accumulation of C to U mutations in RNA-minus strands. Replication and transcription processes lead to the separation of RNA-minus and RNA-plus strands from each other. This separation is temporary. However, during this period of time the RNA-minus strand can be edited by the APOBEC1 enzyme (in case if it is single-stranded). An opposite situation ($3A < 3U$) has been observed in ORFs from the Rubella virus. This fact is consistent with our

hypothesis (see below) that the Rubella virus escapes APOBEC1-editing, unlike other Togaviruses.

3.5 Percentage of variable sites between 22 completely sequenced genomes of Rubella virus

Results of the counting of sites with mutated nucleotides for 22 complete genomes of Rubella virus are shown in Figure 5(a). There can be three types of nucleotide mutation in each site of the alignment. As one can see, the most frequent type of nucleotide mutation in the Rubella virus genome is U ↔ C transition (55%). The second place belongs to another type of transition (A ↔ G). The common percentage of all the types of transversions is just 22%. These data make us sure that the level of cytosine in Rubella virus genome has been elevated mostly due to U to C transitions and not due to A to C or G to C transversions.

Figure 5 Percentage of sites with different types of nucleotide mutation (a) and percentage of mutated nucleotides (b) in 22 completely sequenced genomes of Rubella virus



(a) (b)

The second method of nucleotide mutation number calculation (see Figure 5(b)) gives approximately the same percentage of nucleotide mutation occurrence. This method shows the distribution of polymorphism in the alignment better than the previous one, because every mutated nucleotide in each of the aligned sequences is counted.

For the most of U ↔ C mutations the direction should be from U to C, for the most of A ↔ G mutations the direction should be from A to G. We made these statements because the level of C in third codon positions is much higher than the level of U and the level of G in third codon positions is much higher than the level of A.

Oxidation of guanine is thought to be the main mechanism of GC to AT transversions. There are many works on guanine oxidation in cellular DNA (Gros et al., 2002), in which this kind of lesion (8-oxo-guanine) is the target for reparation. Genomes of single-stranded RNA viruses are not repaired. So, we can hypothesise that the most of transversions in Rubella virus, as well as in other Togaviruses, are of GC to AU direction.

The high percentage of G ↔ C transversions can be caused by two consequent mutational processes. Guanine can be oxidised, leading to G to U transversion, and then uracil can be substituted for cytosine due to ADAR-editing of the complementary strand.

3.6 Strong mutational C-pressure in the genome of Rubella virus

Although the rates of U to C transitions in the Rubella virus should be high (if this process can maintain such an elevated level of 3C), they are rarely fixed in first and second codon positions. In Figure 6 one can see that the level of cytosine in the first and second codon positions of the regions of ORF1 coding for non-structural proteins (RNA-polymerase and p150) and the region of ORF2 coding for capsid protein are quite invariable in 22 Rubella virus strains. The level of cytosine varies among coding regions from 22 Rubella virus strains mostly due to the mutations in third codon positions. This feature is characteristic for regions of ORF2 coding for glycoproteins too (see Figure 7).

Figure 6 Dependences between total cytosine content and cytosine content in three codon positions of the regions coding for RNA-dependent-RNA-polymerase, p150 and capsid protein from 22 completely sequenced genomes of Rubella virus

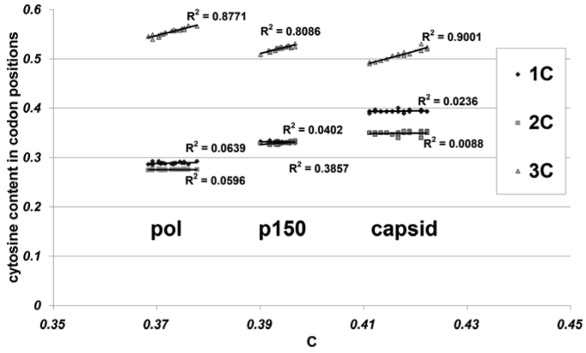
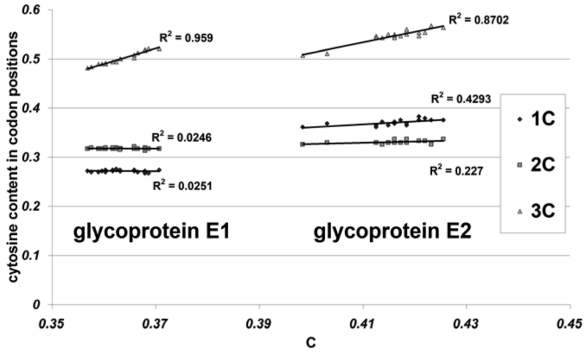


Figure 7 Dependences between total cytosine content and cytosine content in three codon positions of the regions coding for glycoprotein E1 and glycoprotein E2 from 22 completely sequenced genomes of Rubella virus



As one can see in Figures 6 and 7, the level of cytosine usage in third codon positions varies between the strains, but not between the coding regions (levels of 3C are relatively close to each other in regions coding for RNA-polymerase, p150, C, E1 and E2). On the other hand, the levels of cytosine in the first and second codon positions do not vary between the strains, but do vary between the coding regions. This situation is

characteristic for the strong directed mutational pressure (Khrustalev and Barkovsky, 2009).

The lowest levels of C1 and C2 are in the region coding for RNA-polymerase. RNA-polymerase is surely the most evolutionary conserved protein in the Rubella virus. That is why non-synonymous mutations from U to C are rarely fixed in the region coding for RNA-polymerase. The region coding for capsid protein shows the highest total C-content. The probability for the amino acid replacement caused by non-synonymous U to C mutation to bring negative consequences for the fitness of the virus should be lower for capsid protein than for any other Rubella virus protein.

3.7 Hypothesis of APOBEC1-editing escaping by Rubella virus: putative microRNA precursor in its ORF1

As we have suggested before, RNA-minus strands of all the genomes of Togaviruses are edited by ADAR. According to our results, RNA-minus and RNA-plus strands of Alphaviruses should also be edited by APOBEC1. Probably, the Rubella virus somehow inactivates APOBEC1 and not ADAR. There are many specific and non-specific ways for the virus to reduce the activity of any cellular enzyme (Adamo et al., 2008). In recent years a lot of attention has been paid to the production of viral microRNAs and siRNAs (Donaire et al., 2008). These molecules can block the translation of certain cellular mRNAs (Moissiard and Voinnet, 2006). Using 'mirEval' software (Xue et al., 2005) we decided to screen the Rubella virus genome for microRNA precursors and then to test whether those microRNAs can interact with APOBEC1 mRNA.

'MirEval' tool (http://tagc.univ-mrs.fr/mireval/) predicted eight putative pre-microRNAs in the reference Rubella virus genome. Five of them are located in ORF1 and three of them are located in ORF2. As one can see in Table 2, the number of predicted pre-microRNAs in the Rubella virus is relatively low in comparison with Alphaviruses. One cannot say that the genomes of Togaviruses are littered by microRNA precursors: in general, there is one pre-microRNA per 600–1600 nucleotides in their ORFs (see Table 2).

The 'seed' region of microRNA includes at least 6 nucleotides which have to be the perfect complementary sequence to the mRNA (Lewis et al., 2005). There is at least one invariable region of at least 6 nucleotides in length in every pre-microRNA from 22 Rubella virus genomes. With the help of 'CentroidFold' software (http://www.ncrna.org/centroidfold/) (Hamada et al., 2009) we predicted the secondary structure of these regions. Then we cut each region into two 'stems', obtained their reverse complement sequences and aligned them with human APOBEC1 mRNA.

Four from 16 stems of 8 pre-microRNA hairpins predicted by 'mirEval' in the RNA-plus strand of the reference Rubella virus strain contain relatively long putative 'seed' regions complementary to APOBEC1 mRNA (their length varies from 6 to 9 nucleotides).

Only one from these four relatively long regions complementary to APOBEC1 mRNA is invariable among 22 completely sequenced Rubella virus genomes. Predicted pre-microRNA containing an invariable seed region complementary to APOBEC1 mRNA is situated in ORF1 (from 2574 to 2611 nucleotide). In Figure 8(a) one can see the hairpin containing microRNA built for the consensus sequence of 11 completely sequenced Rubella virus genomes. In this figure we showed regions complementary to APOBEC1 mRNA. In Figure 8(b) one can see the hairpin containing the same

microRNA built for the consensus sequence of another 11 completely sequenced Rubella virus genomes ('CentroidFold' can build a consensus secondary structure maximum for 12 homologous RNAs) (Hamada et al., 2009). In Figure 8(b) we showed invariable nucleotides in active stem of predicted pre-microRNA.

Table 2 Number of microRNA precursors predicted by 'mirEval' in two ORFs of Togaviruses

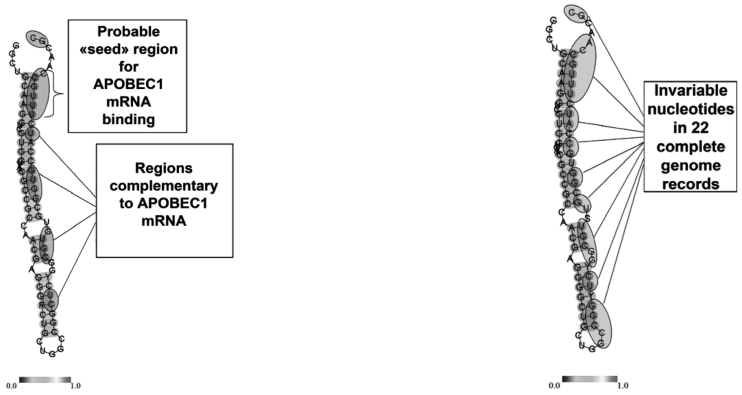
	ORF1		*ORF2*		*ORF1 + ORF2*	
Virus	*Number of microRNA precursors*	*Length, nt*	*Number of microRNA precursors*	*Length, nt*	*Number of microRNA precursors*	*Length, nt*
Rubella virus	5	6618	3	3192	8	9810
Salmon pancreas disease virus	9	7806	5	3963	14	11769
Sleeping disease virus	10	7782	4	3969	14	11751
Eastern equine encephalitis virus	9	7482	3	3726	12	11208
Barmah Forest virus	11	5385 and 1620	7	3720	18	10725
Sindbis virus	5	7542	4	3738	9	11280
Getah virus	6	7404	1	3762	7	11166
Aura virus	5	7497	3	3735	8	11232
O'nyong-nyong virus	8	7545	5	3744	13	11289
Chikungunya virus	10	7425	1	3747	11	11172
Semliki forest virus	10	7296	5	3762	15	11058
Ross River virus	8	7443	4	3765	12	11208
Western equine encephalomyelitis virus	7	7404	3	3711	10	11115
Highlands J virus	14	7353	3	3711	17	11064
Venezuelan equine encephalitis virus	7	7482	4	3768	11	11250
Mayaro virus	8	7314	9	3729	17	11043

Figure 8(c) shows the Watson-Crick base pairs between predicted microRNA and APOBEC1 mRNA. Variable nucleotides are also shown under the consensus sequence of microRNA. In some Rubella virus strains the length of a seed region is 8 nucleotides, while in the most of them its length is 6 nucleotides.

The origin of viral microRNA is surely a random event. But how does and why does this random event fix in a viral population? If the hairpin found by us in Rubella virus is really functional, it should lead to the block of APOBEC1 translation in the infected cells.

The seed region of this microRNA conserves between 22 genomes of different Rubella virus strains. It seems like the loss of this hairpin should bring negative consequences for the fitness of the virus. However, this suggestion should be tested in future in-vitro experiments.

Figure 8 Predicted secondary structure of the putative microRNA precursor from ORF1 of Rubella virus (it is mapped from 2574 to 2611 nucleotide). In Figure 8(a) regions complementary to APOBEC1 mRNA are shown, in Figure 8(b) invariable nucleotides in 22 completely sequenced genomes of Rubella virus are shown. Figure 8(c) shows base-pairing between putative microRNA and APOBEC1 mRNA; variable nucleotides are written under the consensus sequence of putative microRNA



(a) (b)

```
Apobec1   H.S.       GCUCAGGCAAAAACACCACCAAUCACGUGGAAGUUAAU
                     ||   |||||| |   ||||   ||||    ||
ORF1 Rubella virus   CGCAACCGUUUCUACCGUGGCGUGUGCGGUCUCGGCCG
                       GG       U  G  C  U  C     C  A
                                                    U
```

(c)

Indeed, APOBEC3 enzymes decrease the rates of HIV, Hepatitis B and Parvoviruses replication (Narvaiza et al., 2009). Recent experimental works showed that this effect is not only due to the cytosine deaminase activity (Narvaiza et al., 2009). Probably, APOBEC3 enzymes are able to slow the rates of viral replication because they bind single-stranded DNA which is the template for replication. Does it mean that the APOBEC1 enzyme should decrease the rates of viral RNA replication? If so, the block of APOBEC1 translation may be "beneficial" for any virus with RNA genome.

Most of the Alphaviruses, according to the results of this work, are able to replicate in the presence of active APOBEC1. There should be special features of these viruses which help them to continue their lifecycle in case of extensive APOBEC1 editing. Once the pre-microRNA allowing Rubella virus to escape APOBEC1 editing occasionally been formed, all the other special features helping the virus to replicate in the presence of APOBEC1 may become unnecessary and disappear during the course of evolution.

4 Conclusion

Our *in-silico* analyses including the study of nucleotide composition and the calculation of the percentage of nucleotide mutations allowed us to suggest that the down-regulation of APOBEC1 by the Rubella virus does exist (putative microRNA precursor has been found), while ADAR activity is not reduced in cells infected by the Rubella virus.

Acknowledgements

We are grateful to Professor Lubov Vladimirovna Khotileva, academician of Belarussian State Academia of Science, for her productive suggestions during review of our monograph on original bioinformatical methods (all of them have been used in the present work) and her ongoing interest in our findings.

References

Adamo, M.P., Zapata, M. and Frey, T.K. (2008) 'Analysis of gene expression in fetal and adult cells infected with rubella virus', *Virol.*, Vol. 370, pp.1–11.

Berkhout, B. and van Hemert, F.J. (1994) 'The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins', *Nucleic Acids Res*., Vol. 22, pp.1705–1711.

Deichman, A.M., Choi, W.C. and Baryshnikov, A.Y. (2005) *RNA-Editing. Hypothetical Mechanisms, "Practical medicine"*, Russian Federation, Moscow.

Donaire, L., Barajas, D., Martínez-García, B., Martínez-Priego, L., Pagán, I. and Llave, C. (2008) 'Structural and genetic requirements for the biogenesis of tobacco rattle virus-derived small interfering RNAs', *J. Virol*., Vol. 82, pp.5167–5177.

Fujino, T., Navaratnam, N. and Scott, J. (1998) 'Human apolipoprotein B RNA editing deaminase gene (APOBEC1)', *Genomics*, Vol. 47, pp.266–275.

Gros, L., Saparbaev, M.K. and Laval, J. (2002) 'Enzymology of the repair of free radicals-induced DNA damage', *Oncogene*, Vol. 21, pp.8905–8925.

Hamada, M., Kiryu, H., Sato, K., Mituyama, T. and Asai, K. (2009) 'Prediction of RNA secondary structure using generalized centroid estimators', *Bioinformatics*, Vol. 25, pp.465–473.

Iyer, L.M., Koonin, E.V. and Aravind, L. (2003) 'Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases', *BMC Struct. Biol.*, Vol. 3, p.1.

Katow, S. and Matsuno, T. (1980) 'Base composition of Rubella virus RNA', *Arch. Virol*., Vol. 65, pp.67–70.

Khrustalev, V.V. and Barkovsky, E.V. (2008) 'An in-silico study of alphaherpesviruses ICP0 genes: positive selection or strong mutational GC-pressure?', *IUBMB Life*, Vol. 60, pp.456–460.

Khrustalev, V.V. and Barkovsky, E.V. (2009) 'Mutational pressure is a cause of inter- and intragenomic differences in GC-content of simplex and varicello viruses', *Comput. Biol. and Chem.*, Vol. 33, pp.295–302.

Khrustalev, V.V. and Barkovsky, E.V. (2010a) 'Study of completed archaeal genomes and proteomes: hypothesis of strong mutational AT pressure existed in their common predecessor', *Genomics Proteomics Bioinformatics*, Vol. 8, pp.22–32.

Khrustalev, V.V. and Barkovsky, E.V. (2010b) '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%', *J. Theor. Biol.*, Vol. 266, pp.88–98.

Khrustalev, V.V. (2009) 'HIV1 V3 loop hypermutability is enhanced by the guanine usage bias in the part of env gene coding for it', *Silico Biol*., Vol. 9, p.0022.

Lee, J.Y. and Bowden, D.S. (2000) 'Rubella virus replication and links to teratogenicity', *Clin. Microbiol. Rev.*, Vol. 13, pp.571–587.

Lee, R.M., Hirano, K., Anant, S., Baunoch, D. and Davidson, N.O. (1998) 'An alternatively spliced form of apobec-1 messenger RNA is overexpressed in human colon cancer', *Gastroenterology*, Vol. 115, pp.1096–1103.

Lewis, B.P., Burge, C.B. and Bartel, D.P. (2005) 'Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are MicroRNA targets', *Cell*, Vol. 120, pp.15–20.

Liuharles, Y., George, C.X., Patterson, J.B. and Samuel, C.E. (1997) 'Functionally distinct double-stranded RNA-binding domains associated with alternative splice site variants of the interferon-inducible double-stranded RNA-specific adenosine deaminase', *J. Biol. Chem*., Vol. 272, pp.4419–4428.

Liuharles, Y. and Samuel, C.E. (1996) 'Mechanism of interferon action: functionally distinct RNA binding and catalytic domains in the interferon-inducible, double-stranded RNA-specific adenosine deaminase', *J. Virol*., Vol. 70, pp.1961–1968.

Maas, S., Rich, A. and Nishikura, K. (2003) **'**A-to-I RNA editing: recent news and residual mysteries', *J. Biol. Chem*., Vol. 278, pp.1391–1394.

Moissiard, G. and Voinnet, O. (2006) 'RNA silencing of host transcripts by cauliflower mosaic virus requires coordinated action of the four Arabidopsis Dicer-like proteins', *Proc. Natl. Acad. Sci., USA*, Vol. 103, pp.19593–19598.

Mukhopadhyay, D., Anant, S., Lee, R.M., Kennedy, S., Viskochil, D. and Davidson, N.O. (2002) 'C→U editing of neurofibromatosis 1 mRNA occurs in tumors that express both the type II transcript and apobec-1, the catalytic subunit of the Apolipoprotein B mRNA–editing enzyme', *Am. J. Hum. Genet*., Vol. 70, pp.38–50.

Narvaiza, I., Linfesty, D.C., Greener, B.N., Hakata, Y., Pintel, D.J., Logue, E., Landau, N.R. and Weitzman, M.D. (2009) 'Deaminase-independent inhibition of parvoviruses by the APOBEC3A cytidine deaminase', *PLoS Pathog*, Vol. 5, p.e1000439.

Pauli, E.K., Schmolke, M., Hofmann, H., Ehrhardt, C., Flory, E., Münk, C. and Ludwig, S. (2009) 'High level expression of the anti-retroviral protein APOBEC3G is induced by influenza A virus but does not confer antiviral activity', *Retrovirology*, Vol. 6, p.38.

Petit, V., Guétard, D., Renard, M., Keriel, A., Sitbon, M., Wain-Hobson, S. and Vartanian, J.P. (2009) 'Murine APOBEC1 is a powerful mutator of retroviral and cellular RNA in vitro and in vivo', *J. Mol. Biol*., Vol. 385, pp.65–78.

Strauss, J.H. and Strauss, E.G. (1994) 'The alphaviruses: gene expression, replication, and evolution', *Microbiol. Rev*., Vol. 58, pp.491–562.

Sueoka, N. (1988) 'Directional mutation pressure and neutral molecular evolution', *Proc. Natl. Acad. Sci. USA*, Vol. 85, pp.2653–2657.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) 'MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0', *Mol. Biol. Evol*., Vol. 24, pp.1596–1599.

Tzeng, W.P. and Frey, T.K. (2005) 'Rubella virus capsid protein modulation of viral genomic and subgenomic RNA synthesis', *Virol.*, Vol. 337, pp.327–334.

Vartanian, J.P., Henry, M., Marchio, A., Suspène, R., Aynaud, M.M., Guétard, D., Cervantes-Gonzalez, M., Battiston, C., Mazzaferro, V., Pineau, P., Dejean, A. and Wain-Hobson, S. (2010) 'Massive APOBEC3 editing of Hepatitis B Viral DNA in Cirrhosis', *PLoS Pathog*, Vol. 6, p.e1000928.

Xue, C., Li, F., He, T., Liu, G.P., Li, Y. and Zhang, X. (2005) 'Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine', *BMC Bioinformatics*, Vol. 6, p.310.