(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Journal of Theoretical Biology 282 (2011) 71–79

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Percent of highly immunogenic amino acid residues forming B-cell epitopes is higher in homologous proteins encoded by GC-rich genes

Vladislav V. Khrustalev*, Eugene V. Barkovsky

Department of General Chemistry, Belarussian State Medical University, Belarus, Minsk 220022, Dzerzinskogo 83, Belarus

ARTICLE INFO

Article history:
Received 7 February 2011
Received in revised form
6 May 2011
Accepted 9 May 2011
Available online 23 May 2011

Keywords:
GC-content
Amino acid replacements
B-cell epitopes
Simplex virus
Varicello virus

ABSTRACT

We analyzed the dependence of the percent of highly immunogenic amino acid residues included in B-cell epitopes of homologous proteins on the GC-content (G+C) of genes coding for them in twenty-seven lineages of proteins (and subsequent genes), which belong to seven Varicello and five Simplex viruses. We found out that proteins encoded by genes of a high GC-content usually contain more targets for humoral immune response than their homologs encoded by GC-poor genes. This tendency is characteristic not only to the lineages of glycoproteins, which are the main targets for humoral immune response against Simplex and Varicello viruses, but also to the lineages of capsid proteins and even "housekeeping" enzymes. The percent of amino acids included in linear B-cell epitopes has been predicted for 324 proteins by BepiPred algorithm (www.cbs.dtu.dk/services/BepiPred), the percent of highly immunogenic amino acids included in discontinuous B-cell epitopes and the percent of exposed amino acid residues have been predicted by Epitopia algorithm (http://epitopia.tau.ac.il/). Immunological consequences of the directional mutational GC-pressure are mostly due to the decrease in the total usage of highly hydrophobic amino acids and due to the increase in proline and glycine levels of usage in proteins. The weaker the negative selection on amino acid substitutions caused by symmetric mutational pressure, the higher the slope of direct dependence of the percent of highly immunogenic amino acids included in B-cell epitopes on G+C.

1. Introduction

Symmetric directional mutational pressure is a situation when nucleotide mutations of AT to GC direction occur with rates unequal to the rates of GC to AT mutations (Sueoka, 1988). Mutations caused by certain biochemical processes forming mutational pressure (either AT-pressure or GC-pressure) may be fixed in case if they are neutral or positive (Sueoka, 1988). Negative selection eliminates those mutations in coding regions, which significantly decrease the fitness of an organism (radical mutations) (Kimura, 1983). In other words, if amino acid replacement in a certain site of a given protein leads to that kind of change in structure and function of this protein, which in turn decreases the fitness of an organism, it will be eliminated from the population by negative selection (Kimura, 1983). Some amino acids encoded by GC-poor codons can be substituted with amino acids encoded by GC-rich codons and vice versa without any significant changes in function and structure of a protein (Khrustalev and Barkovsky, 2009a). Although mutational pressure causes both neutral and radical amino acid mutations, neutral amino acid substitutions are fixed more frequently than radical substitutions, because most of radical mutations are eliminated from population by negative selection (Khrustalev and Barkovsky, 2009a).

Our recent studies (Khrustalev, 2009, 2010) showed that along with many other effects such as changes in structure and so in stability of mRNAs (Cristillo et al., 2001), simplification of amino acid content (Khrustalev and Barkovsky, 2009a), changes in probability of synonymous mutation occurrence (Khrustalev and Barkovsky, 2009b), etc., directional mutational pressure should bring immunological consequences. Results obtained in sessions of "in-silico directed mutagenesis" (Khrustalev, 2009, 2010) performed with the help of BepiPred algorithm (Larsen et al., 2006) helped us discover the following tendencies. Nonsynonymous nucleotide substitutions of AT to GC direction should cause elongation of previously existing linear B-cell epitopes and formation of new epitopes at a high probability (Khrustalev, 2009). For two long epitope free regions of Herpes simplex virus type 1 glycoprotein B (predicted by BepiPred) the probability of elongation or appearance of a new epitope due to a single AT to GC nonsynonymous mutation is equal to 25% (Khrustalev, 2009). Single nonsynonymous nucleotide mutation of an opposite (GC to AT) direction causes partial or complete "destruction" of linear

* Corresponding author. Belarussian State Medical University, Department of General Chemistry, Communisticheskaya 7-24, 220029 Minsk, Belarus.
Tel.: +375 80172845957.
E-mail address: vvkhrustalev@mail.ru (V.V. Khrustalev).

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.05.018

B-cell epitopes in full-length gp120 of human immunodeficiency virus type 1 at even higher probability (35%) (Khrustalev, 2010).

In the present work we tested a hypothesis that some of those types of amino acid replacements caused by directional mutational pressure, which influence antigenic properties of proteins (increase or decrease the number and length of B-cell epitopes), have undergone fixation in populations of Simplex and Varicello viruses.

It was shown that the percent of pathogens is higher among bacteria with GC-poor genomes than in bacteria with GC-rich genomes (Rocha and Danchin, 2002). Indeed, according to our hypothesis, mutations of GC to AT direction should lead to frequent destruction of B-cell epitopes (Khrustalev, 2010) and help those GC-poor pathogenic bacteria escape humoral immune answer. However, there are, of course, exceptions of this tendency. Average GC-content of 1223 viruses has been shown to be equal to 44.0%, average GC-content of 16 pathogenic intracellular bacteria has been shown to be equal to 44.1%, while average GC-content of 48 non-pathogenic bacteria (53.4%) has been shown to be significantly higher than two previous ones (Calis et al., 2010). Coming back to viruses we have to admit that the term "GC-content" is not entirely suitable for any singlestranded viral genome, because the level of C inside it is rarely equal to the level of G (Khrustalev and Barkovsky, 2011a).

Among 11 HLA-A molecules studied by Calis et al. (2010), 9 showed significant preference for peptides from proteins encoded by GC-poor genes. This finding has been interpreted as an evidence of T-cell immune answer adaptation to proteins of intracellular pathogens, which are usually encoded by GC-poor genes (Calis et al., 2010). On the other hand, among 10 HLA-B molecules studied, 4 showed significant preference for peptides from proteins encoded by GC-rich genes and only 1 for peptides from proteins encoded by GC-poor genes (Calis et al., 2010).

Each protein is composed of hydrophobic core and relatively hydrophilic surface, which is in contact with solvent (Kyte and Doolittle, 1982). Antibodies can recognize only those parts of a native protein that are exposed on its surface (Rubinstein et al., 2008). Theoretically, antibodies may be synthesized against the "inner" parts of a protein in case if it undergoes denaturation, cleavage or fragmentation. However, these antibodies will never bind to the native protein. So, the main criterion for distinguishing epitopes from the remaining part of a protein is the solvent accessibility (Kyte and Doolittle, 1982; Rubinstein et al., 2008, 2009).

B-cell epitopes are not distributed equally on the surface of a protein (Rubinstein et al., 2008). They show a general preference for charged and polar amino acids (Rubinstein et al., 2008). These amino acid residues are hydrophilic (Kyte and Doolittle, 1982) and so they should be in contact with water molecules in case if they are not surrounded by strongly hydrophobic amino acids. Epitopes were also found to be significantly enriched with tyrosine and tryptophan (Rubinstein et al., 2008). These hydrophobic amino acid residues become highly immunogenic in case if they are surrounded by acrophilic amino acid residues (by those amino acid residues that are prone to be situated on a surface of protein globules) (Hopp and Woods, 1983).

It is known that protruding parts of a protein usually presented by beta-turns (Chou and Fasman, 1978) serve as good targets for antibodies (Rubinstein et al., 2008). Beta-turns are usually formed by proline and glycine (Chou and Fasman, 1978). The higher the usage of glycine and proline, the more the formation of beta-turns.

Epitopes were shown to be enriched with unorganized secondary structure elements that render them flexible (Rubinstein et al., 2008). The most flexible amino acid residue is glycine since it has no side chain (Chou and Fasman, 1978).

Simplex and Varicello viruses belong to Alphaherpesvirinae subfamily of Herpesviridae family (Khrustalev and Barkovsky, 2009b). Genomes of these viruses are presented by doublestranded DNA. They contain about 75 open reading frames. Most of the open reading frames found in completely sequenced genomes of Simplex and Varicello viruses have homologs in the properly studied genome of human simplex virus type 1. Although Simplex and Varicello viruses are close relatives, GC-content in their genomes varies greatly (Khrustalev and Barkovsky, 2009b). In general, changes in GC-content of genes and genomes of alphaherpesviruses are caused by the directional mutational pressure (Khrustalev and Barkovsky, 2009b). These variations in nucleotide content of coding regions have a significant impact on amino acid content of viral proteins (Khrustalev and Barkovsky, 2009a). Possible biochemical causes of mutational pressure in genomes of different alphaherpesviruses have been discussed in our previous papers (Khrustalev and Barkovsky, 2009b).

In our opinion, groups of closely related but different in the respect of GC-content genes form a perfect material for the present work.

With the help of two different methods (Rubinstein et al., 2009; Larsen et al., 2006) we demonstrated that such immunological features of proteins such as (i) the percent of amino acids included in linear B-cell epitopes, (ii) the percent of highly immunogenic amino acid residues, (iii) the percent of amino acids included in epitopes at a high probability, (iv) the percent of exposed (and buried) amino acid residues and (v) the percent of amino acids in 5 highly immunogenic stretches frequently show a direct linear dependence on GC-content of viral genes coding for glycoproteins, capsid proteins and even "housekeeping" enzymes.

2. Materials and methods

As a material we used GenBank records describing completely sequenced genomes of five simplex viruses: macacine herpesvirus 1 (MaHV1) [**NC_004812**], cercopithecine herpesvirus 2 (CeHV2) [**NC_006560**], papiine herpesvirus 2 (PaHV2) [**NC_007653**], human herpesvirus 1 (HSV1) [**NC_001806**], human herpesvirus 2 (HSV2) [**NC_001798**]; seven varicello viruses: 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**].

There are nine glycoproteins that have not been lost in any completely sequenced genome of Simplex or Varicello virus. Names and locations in unique long (UL) and unique short (US) regions of HSV1 genome of genes coding for studied glycoproteins are written below: glycoprotein L (UL1), glycoprotein M (UL10), glycoprotein H (UL22), glycoprotein B (UL27), glycoprotein C (UL44), glycoprotein N (UL49A), glycoprotein K (UL53), glycoprotein I (US7) and glycoprotein E (US8).

We also studied nine capsid and tegument proteins encoded by all completely sequenced genomes of Simplex and Varicello viruses as well as nine nonstructural enzymes. Names and locations in UL of HSV1 genome of genes coding for studied capsid and tegument proteins are listed below: capsid portal protein (UL6), DNA packaging tegument protein (UL17), capsid triplex subunit 2 (UL18), major capsid protein (UL19), DNA packaging capsid associated tegument protein (UL25), small capsid protein (UL35), large tegument protein (UL36), tegument protein (UL37) and capsid triplex subunit 1 (UL38).

Names and locations in UL of HSV1 genome of genes coding for nonstructural (housekeeping) viral enzymes used in this study are uracil-DNA glycosylase (UL2), DNA replication origin-binding helicase (UL9), deoxyribonuclease (UL12), thymidine kinase (UL23), singlestranded DNA-binding protein (UL29), DNA polymerase catalytic subunit (UL30), ribonucleotide reductase

subunit 1 (UL39), DNA polymerase processivity subunit (UL42) and deoxyuridine triphosphatase (UL50).

The percent of amino acids situated in linear B-cell epitopes has been calculated for every studied protein. Linear B-cell epitopes have been predicted by the special software BepiPred (www.cbs.dtu.dk/services/BepiPred) (Larsen et al., 2006). Experimental works on the mapping of epitopes in HSV and VZV glycoproteins are concentrated mostly on the discovery of type-specific epitopes, which can be used in diagnostics (Tunbäck et al., 2000), or on the discovery of neutralizing epitopes. In this work we have not concentrated on certain epitopes, but calculated the percent of amino acids in all linear B-cell epitopes mapped in each protein.

Percent of amino acids exposed to a solvent has been calculated by Epitopia server (Rubinstein et al., 2009). Making a decision on whether the given amino acid is exposed to a solvent or buried Epitopia server predicts a tertiary structure using the PredictProtein Program (Rost et al., 2004) and then solvent accessibility is calculated by the Surface Racer Program (Tsodikov et al., 2002).

As to another index calculated by Epitopia, we decided to estimate a threshold for highly immunogenic amino acid residues at the point of −20.00. It means that the percent of highly immunogenic amino acid residues designated as "*I*" in the figures is the percent of amino acid residues with immunogenicity score higher than −20.00 according to the Epitopia prediction. A threshold for the probability score also provided by Epitopia was estimated at the point of 0.05. So, the percent of amino acid residues with probability score higher than 0.05 is represented as the letter "*P*" in the figures.

Epitopia also determines five most immunogenic amino acid stretches in the protein (Rubinstein et al., 2009). Maximal length of each stretch is 29 amino acid residues. The percent of amino acids included in these five stretches is represented in the figures as the letter "*S*".

Total GC-content and GC-content in first (1GC), second (2GC) and third (3GC) codon positions in each coding region and amino acid usage in each protein have been calculated by our original MS Excel spreadsheet "VVK in group" (www.barkovsky.hotmail.ru) (Khrustalev and Barkovsky, 2010). This spreadsheet calculates all these indexes right after the pasting of nucleotide sequences in designated cells on its "sequences" list. Coefficients of correlation (*R*) and slopes of the linear dependences have been calculated with the help of MS Excel functions.

3. Results

3.1. Estimation of the general level of neutrality for the evolution of a protein under the influence of symmetric mutational pressure according to the method proposed by Sueoka

According to the work of Sueoka (1988), one can estimate a level of neutrality for the evolution of the lineage of homologous genes using their levels of GC-content in three codon positions. He decided to build a graph with levels of GC-content in third codon positions situated on *X*-axis and average levels of GC-content in first and second codon positions situated on the *Y*-axis. The example of this graph made for the genes coding for capsid portal protein (UL6) can be seen in Fig. 1.

Most of the nucleotide mutations in third codon positions are synonymous, while all nucleotide mutations in second codon positions and most of them in first codon positions are nonsynonymous (Kimura, 1983; Sueoka, 1988). In general, mutations in third codon positions should be fixed faster than mutations in first and second codon positions (Kimura, 1983; Sueoka, 1988). Sueoka (1988) calculated GC-content in first, second and third codon positions of 56 from 64 codons (three stop-codons, three codons coding for isoleucine, single codon coding for tryptophan and single codon coding for methionine were excluded from calculations). Those corrected indexes are usually abbreviated as "GC1", "GC2" and "GC3" or "P_1", "P_2" and "P_3". In our work we prefer to calculate GC-content in all the 64 codons. This is why we always use similar but not identical abbreviations ("1GC", "2GC" and "3GC") to highlight the difference (Khrustalev and Barkovsky, 2009b, 2011b). For corrections, for the discrimination between biases in the rates of transversions and transitions, as well as for the detection of "asymmetric" negative selection the set of other indexes describing GC-content in fourfold and twofold degenerated sites situated in third codon positions has been proposed (Khrustalev and Barkovsky, 2010).

The higher is the slope of the dependence between 3GC and "(1GC+2GC)/2" in a lineage of homologous genes (see Fig. 1), the higher is the level of neutrality for the fixation of amino acid substitutions caused by mutational pressure (Sueoka, 1988).

The ratio between of length of the bold line in Fig. 1 to the length of the *Y*-axis (the slope of the trend of linear dependence) is the level of neutrality (Sueoka, 1988). The length of the dotted line reflects the strength of negative selection on amino acid

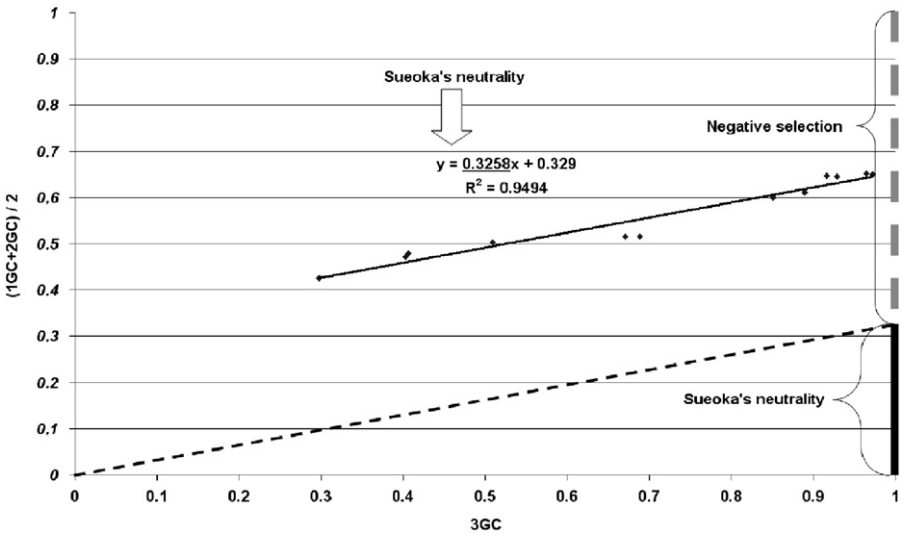

Fig. 1. Dependence of the average GC-content in first and second codon positions ((1GC+2GC)/2) on the GC-content in third codon positions (3GC) in genes coding for capsid portal protein (UL6) from 5 Simplex and 7 Varicello viruses.

Table 1
Slopes of the dependences of (i) the percent of amino acid residues included in linear B-cell epitopes (*L*), (ii) the percent of exposed (*E*) and buried (*B*) amino acid residues, (iii) the percent of amino acid residues with immunogenicity score higher than −20.00 (*I*) and probability score higher than 0.05 (*P*), as well as (iv) of the percent of amino acid residues included in five most immunogenic stretches (*S*) in each of 27 lineages of proteins from 5 Simplex and 7 Varicello viruses on the GC-content (G+C) of genes coding for them. Levels of Sueoka's neutrality (*N*) are given for each lineage too. Coefficients of correlation of "*L*", "*E*", "*B*", "*I*", "*P*" and "*S*" on "*N*" are provided.

Name of the protein	*N*	*L*	*E*	*B*	*I*	*P*	*S*
UL19	0.1827	0.1229	0.0191	−0.0191	0.1077	0.1483	0.0300
UL35	0.1968	0.2418	0.1140	−0.1140	0.1589	0.2471	−0.1390
UL18	0.2433	0.2232	0.0202	−0.0202	0.0389	−0.0234	0.0341
UL25	0.2944	0.3759	0.0382	−0.0382	0.2058	0.1389	0.0718
UL38	0.3082	0.3948	0.1156	−0.1156	0.2622	0.2429	0.0812
UL6	0.3258	0.2995	0.1000	−0.1000	0.3177	0.2856	0.1181
UL37	0.3345	0.5141	0.1776	−0.1776	0.2698	0.2905	0.0822
UL17	0.3730	0.4348	0.0189	−0.0190	0.3132	0.2906	0.0120
UL36	0.4124	0.4872	0.1543	−0.1543	0.2845	0.2687	0.0095
gB	0.1915	0.1663	−0.0015	0.0015	0.1623	0.1491	0.0232
gK	0.3206	0.1874	0.0413	−0.0413	0.1605	0.0995	0.1745
gM	0.3604	0.3523	0.0798	−0.0798	0.2557	0.3431	−0.1142
gN	0.3926	0.8678	0.4311	−0.4311	0.3479	0.5243	0.2715
gE	0.3929	0.5153	0.0483	−0.0483	0.3007	0.3695	0.0720
gH	0.3968	0.4735	0.1490	−0.1490	0.2422	0.2568	0.0154
gC	0.4029	0.4387	0.0501	−0.0501	0.2592	0.2362	0.0193
gL	0.4270	0.6115	0.2000	−0.2000	0.2051	0.3762	0.4013
gI	0.4601	0.4785	0.0584	−0.0584	0.2767	0.2974	−0.0042
UL29	0.1844	0.0627	0.0464	−0.0464	0.1068	0.0755	0.0256
UL30	0.2617	0.2646	0.0390	−0.0390	0.2359	0.1297	−0.0282
UL39	0.2347	0.4365	0.0787	−0.0787	0.1367	0.0497	0.0059
UL9	0.2749	0.3642	0.0674	−0.0674	0.2456	0.2982	−0.0317
UL2	0.3080	0.4120	0.0655	−0.0655	0.2698	0.2571	0.0126
UL23	0.3530	0.6603	0.0376	−0.0376	0.2896	0.3336	0.1140
UL12	0.3742	0.5713	0.1770	−0.1770	0.2501	0.2111	−0.0185
UL42	0.3885	0.6295	0.1803	−0.1803	0.2065	0.2143	0.1231
UL50	0.4213	0.6101	0.0493	−0.0493	0.2155	0.1695	0.1216
Correlation with *N*		0.7662	0.3917	−0.3917	0.6740	0.5951	0.3758
Correlation with *N* for groups with strong dependence of a given index on G+C		0.7180	0.5946	−0.5947	0.6774	0.5401	0.6293
% of groups with strong dependence on G+C		92.59	59.26	59.26	96.30	92.59	40.74

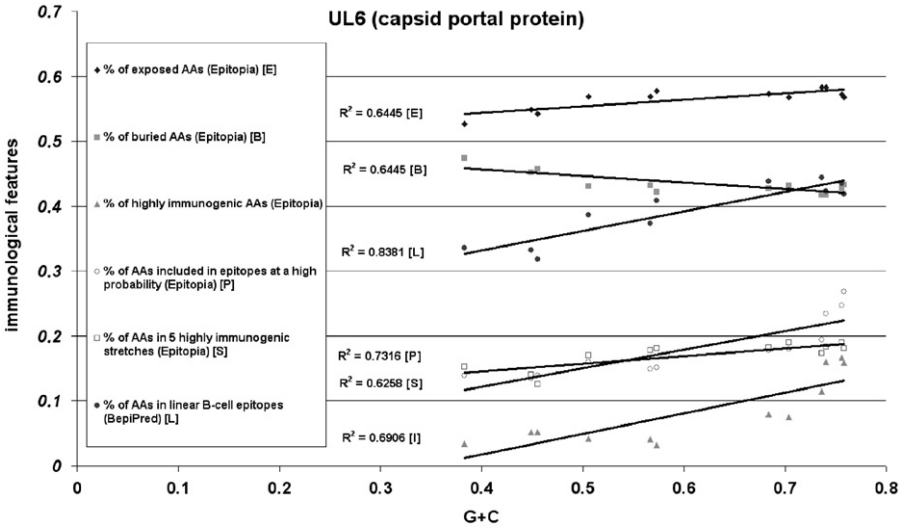

Fig. 2. Dependences of (i) the percent of amino acid residues included in linear B-cell epitopes (*L*), (ii) the percent of exposed (*E*) and buried (*B*) amino acid residues, (iii) the percent of amino acid residues with immunogenicity score higher than −20.00 (*I*) and probability score higher than 0.05 (*P*), as well as (iv) of the percent of amino acid residues included in five most immunogenic stretches (*S*) in capsid portal protein (UL6) from 5 Simplex and 7 Varicello viruses on the GC-content (G+C) of genes coding for it. Level of "*L*" has been calculated according to BepiPred prediction; levels of "*E*", "*B*", "*I*", "*P*" and "*S*" have been calculated according to the Epitopia results.

replacements. As one can see in Fig. 1, the level of neutrality for capsid portal protein evolution under the influence of symmetric mutational pressure is equal to 0.4601 (46.01%).

Levels of neutrality and slopes of the dependences between G+C and immunological features have been calculated for every lineage of homologous genes (see Table 1). We also calculated slopes of the dependences between every amino acid usage and GC-content of homologous genes.

In the figures one can see the levels of "R^2". "R^2" is the square of the coefficient of correlation (R). The level of R^2 lower than 0.25

is the evidence that the strength of the dependence between two variables is low or that there is no dependence at all. The level of R for the dependence between 3GC and "(1GC+2GC)/2" is higher than 0.5 in all 27 cases.

Differences between average levels of neutrality (N) for lineages of glycoproteins, capsid proteins and housekeeping enzymes are not significant according to the results of t-test. Indeed, variations in levels of neutrality inside each of the three groups of lineages are wide (see Table 1). That is why we cannot state that all the glycoproteins are under the influence of weaker (or stronger) negative selection than all the capsid proteins or housekeeping enzymes.

There are no significant differences (according to the results of t-test) in average levels of all the immunological features represented in Table 1 (L, E, B, I, P, S) between lineages of glycoproteins, capsid proteins and housekeeping genes. It means that we can study all the 27 lineages together as a common group.

3.2. Viruses with GC-rich genomes encode proteins which contain more linear and discontinuous antigenic determinants than their homologs from viruses with GC-poor genomes

As one can see in Fig. 2, the higher is the GC-content of a gene coding for capsid portal protein, the higher is the percent of amino acids included in linear B-cell epitopes of this protein (L), the higher is the percent of highly immunogenic amino acids (I), the percent of amino acids with high probability score (P) and the percent of those included in five highly immunogenic stretches (S). The percent of exposed amino acid residues (E) also shows direct linear dependence on G+C (see Fig. 2).

In Fig. 3 we placed the average level of GC-content in 27 genes for each of the 12 viruses studied on the X-axis and the average levels of several immunological characteristics on the Y-axis. As one can see in Fig. 3, there is a strong direct linear dependence of the average percent of exposed amino acid residues (E) in proteins on the average GC-content of genes. The average percent of buried amino acid residues (B) in proteins shows inversed linear dependence on G+C.

The main result of the present work is as follows: viruses with GC-rich genomes encode proteins that contain more linear antigenic determinants (L), more highly immunogenic amino acid residues (I) and residues with high score of the probability to be included in epitope (P) than their homologs from viruses with GC-poor genomes (see Fig. 3). Even the percent of amino acid residues included in five highly immunogenic stretches (S) is higher for the proteins encoded by GC-rich genes than for their homologs from GC-poor genomes (see Fig. 3).

3.3. Strong negative selection on amino acid substitutions occurring due to mutational pressure makes the direct dependence of the antigenicity of a protein on G+C of a gene weaker

In sixteen from twenty-seven (~60%) groups of homologous proteins the percent of exposed amino acids shows strong correlation ($R > 0.5$) with G+C of subsequent genes (see Table 1). The slope of this dependence is not constant for all these sixteen lineages. For the rest of the lineages the coefficient of correlation between E and G+C is low ($R < 0.5$).

Even so the percent of exposed amino acid residues strongly depends on G+C only in 60% of lineages, the percent of amino acids included in linear B-cell epitopes (L) and the percent of amino acids with high score of the probability to be included in epitope (P) show strong linear dependence on G+C in 25 from 27 lineages studied (see Table 1). The percent of highly immunogenic amino acid residues (I) shows no dependence on G+C only in a single lineage.

The percent of amino acid residues in five most immunogenic amino acid stretches shows linear dependence on G+C only in 11 from 27 lineages, probably, due to the limit of the length of each stretch (it cannot be higher than 29 amino acid residues) (Rubinstein et al., 2009).

The slope of the dependence of the percent of exposed amino acid residues (E) on G+C is higher in lineages with high level of Sueoka's neutrality (N). The correlation between E and N is strong in those lineages in which E shows strong correlation with G+C (see Fig. 4). In other words, the percent of exposed amino acid residues grows under the influence of GC-pressure and decreases under the influence of AT-pressure in proteins evolving under the weak negative selection.

Slopes of the dependences on G+C of (i) the percent of amino acids included in linear B-cell epitopes (L), (ii) the percent of

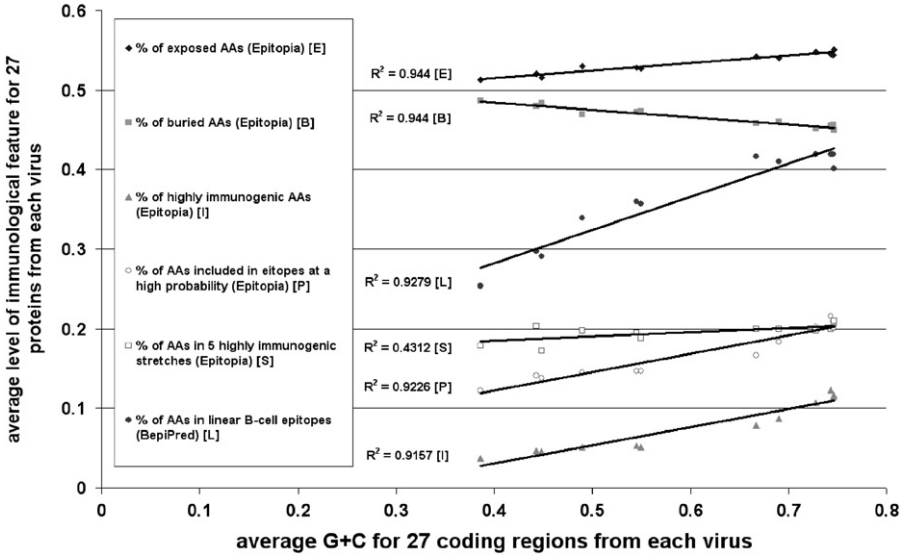

Fig. 3. Dependences of (i) the average percent of amino acid residues included in linear B-cell epitopes (L), (ii) the average percent of exposed (E) and buried (B) amino acid residues, (iii) the average percent of amino acid residues with immunogenicity score higher than −20.00 (I) and probability score higher than 0.05 (P), as well as (iv) of the average percent of amino acid residues included in five most immunogenic stretches (S) in 27 proteins from 5 Simplex and 7 Varicello viruses on the average GC-content (G+C) of genes coding for them. Level of "L" has been calculated according to BepiPred prediction; levels of "E", "B", "I", "P" and "S" have been calculated according to the Epitopia results.

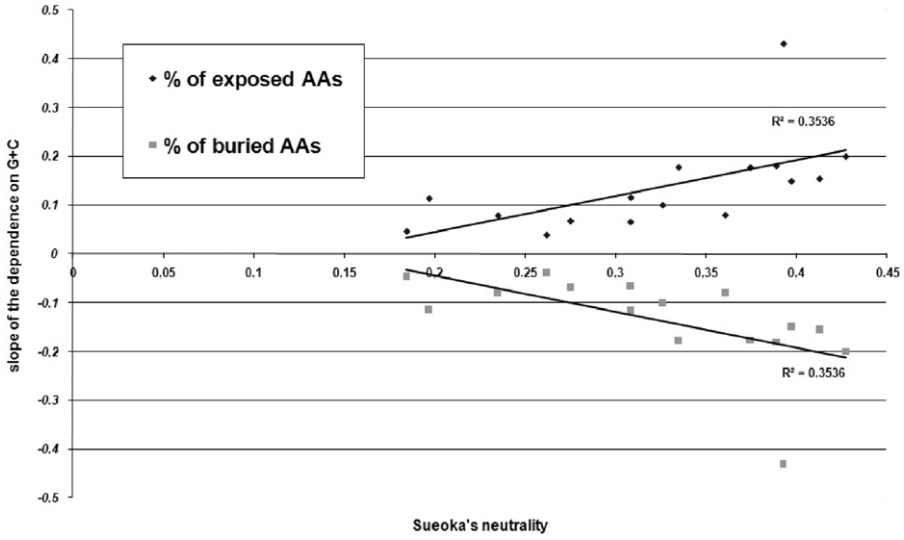

Fig. 4. Dependence of the slope of the dependence of the percent of exposed and buried amino acid residues and G+C on the level of Sueoka's neutrality for 16 from 27 lineages of viral proteins (and subsequent genes). For these 16 lineages the coefficient of correlation between the percent of exposed amino acid residues and G+C is higher than 0.5.

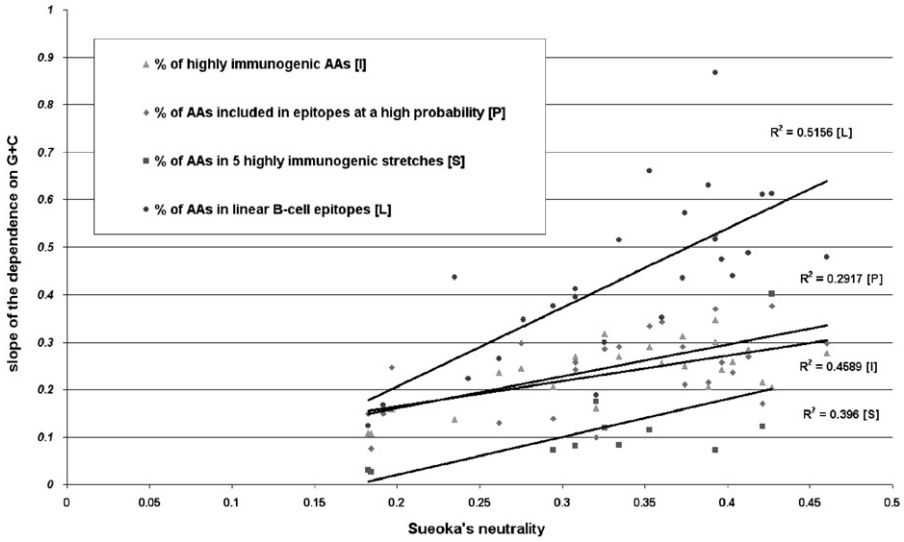

Fig. 5. Dependences of the slopes of the dependences between (i) the percent of amino acid residues included in linear B-cell epitopes (*L*), (ii) the percent of amino acid residues with immunogenicity score higher than −20.00 (*I*) and probability score higher than 0.05 (*P*), as well as (iv) of the percent of amino acid residues included in five most immunogenic stretches (*S*) and G+C on the level of Sueoka's neutrality for those lineages in which coefficients of correlation between "*L*", "*I*", "*P*" and "*S*" on G+C are higher than 0.5.

highly immunogenic amino acid residues (*I*) and (iii) the percent of amino acids situated in epitopes at a high probability (*P*) show strong correlation with the level of Sueoka's neutrality (see Fig. 5 and Table 1).

In general, mutational GC-pressure seems to increase (i) the percent of exposed amino acid residues and (ii) the percent of amino acids included in continuous and discontinuous B-cell epitopes. The latter effect is usually present even in those lineages in which the percent of exposed amino acid residues does not correlate with G+C.

4. Discussion

There is no need for alphaherpesviruses to escape immune answer of their hosts to continue their lifecycles (Kaufman et al., 2005). Genomes of these viruses exist in latency during the host's lifetime (Kaufman et al., 2005; Liljeqvist et al., 2009). Certain amount of HSV1 virions produced due to reactivation is excreted by a host without any sign of clinically recognizable relapse (Scott et al., 1997; Kaufman et al., 2005), as well as certain amount of HSV2 virions (Miller and Danaher, 2008). This amount is thought to be sufficient to cause primary infection in nonimmune organisms (Kaufman et al., 2005). Cases of asymptomatic reactivation and viral shed of Varicella-zoster (HHV3) have been recently approved in astronauts (Cohrs et al., 2008).

On the other hand, GC-pressure in genomes of Simplex viruses infecting human and other primates may sometimes lead to the occurrence of new targets for immune answer (Khrustalev, 2009). In case if this occurrence is associated with the destruction or the modification of previously existed epitopes, immune escaping may happen (Khrustalev, 2009). Although this immune escaping is not necessary for persistence and recirculation of the virus, it may cause clinically recognizable relapse. Antibodies of IgG class against HSV1 antigens can be found in infected persons during latency (Kuhn et al., 1987). However, during the relapse of HSV1 infection IgM molecules are synthesized in 68% of persons (Hashido and Kawana, 1997). At least some part of those IgM molecules, which are detected during each relapse, may be

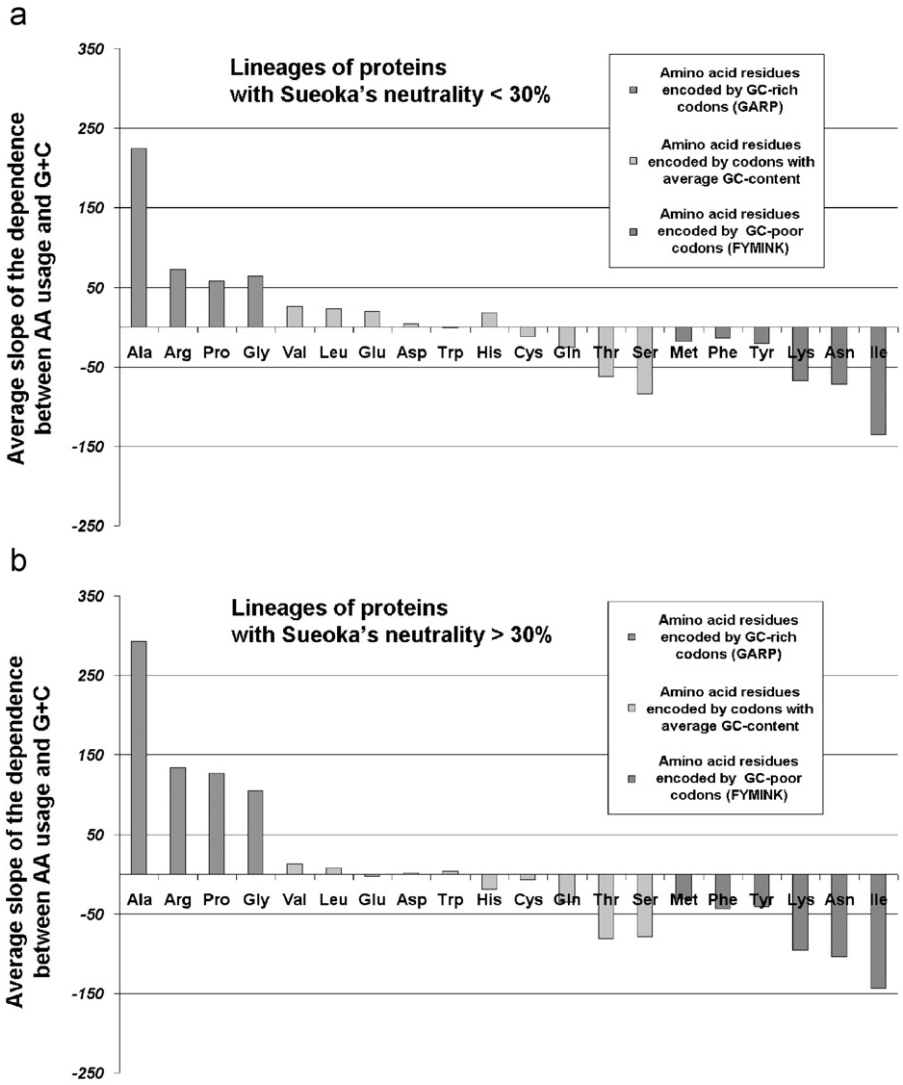

Fig. 6. Average slopes of the dependences of each amino acid usage on G+C for 9 lineages of viral proteins (and subsequent genes) from 5 Simplex and 7 Varicello viruses with Sueoka's neutrality levels lower than 30% (a) and for 18 lineages with Sueoka's neutrality levels higher than 30% (b).

synthesized to new antigenic determinants occurred due to mutations.

Virions of alphaherpesviruses are enveloped. Glycoproteins are situated on the surface of viral participles, while capsid proteins are situated under the envelope (Kuhn et al., 1987). However, antibodies against capsid proteins (including those against major capsid protein) are present in persons with acute herpetic infection (Kuhn et al., 1987).

It is hard to imagine that evolution of genes coding for such viral "housekeeping" enzymes as thymidine kinase, deoxyuridine triphosphatase, deoxyribonuclease or processivity subunit of viral DNA polymerase captured by common ancestor of alphaherpesviruses from the genome of its host is driven by the process of immune escaping, while their immunological features seem to show high slopes of the dependence on GC-content of genes coding for them.

In proteins that are under the influence of a weak negative selection the percent of amino acids included in epitopes grows more steeply under the influence of GC-pressure and falls more steeply under the influence of AT-pressure, in comparison with the proteins that are under the influence of a strong negative selection. We have analyzed Fig. 6 to understand the causes of this phenomenon.

As one can see in Fig. 6A and B, the highest slope is characteristic to the dependence of alanine usage on G+C (Khrustalev and Barkovsky, 2009a) in both lineages of proteins under the strong ($N<0.3$) and under the weak ($N>0.3$) negative selection. Slopes of the dependences on G+C for the levels of three other amino acids encoded by GC-rich codons (proline, glycine and arginine) are lower (Khrustalev and Barkovsky, 2009a). In lineages of the proteins that are under the influence of the strong negative selection slopes of proline, glycine and arginine levels (see Fig. 6a) are much lower than those slopes in lineages of the proteins evolving under the weak negative selection (see Fig. 6b).

As to the levels of amino acids encoded by GC-poor codons, three of them (isoleucine, lysine and asparagine) demonstrate slopes of the dependences on G+C comparable with those for proline, glycine and arginine (see Fig. 6), but in a different direction (Khrustalev and Barkovsky, 2009a). Slopes of the dependences on G+C for methionine, phenylalanine and tyrosine levels are not as steep as those for isoleucine, lysine and asparagine levels (Khrustalev and Barkovsky, 2010). Slopes of the dependences of FYMINK amino acids levels on G+C in lineages of proteins evolving under the weak negative selection (see Fig. 6B) are steeper than those in lineages of the proteins that are under the influence of the strong negative selection (see Fig. 6A).

Interestingly, levels of serine, threonine and glutamine demonstrate negative slopes of the dependence on G+C in both kinds of

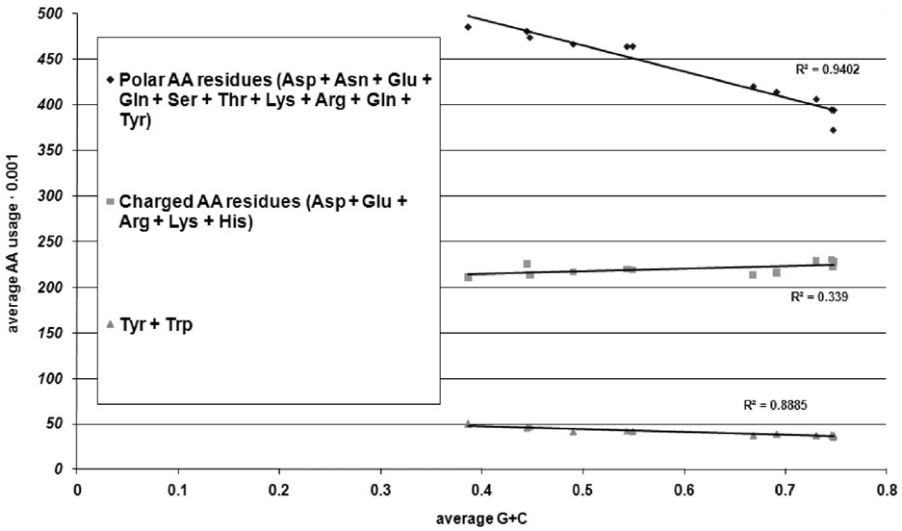

Fig. 7. Dependences (i) of the average level of polar amino acids usage and (ii) of the average level of charged amino acids usage and (iii) of the average level of tyrosine and tryptophan usage in 27 proteins from 5 Simplex and 7 Varicello viruses on the average GC-content (G+C) of genes coding for them.

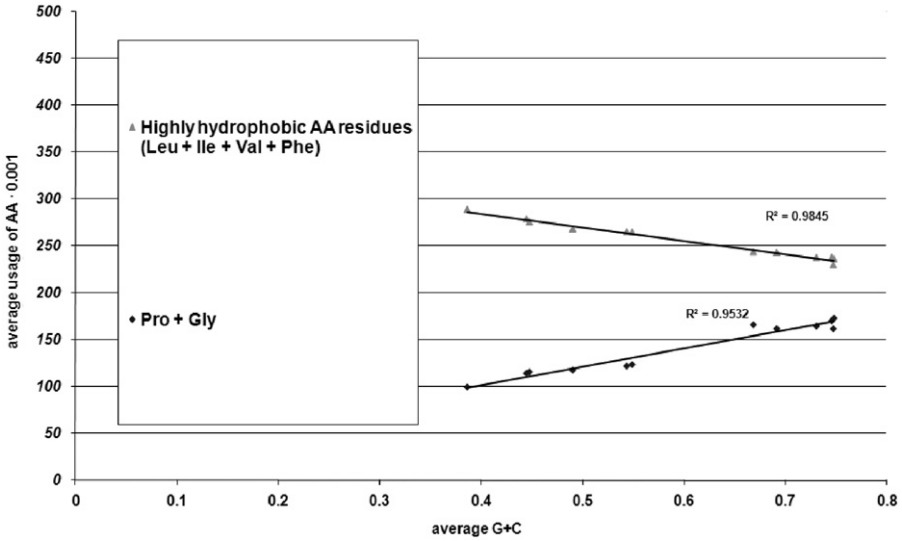

Fig. 8. Dependences (i) of the average level of highly hydrophobic amino acids usage and (ii) of the average level of proline and glycine usage in 27 proteins from 5 Simplex and 7 Varicello viruses on the average GC-content (G+C) of genes coding for them.

lineages (see Figs. 6A and B), although they are encoded by codons with average GC-content.

Which kinds of amino acid substitutions allowed by the weak negative selection result in the growth of the number and lengths of B-cell epitopes?

In Fig. 7 we showed that GC-pressure leads to the decrease in the usage of polar amino acid residues. The usage of charged amino acid residues does grow under the influence of GC-pressure but with practically unrecognizable slope (see Fig. 7). It means that the growth of the immunogenicity of proteins encoded by GC-rich genes cannot be caused by the growth of polar and charged (i.e. highly hydrophilic) amino acid residues.

The growth of the immunogenicity of proteins under the influence of GC-pressure cannot be caused by the increase in tyrosine and tryptophan usages too (see Fig. 7).

As the usage of highly hydrophobic amino acid residues decreases under the influence of GC-pressure and grows under the influence of AT-pressure (see Fig. 8), the percent of amino acids included in highly hydrophobic core of the protein (i.e. buried residues) should be lower in the proteins encoded by GC-rich genes. The sum of the levels of usage for highly hydrophobic amino acid residues shows inversed correlation with G+C mostly due to the changes in isoleucine usage (Khrustalev and Barkovsky, 2009a, 2010) (see Fig. 6).

Levels of proline and glycine usage show strong linear dependence on G+C because they both are encoded by GC-rich codons (Khrustalev and Barkovsky, 2009a; Singer and Hickey, 2000) (see Fig. 8). These amino acid residues usually form protruding parts of proteins (Chou and Fasman, 1978). Even though proline has nonpolar side chain and glycine has no side chain at all, they are usually situated on the surface of a protein (Hopp and Woods, 1983).

In our opinion, the growth of the percent of exposed amino acid residues as well as the growth of the percent of amino acids included in epitopes under the influence of GC-pressure is caused by the increase in proline and glycine levels of usage in GC-rich genes.

The level of the neutrality for amino acid replacements leading to proline and glycine appearance under the influence of GC-pressure and their disappearance under the influence of AT-pressure is the main factor responsible for the linear dependences given in Figs. 4 and 5. However, the strength of the negative selection for

proline and glycine appearance or disappearance in each protein under the influence of directional mutational pressure is primarily determined by structural and functional limitations.

GC-pressure is responsible for the "externalization" of certain parts of a protein, which have been "buried" in its hydrophobic core. In general, this GC-pressure-associated "externalization" happens due to substitutions of acrophobic amino acid residues (those that are usually "buried" in hydrophobic core) with acrophilic (those that are usually exposed on a surface), but not hydrophilic ones.

Even if the "externalization" has not happened in the proteins encoded by GC-rich genes, the percent of amino acids included in linear B-cell epitopes and 3D epitopes is usually higher in them than in their homologs encoded by GC-poor genes. This effect is also caused by the increase in the number of protruding parts of a protein (containing proline and/or glycine) under the influence of GC-pressure.

Mutational AT-pressure in genomes of pathogens may facilitate immune escaping by the way of frequent destruction of B-cell epitopes (Khrustalev, 2010). Mutational GC-pressure may facilitate immune escaping by the way of creation of new B-cell epitopes (associated with the modification of previously existed ones) and by the way of the enlargement of previously existing ones (Khrustalev, 2009). However, our present findings showed that immunogenic features of viral glycoproteins depend on GC-content of subsequent genes in the same way and with the same strength as immunogenic features of "housekeeping" enzymes, which are surely not the main targets for B-cell immune response. It means that the difference in immunogenicity of proteins is one of the multiple consequences of the directional mutational pressure. In general, difference in immunogenicity of proteins associated with the GC-content of genes is not caused by immune pressure, while it undoubtedly may play a role in interactions between host and pathogen.

5. Conclusion

Mutational GC-pressure usually leads to the increase in the percent of amino acids included in discontinuous (and linear) B-cell epitopes, while mutational AT-pressure usually leads to the decrease in this percent. This general tendency reflects characteristic features of the tertiary structure of proteins encoded by GC-rich and AT-rich genes: the number and length of protruding parts of a protein situated on its surface grows due to GC-pressure and decrease due to AT-pressure. The impact of mutational pressure on the growth and on the decrease of immunogenicity of a protein is "controlled" by negative selection on the fixation of amino acid replacements leading to proline and glycine appearance or disappearance.

References

Calis, J.J.A., Sanchez-Perez, C.F., Kesmir, C., 2010. MHC class I molecules exploit the low G+C content of pathogen genomes for enhanced presentation. Eur. J. Immunol. 40, 2699–2709.

Chou, P.Y., Fasman, G.D., 1978. Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. Relat. Areas Mol. Biol. 47, 45–148.

Cohrs, R.J., Mehta, S.K., Scott, S.D., Gilden, D.H., Pierson, D.L., 2008. Asymptomatic reactivation and shed of infectious varicella zoster virus in astronauts. J. Med. Virol. 80, 1116–1122.

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.

Hashido, M., Kawana, T., 1997. Herpes simplex virus-specific IgM, IgA and IgG subclass antibody responses in primary and nonprimary genital herpes patients. Microbiol. Immunol. 41, 415–420.

Hopp, T.P., Woods, K.R., 1983. A computer program for predicting protein antigenic determinants. Mol. Immunol. 20, 483–489.

Kaufman, H.E., Azcuy, A.M., Varnell, E.D., Sloop, G.D., Thompson, H.W., Hill, J.M., 2005. HSV-1 DNA in tears and saliva of normal adults. Invest. Ophthalmol. Vis. Sci. 46, 241–247.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press.

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. Bioinf. 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, 295–302.

Khrustalev, V.V., Barkovsky, E.V., 2010. 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. 266, 88–98.

Khrustalev, V.V., Barkovsky, E.V., 2011a. Unusual nucleotide content of Rubella virus genome as a consequence of biased RNA-editing: comparison with Alphaviruses. Int. J. Bioinf. Res. Appl. 7, 82–100.

Khrustalev, V.V., Barkovsky, E.V., 2011b. "Protoisochores" in certain archaeal species are formed by replication-associated mutational pressure. Biochimie 93, 160–167.

Khrustalev, V.V., 2009. Can mutational GC-Pressure create new linear B-cell epitopes in Herpes Simplex virus type 1 glycoprotein B? Immunol. Invest. 38, 613–623.

Khrustalev, V.V., 2010. Mutational pressure makes HIV1 gp120 linear B-cell epitopes shorter and may lead to their disappearance. Mol. Immunol. 47, 1635–1639.

Kuhn, J.E., Dunkler, G., Munk, K., Braun, R.W., 1987. Analysis of the IgM and IgG antibody response against herpes simplex virus type 1 (HSV-1) structural and nonstructural proteins. J. Med. Virol. 23, 135–150.

Kyte, J., Doolittle, R., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.

Larsen, J.E.P., Lund, O., Nielsen, M., 2006. Improved method for predicting linear B-cell epitopes. Immunome Res. 2, 2.

Liljeqvist, J.A., Tunback, P., Norberg, P., 2009. Asymptomatically shed recombinant herpes simplex virus type 1 strains detected in saliva. J. Gen. Virol. 90, 559–566.

Miller, C.S., Danaher, R.J., 2008. Asymptomatic shedding of herpes simplex virus (HSV) in the oral cavity. Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod. 105, 43–50.

Rubinstein, N.D., Mayrose, I., Halperin, D., Yekutieli, D., Gershoni, J.M., Pupko, T., 2008. Computational characterization of B-cell epitopes. Mol. Immunol. 45, 3477–3489.

Rubinstein, N.D., Mayrose, I., Martz, E., Pupko, T., 2009. Epitopia: a web-server for predicting B-cell epitopes. BMC Bioinf. 10, 287.

Rocha, E.P., Danchin, A., 2002. Base composition bias might result from competition for metabolic resources. Trends Genet. 18, 291–294.

Rost, B., Yachdav, G., Liu, J., 2004. The PredictProtein server. Nucl. Acids Res. 32, 321–326.

Scott, D.A., Coulter, W.A., Lamey, P.J., 1997. Oral shedding of herpes simplex virus type 1: a review. J. Oral Pathol. Med. 26, 441–447.

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.

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

Tsodikov, O.V., Record, M.T., Sergeev, Y.V., 2002. Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. J. Comput. Chem. 23, 600–609.

Tunbäck, P., Liljeqvist, J.A., Löwhagen, G.B., Bergström, T., 2000. Glycoprotein G of herpes simplex virus type 1: identification of type-specific epitopes by human antibodies. J. Gen. Virol. 81, 1033–1040.