Immunological Investigations, 39:551–569, 2010
Copyright © Informa Healthcare USA, Inc.
ISSN: 0882-0139 print / 1532-4311 online
DOI: 10.3109/08820131003706313

informa
healthcare

Levels of HIV1 gp120 3D B-cell Epitopes Mutability and Variability: Searching for Possible Vaccine Epitopes

Vladislav Victorovich Khrustalev

Department of General Chemistry, Belarussian State Medical University, Minsk 220000, Belarus

We used a DiscoTope 1.2 (http://www.cbs.dtu.dk/services/DiscoTope/), Epitopia (http://epitopia.tau.ac.il/) and EPCES (http://www.t38.physik.tu-muenchen.de/programs.htm) algorithms to map discontinuous B-cell epitopes in HIV1 gp120. The most mutable nucleotides in HIV genes are guanine (because of G to A hypermutagenesis) and cytosine (because of C to U and C to A mutations). The higher is the level of guanine and cytosine usage in third (neutral) codon positions and the lower is their level in first and second codon positions of the coding region, the more stable should be an epitope encoded by this region. We compared guanine and cytosine usage in regions coding for five predicted 3D B-cell epitopes of gp120. To make this comparison we used GenBank resource: 385 sequences of *env* gene obtained from ten HIV1-infected individuals were studied (http://www.barkovsky.hotmail.ru/Data/Seqgp120.htm). The most protected from nonsynonymous nucleotide mutations of guanine and cytosine 3D B-cell epitope is situated in the first conserved region of gp120 (it is mapped from 66th to 86th amino acid residue). We applied a test of variability to confirm this finding. Indeed, the less mutable predicted B-cell epitope is the less variable one. MEGA4 (standard PAM matrix) was used for the alignments and "VVK Consensus" algorithm (www.barkovsky.hotmail.ru) was used for the calculations.

Keywords HIV1, env, gp120, Discontinuous B-cell epitopes, Epitope stability, Mutational pressure, HIV1 vaccine.

INTRODUCTION

Selection of stable epitopes for anti-HIV1 vaccine production is one of the main aims of recent molecular immunology (Hoxie, 2009). Unfortunately, the rates of amino acid substitutions in HIV1 proteins and so in their epitopes are

Address correspondence to Vladislav Victorovich Khrustalev, 7-24 Communisticheskaya street, Minsk 220029, Belarus; E-mail: vvkhrustalev@mail.ru

too high (Yang, 2009). The last variant of anti-HIV1 vaccine consists of several T-cell epitopes from different HIV1 subtypes (Berkhout and Paxton, 2009). Although, some investigators believe that T-cell immune response is not sufficient to prevent HIV1-infection, and B-cell immune response should also be involved in the protective immunity (Haynes and Montefiori, 2006; Hoxie, 2009).

B-cell epitopes are parts of proteins or other molecules recognized and bound by antibodies produced by B-cells (Larsen et al., 2006). Discontinuous B-cell epitopes are composed of different parts of the polypeptide chain that are brought into spatial proximity by the folding of the protein (Larsen et al., 2006). That is why they are usually called 3D-epitopes. There is well-known software – DiscoTope 1.2 – for the prediction of 3D-epitopes (Andersen, 2006). This software works with known 3D structure of a protein. The information about 3D structure is stored in PDB database. In this work we predicted the boarders of B-cell 3D-epitopes of HIV1 gp120 with the help of DiscoTope 1.2 using the information about eleven 3D structures of the HIV1 gp120 core.

Results of DiscoTope 1.2 predictions have been confirmed by two recently created methods for 3D-epitopes predictions: Epitopia (Rubinstein et al., 2009) and EPCES (Liang et al., 2009).

What makes the structure of HIV1 gp120 3D-epitopes so unstable? What kinds of molecular processes cause high rates of amino acid substitutions in HIV1 proteins? Cellular cytidine deaminases from APOBEC3 family deaminate cytosine residues in HIV1 DNA minus strands (Izumi et al., 2008). Cytosine is complementary to guanine, while the product of cytosine deamination (uracil) is complementary to adenine. So, during the synthesis of RNA on the matrix of HIV1 DNA minus strands frequent G to A nucleotide substitutions occur. This phenomenon is known under the term “G to A hypermutagenesis” (Izumi et al., 2008).

Recent experimental works confirmed that cellular cytosine deaminases from APOBEC1 family are able to deaminate cytosine in HIV1 RNA, at least, in mouse (Petit et al., 2009) and rat models (Bishop et al., 2004), producing C to U transitions.

The reverse transcriptase of HIV1 was shown to incorporate oxidized guanine (8-oxo-G) preferably in front of cytosine (Kamath-Loeb et al., 1997). Being unrepaired this kind of mispair (8-oxoG:C) should lead to C to A transversion in HIV1 RNA. The most of cellular polymerases, including RNA polymerases, preferably incorporate adenine in front of 8-oxo-G (Kamath-Loeb et al., 1997). So, if 8-oxo-G is incorporated into HIV DNA minus strand against cytosine during the reverse transcription, adenine will occur in front of this 8-oxo-G during the “simple” transcription of HIV RNA.

The most mutable nucleotides in HIV1 genes should be guanine (G) and cytosine (C). This statement is confirmed by the mutational pressure theory (Sueoka, 1988). Levels of both guanine and cytosine usage in HIV1 *env* gene are lower than 25%, especially in third codon positions (see the results section).

In our recent article we showed that the variability of HIV1 gp120 V3 region is enhanced by the guanine usage bias in the region coding for it (Khrustalev, 2009a). Indeed, the higher is the level of "hot spots for the mutation" in the coding region, the more unstable the encoded protein should be. However, these "hot spots" (mutable nucleotides) can be situated in third codon positions. In most of the cases mutation in third codon position will not lead to the amino acid replacement in the protein. So, the higher is the level of mutable nucleotides in "neutral" third codon positions, the higher is the probability of synonymous mutation occurrence (Khrustalev and Barkovsky, 2008). Mutable nucleotides situated in first and second codon positions should be protected from mutations in case if the level of those nucleotides in third codon positions is relatively high (Khrustalev, 2009a). So, we used the term "protective buffer" to refer to the level of guanine in third codon positions of HIV genes (Khrustalev, 2009a).

In the present work we calculated the level of guanine and cytosine usage in regions coding for 5 predicted discontinuous B-cell epitopes of HIV1 gp120. To make our calculations reliable we used a collection of sequences coding for HIV1 gp120 obtained from GenBank. There were 10 groups of sequences in our collection. Each group of sequences was derived from a periodic sampling of viral population infecting a single patient (Bailey et al., 2006; Bunnik et al., 2008; Fransen et al., 2008). We predicted the probability of nonsynonymous mutation of GC to AT direction in those coding regions. The region coding for predicted B-cell epitope from the first conserved region of gp120 (C1) is protected from G to A, C to U and C to A mutations much better than regions coding for other epitopes.

In the final part of our study we performed a simple test to estimate levels of variability of five predicted B-cell epitopes in viruses from ten patients (the total number of studied full length *env* sequences is 385). In all of these ten groups the region coding for the predicted B-cell epitope from the first conserved region of gp120 (C1) is much more stable than regions coding for other epitopes.

Our results confirm that mutational pressure theory (Sueoka, 1988; Khrustalev and Barkovsky 2008) can be successfully used for the prediction of the mutability (and stability) of regions coding for HIV1 epitopes. The bioinformatical method described in this article can be simply adapted to B-cell epitopes from other HIV1 proteins as well as to T-cell epitopes.

EXPERIMENTAL PROCEDURES

To map the boarders of gp120 3D epitopes we chose 11 PDB records. We selected the records describing longest gp120 polypeptides. Ten of those records describe the structure of the core of gp120 in the interaction with CD4 molecule and the neutralizing antibody (PDB accession numbers: 2NY0 – 2NY7, 2NXY and 2NXZ). Variable loops have been partially cut off from the studied gp120 polypeptides (Zhou et al., 2007). The 11th record describes the core of gp120 with V3 region (Huang et al., 2005); its PDB accession number is 2B4C.

We predicted the boarders of 3D B-cell epitopes in the core of gp120 with the help of DiscoTope 1.2 computer algorithm (Andersen, 2006). DiscoTope 1.2 uses a combination of amino acid statistics, spatial information, and surface exposure. It was trained on a compiled data set of discontinuous epitopes from 76 X-ray structures of antibody/antigen protein complexes (Andersen, 2006).

We found out that there are four relatively long 3D B-cell epitopes predicted in all records describing gp120 core without variable loops. However, there are some conformational changes in predicted discontinuous epitopes (see Figures 1–5).

Results of Epitopia server predictions showed that the most of amino acids included in discontinuous B-cell epitopes by DiscoTope 1.2 are exposed to solvent. This criterion is the crucial one for the distinguishing between epitopes (which may have different scores of immunogenicity) and “buried” residues (which cannot actively participate in the antigen/antibody interactions) for Epitopia (Rubinstein et al., 2009).

According to the results of EPCES calculations there are always several amino acid residues in each epitope predicted by DiscoTope 1.2 which have very high antigenicity score (>80). Six terms (residue epitope propensity, conservation score, side chain energy score, contact number, surface planarity score and secondary structure composition) are used for antibody binding site prediction by EPCES (Liang et al., 2009).

The first B-cell epitope is mapped in the first conserved region of gp120 (C1). The average length of this epitope is 21 amino acids (see Figure 1).

V3 loop is a well known strong B-cell epitope (Khrustalev, 2009a). That is why we decided to include full length V3 loop in our study. Indeed, DiscoTope 1.2 mapped B-cell epitope in the V3 region of the core of gp120 including V3 loop (see Figure 2).

The third long B-cell epitope is mapped in the third conserved region of gp120 (C3). In fact, there are at least three short amino acid stretches situated near each other (see Figure 3).

The fourth predicted B-cell epitope is situated in the V4 variable loop of gp120. As one can see in Figure 4, V4 loop has been cut off from the core of gp120 only partially.

The last relatively long predicted B-cell epitope of gp120 (see Figure 5) is situated not only in the V5 region of gp120, but also in flanking parts of C4 and C5. Canonical V5 region of gp120 is really short (eight amino acid residues in length).

```
                                   3D epitope#1
                                 66            86
HIV1 reference strain   VVLVNVTENFNMWKNDMVEQMHEDIISLWDQ
2B4C                    VVLQNVT-H-NMWKNDMVEQMQED--S---Q
2NY7                    -------------KND-VEQMHE--------
2NY6                    -------------KND--EQ--E--------
2NY5                    -------------KND--EQ-HE--------
2NY4                    ------------WKNDMVEQMHED-IS---Q
2NY3                    ------------WKNDMVEQMHED-IS---Q
2NY2                    ----------N-WKNDMVEQMHED--S---Q
2NY1                    ----------N-WKND-VEQ-HE--------
2NY0                    -------------KND--EQ-HED-IS---Q
2NXZ                    ----------N-WKNDMVEQMHED-IS---Q
2NXY                    ----------N-WKNDMVEQMHED-IS---Q
```

Figure 1: Amino acids included in predicted B-cell epitope situated in the first conserved region of gp120. Amino acids that were not included in B-cell epitopes by DiscoTope 1.2 are substituted by the symbol "-". Amino acids included in B-cell epitopes by Epitopia are written in **bold font**. Amino acids with antigenicity score higher than 80 according to EPCES prediction are underlined.

```
                                V3 loop = 3D epitope#2
                         286                                303
HIV1 reference strain    CTRPNNNTRKRIRIQRGPGRAFVTIGKI0GNMRQAHC
2B4C                     ---PNQNTRKSIHIGPGR00AFYTTGEIIGDI---
```

Figure 2: Amino acids included in predicted B-cell epitope situated in the third variable region of gp120. Amino acids that were not included in B-cell epitope by DiscoTope 1.2 are substituted by the symbol "-". Gaps are shown by a symbol "**0**". Amino acids included in B-cell epitopes by Epitopia are written in **bold font**. Amino acids with antigenicity score higher than 80 according to EPCES prediction are underlined.

To test mutability and variability levels of five predicted 3D epitopes we decided to make a collection of *env* gene sequences coding for gp120. Ten groups of *env* gene sequences have been obtained from GenBank. Each group of sequences has been derived from a single HIV1-infected patient.

Two groups of sequences were derived from the study of viral tropism (Fransen et al., 2008), their names are "S33" (50 sequences) and "S35" (40 sequences), their GenBank accession numbers are: EU604549 – EU604642.

Five groups of sequences used in this study had been sequenced during the work on viral immune escape from humoral immunity (Bunnik et al., 2008), their names are "H1" – "H5" (41; 41; 42; 49 and 30 sequences in each of these five groups, respectively), their GenBank accession numbers are: EU743973 – EU744175.

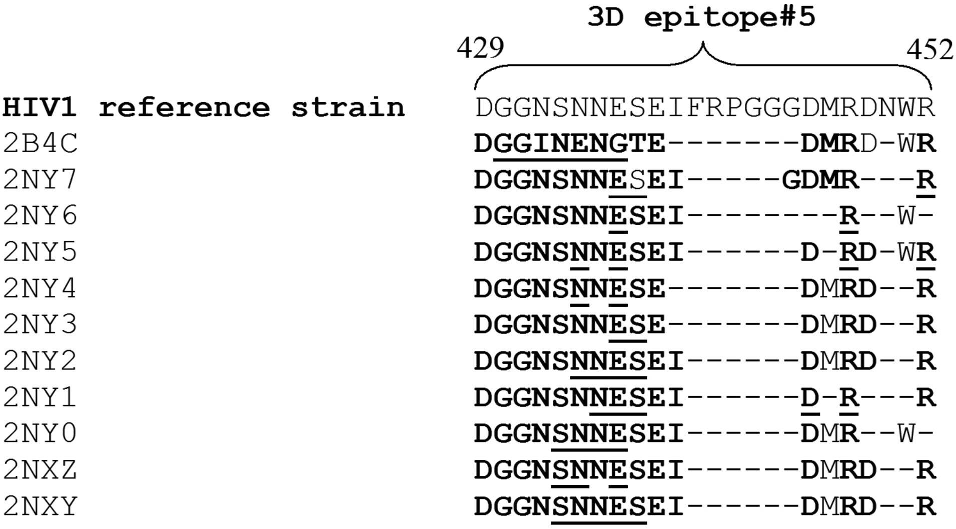
```
                                          3D epitope#3
                                   308                        340
HIV1 reference strain              AKWNNTLKQIASKLREQFGNNKTIIFKQSSGGD
2B4C                               A--ND--K------REQFEN?KT------SGGD
2NY7                               ---NN---------REQFGNNK-------SGGD
2NY6                               ---NN---------REQFGNNKT------SGGD
2NY5                               ---NN--K---S--REQFGNNKT------SGGD
2NY4                               ---NN---------REQFGNNKT------SGGD
2NY3                               A--NN---------REQFGNNKT------SGGD
2NY2                               ---NN---------REQFGNNKT------SGGD
2NY1                               A--NN--K------REQFGNNKT------SGGD
2NY0                               ---NN---------REQFGNNKT------SGGD
2NXZ                               ---NN------S--REQFGNNKT------SGGD
2NXY                               A--NN------S--REQFGNNKT------SGGD
```

Figure 3: Amino acids included in predicted B-cell epitope situated in the third conserved region of gp120. Amino acids that were not included in B-cell epitope by DiscoTope 1.2 are substituted by the symbol "-". Amino acids absent in PDB file are shown by a symbol "?". Amino acids included in B-cell epitopes by Epitopia are written in **bold font**. Amino acids with antigenicity score higher than 80 according to EPCES prediction are underlined.

```
                                     3D epitope#4
                               365                384
HIV1 reference strain          STWFNSTWSTEGSNNTEGSD
2B4C                           STWNN???NTEGSNNTEGN-
2NY7                           S---NSTWS????????GSD
2NY6                           ---FN????????????GSD
2NY5                           S-WFN????????????GSD
2NY4                           ---FNST??????????GSD
2NY3                           ---FNS???????????GSD
2NY2                           ---FN????????????GSD
2NY1                           --WFN????????????GSD
2NY0                           ---FNST??????????GSD
2NXZ                           ---FNST??????????GSD
2NXY                           ---FN????????????GSD
```

Figure 4: Amino acids included in predicted B-cell epitope situated in the fourth variable region of gp120. Amino acids that were not included in B-cell epitope by DiscoTope 1.2 are substituted by the symbol "-". Amino acids absent in PDB file are shown by a symbol "?". Amino acids included in B-cell epitopes by Epitopia are written in **bold font**. Amino acids with antigenicity score higher than 80 according to EPCES prediction are underlined.

Three groups of sequences are from the viruses infecting so-called "elite controllers". They were originally analyzed in the study performed by Bailey et al. in 2006. We used only three groups from this study, because the number of sequences in other groups is too small. The names of these groups are "ES4"



```
                                3D epitope#5
                          429                    452
HIV1 reference strain     DGGNSNNESEIFRPGGGDMRDNWR
2B4C                      DGGINENGTE-------DMRD-WR
2NY7                      DGGNSNNESEI-----GDMR---R
2NY6                      DGGNSNNESEI--------R--W-
2NY5                      DGGNSNNESEI------D-RD-WR
2NY4                      DGGNSNNESE-------DMRD--R
2NY3                      DGGNSNNESE-------DMRD--R
2NY2                      DGGNSNNESEI------DMRD--R
2NY1                      DGGNSNNESEI------D-R---R
2NY0                      DGGNSNNESEI------DMR--W-
2NXZ                      DGGNSNNESEI------DMRD--R
2NXY                      DGGNSNNESEI------DMRD--R
```

Figure 5: Amino acids included in predicted B-cell epitope situated in the fourth variable region of gp120. Amino acids that were not included in B-cell epitope by DiscoTope 1.2 are substituted by the symbol "-". Amino acids included in B-cell epitopes by Epitopia are written in **bold font**. Amino acids with antigenicity score higher than 80 according to EPCES prediction are underlined.

(18 sequences DQ410069 – DQ410086), "ES8" (47 sequences DQ410131 – DQ410178) and "ES9" (27 sequences DQ410188 – DQ410214).

We aligned predicted epitopes with the amino acid sequence of gp120 from HIV1 reference strain (NC_001802). Then we separately aligned amino acid sequences of gp120 derived from each patient and mapped positions of predicted 3D epitopes in them. Alignments have been performed by MEGA 4 program (Tamura et al., 2007); standard PAM matrix included in MEGA 4 program has been used.

Regions coding for each predicted epitope have been cut from the alignment of full length *env* genes. The collection of alignments made during this study (in a downloadable form) can be found on the following web page (http://www.barkovsky.hotmail.ru/Data/Seqgp120.htm).

The next step of calculations has been performed by "VVK Consensus" algorithm (www.barkovsky.hotmail.ru). This MS Excel spreadsheet is able to calculate the nucleotide content in the inserted sequences and to count the number of nucleotide substitutions between them (Khrustalev, 2009a). To use this algorithm one should copy the massive of aligned nucleotide sequences (without their names) from MEGA4 alignment explorer to any text processor. Then one should substitute "-" (symbol used to designate a gap by MEGA4) with a Latin letter "N" (this operation can be done easily by MS Word). After this procedure one should paste the massive of sequences into marked cells from "sequences" list of "VVK Consensus" spreadsheet. Nucleotide content in every codon position for each inserted sequence is calculated on the "content" list of "VVK Consensus" spreadsheet.

"VVK Consensus" creates consensus sequence for all the sequences inserted and counts the number of sites with nucleotide substitutions (counts mutations *per site*) and the number of substituted nucleotides (counts mutations *per nucleotide*) from this consensus sequence. These data can be found on the "results" list of "VVK Consensus" spreadsheet.

In case if "VVK Consensus" algorithm counts the number of sites with all possible types of nucleotide mutations in the alignment (counts mutations *per site*), the maximal number of nucleotide mutations is equal to "3 mutation *per site*".

In case if "VVK Consensus" algorithm counts the number of mutated nucleotides in the alignment (counts mutations *per nucleotide*), the maximal number of nucleotide mutations depends on the number of sequences (n) in the alignment (it is equal to "0.75n – 1 mutation *per nucleotide*").

To calculate the rates of nucleotide substitutions in regions coding for different epitopes we divided i) number of sites with different kinds of nucleotide mutations and ii) number of mutated nucleotides in the alignment by the length of each subsequent region coding for an epitope. Nucleotide content and the rates of nucleotide substitutions have been calculated separately in third and in first and second codon positions.

To compare the rates of nucleotide substitutions between regions coding for five predicted B-cell epitopes we performed paired differences test. This test compares measurements within subjects, rather than across subjects, and will generally have greater power than an unpaired test.

The time passed between the first and last sequencing of HIV1 *env* gene varies between ten patients. It means that we cannot calculate the average rates of nucleotide mutations in each of the epitopes for all ten groups. That is why we decided to calculate the differences in the rates of nucleotide mutations between regions coding for epitopes inside every group. The algorithm of paired differences test is as follows: first paired differences are calculated inside every group and then they are treated as variables in t-test.

So, according to the paired differences test, all the differences between the rates of mutations are statistically significant ($P < 0.05$), except the differences between their rates in regions coding for epitopes from C4-V5-C5 and V3, from C4-V5-C5 and C3, from C4-V5-C5 and V4 and from C3 and V4 (the last one – only for mutations counted *per nucleotide*).

Paired differences test has also been applied to nucleotide content comparisons. The level of guanine in first and second codon positions is significantly lower in regions coding for B-cell epitope from C1 and B-cell epitope from C3 than in regions coding for epitopes from V3, V4 and C4-V5-C5 ($P < 0.05$).

There is no significant difference between the level of guanine in first and second codon positions in regions coding for B-cell epitope from C1 and B-cell epitope from C3. However, the level of guanine in third codon positions is much higher ($P < 0.05$) in the region coding for B-cell epitope from C1 (see Figure 6a) in comparison with the region coding for B-cell epitope from C3 (see Figure 6c), as well as in comparison with three other regions. The level of cytosine in first and second codon positions of the region coding for B-cell epitope from C1 is significantly lower than in any other region ($P < 0.05$).

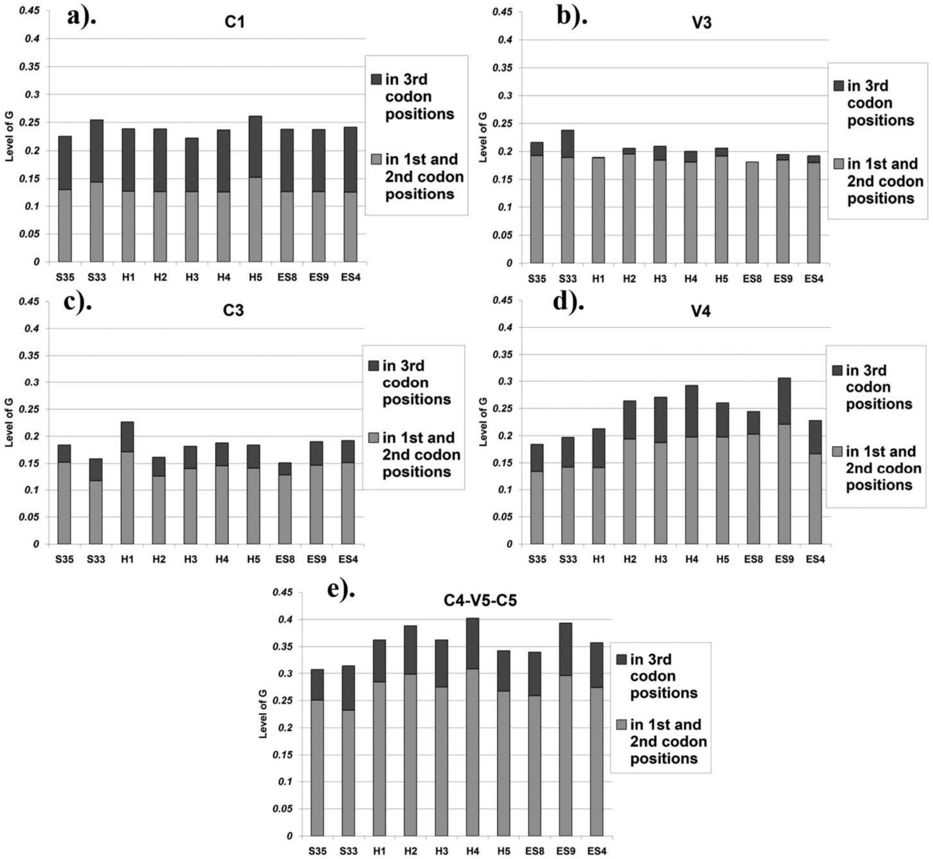


Figure 6: Average level of guanine usage (relative abundance of guanine) and percentage of guanine situated in third codon positions in 10 groups of sequences coding for 5 predicted B-cell epitopes of HIV1 gp120: a) for the epitope situated in C1; b) for the V3 loop; c) for the epitope situated in C3; d) for the epitope situated in V4; e) for the epitope situated in C4-V5-C5.

RESULTS

Before the proper description of the results some terminological issues should be clarified. The term “mutability” in the context of our research is used to describe the probability for the nucleotide substitution of GC to AT direction to occur in first or second codon position (i.e. to be nonsynonymous). The higher is the level of guanine and cytosine in first and second codon positions of the region coding for HIV1 epitope and the lower is the level of guanine and cytosine in its third codon positions, the higher is the *mutability* of this coding region (Khrustalev, 2009a). In other words, the term “mutability” is used to describe the expected (predicted) level of sequence diversity.

The term “variability” is used to describe the factual level of sequence diversity (the number of sites with different kinds of mutations or the number of mutated nucleotides in the region coding for an epitope from the given group of sequences divided by the length of this region). Finally we showed that predicted B-cell epitope from HIV1 gp120 first conserved region (C1) is coded by the region with *the lowest level of mutability* and that this predicted B-cell epitope is *significantly less variable* than other predicted B-cell epitopes of gp120.

In Figure 6 one can see average level of total guanine usage (relative abundance of guanine) in regions coding for 5 discontinuous B-cell epitopes of gp120 in each of the 10 studied groups of sequences. Each bar representing an average total level of G in sequences from the single patient coding for the given epitope is divided into two parts. The upper part of the bar represents the percentage of G situated in third (neutral) codon positions. The lower part of the bar represents the percentage of G sitiuated in first and second codon positions.

The lowest level of guanine situated in first and second codon positions is characteristic to regions coding for B-cell epitope from C1 and B-cell epitope from C3 (see Figure 6), but the level of guanine in third codon positions is higher in region coding for B-cell epitope from C1 than for B-cell epitope from C3. So, the region coding for B-cell epitope from C1 is protected from G to A hypermutagenesis better than any other region coding for an epitope.

As one can see in Figure 6b, “protective buffer” of guanine situated in third codon positions of the region coding for V3 loop is really the smallest one among other regions coding for predicted 3D epitopes (Khrustalev, 2009a).

In Figure 7 one can see the level of cytosine usage (relative abundance of cytosine) in regions coding for five B-cell epitopes of gp120. The lowest level of cytosine usage in first and second codon positions is characteristic to the region coding for B-cell epitope from C1. The highest level of cytosine in third codon positions is characteristic to the region coding for B-cell epitope from C4-V5-C5 (see Figure 7e). However, Figure 7e demonstrates that

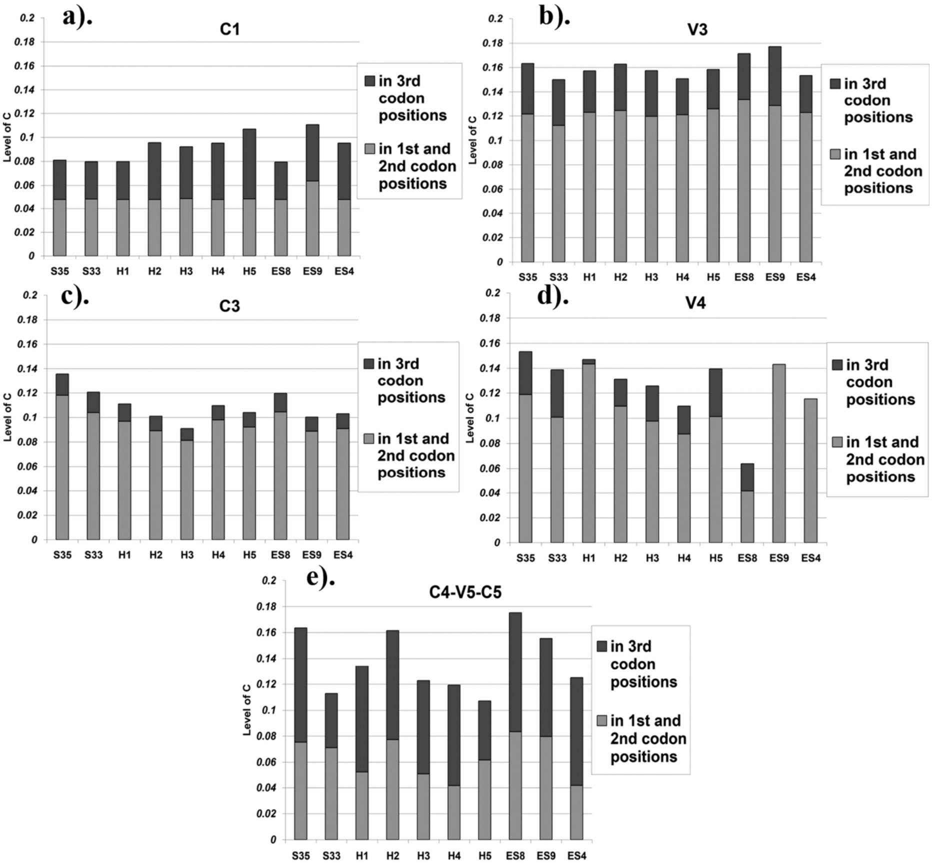


Figure 7: Average level of cytosine usage (relative abundance of cytosine) and percentage of cytosine situated in third codon positions in 10 groups of sequences coding for five predicted B-cell epitopes of HIV1 gp120: a) for the epitope situated in C1; b) for the V3 loop; c) for the epitope situated in C3; d) for the epitope situated in V4; e) for the epitope situated in C4-V5-C5.

cytosine usage (both the total level and the percentage of cytosine in third codon positions) varies greatly between ten groups of sequences. The variations of cytosine usage in regions coding for B-cell epitopes from C3 and V4 are also significant. Coming back to Figure 6 one can admit that guanine usage is also quite unstable in regions coding for B-cell epitopes from V4 and C4-V5-C5.

In Table 1 we showed average level of cytosine and guanine usage in regions coding for B-cell epitopes from all studied sequences. We have to state that the B-cell epitope from C1 region of gp120 is protected from nonsynonymous G to A, C to U and C to A substitutions much better than other regions coding for B-cell epitopes.

Table 1: Average level of guanine and cytosine usage, their percentage in third codon positions of the regions coding for 5 predicted B-cell epitopes.

Region coding for B-cell epitope	#1(C1)	#2(V3)	#3(C3)	#4(V4)	#5(C4-V5-C5)
Average level of G	0.239	0.203	0.182	0.246	0.357
Average probability for the mutation of G to occur in third codon position	45%	8%	22%	28%	23%
Average level of C	0.092	0.160	0.137	0.127	0.138
Average probability for the mutation of C to occur in third codon position	46%	23%	12%	16%	54%

Table 2: Information on gaps found in aligned amino acid sequences of the HIV1 *env* gene regions coding for predicted B-cell epitopes.

Group of sequences	Length of indels in predicted B-cell epitopes				
	C1	V3	C3	V4	C4-V5-C5
S35				5AA	1AA; 1AA
S33			1AA	9AA; 2AA; 2AA	
H1				6 – 24AA;1 – 2AA	1AA; 1AA; 1AA
H2			2AA; 1AA	4AA	2AA; 1AA; 1AA; 1AA
H3				3 – 4AA; 3AA	2 – 4AA; 1AA
H4			1AA	1AA; 1AA	3AA; 1AA; 1AA; 1AA
H5			1AA	3 – 6AA; 1AA	2 – 3AA;1 – 2AA
ES8					1 – 3AA; 1AA; 1AA
ES9				2 – 10AA	2AA
ES4				6AA; 1AA	

Sequences from each group have been aligned separately (standard PAM matrix included in MEGA 4 has been used).

The information contained in Table 2 helps us to understand why the usage of cytosine and guanine varies so greatly between ten groups of sequences in regions coding for B-cell epitopes from C3, V4 and C4-V5-C5. These variations in nucleotide content are enhanced by insertions and deletions of different kind. For example, the level of guanine in one of the regions coding for an epitope from V4 loop 25 codons in length is equal to 18.67%, the level of cytosine is equal to 17.33%. Another sequence from the same group (S35) has a shorter length (15 codons). If we introduce the same kind of deletion (5 codons in length) into the longer sequence, the level of guanine inside it (16.67%), as well as the level of cytosine (16.67%), will become lower. In other words, deletion of a short relatively GC-rich sequence will decrease the level of GC-content in the coding region, while insertion of a short relatively GC-rich sequence will increase the total GC-content of the coding region (Khrustalev, 2009a).

Interestingly, (see Table 2) epitopes from C1 and V3 regions of gp120 do not contain any insertion or deletion (we looked for indels inside each group of aligned sequences). The most of indels in three other predicted B-cell epitopes were short (from 1 to 2 amino acid residues), while in a few sequences they were relatively long (see Table 2).

To determine the level of factual diversity among monophyletic nucleotide sequences we used "VVK Consensus" algorithm.

In Figure 8 the factual levels of variability are shown. Figure 8a shows total number of sites with different kinds of nucleotide mutations in every group of sequences coding for each B-cell epitope divided by the length of the subsequent region. Figure 8b shows the rates of mutations counted *per site* only in first and second codon positions. One cannot compare the rates of mutations between groups, but only inside each group (between five regions coding for B-cell epitopes). That is why we used a paired differences test mentioned above.

The information given in Figure 9 should be even more helpful for the estimation of epitope stability than the information given in Figure 8. Indeed, Figure 9a shows total number of mutated nucleotides in every group of sequences coding for each B-cell epitope divided by the length of the subsequent region. Figure 9b shows the rates of mutations counted *per nucleotide* only in first and second codon positions. The data presented in Figure 9, unlike the data presented in Figure 8, illustrate the distribution of polymorphism in aligned sequences.

The result of our work is the following. The rates of nucleotide mutations in region coding for epitope from C1 and in V3 region are significantly lower than the rates of nucleotide mutations in regions coding for three other predicted B-cell epitopes. This fact should be connected with frequently occurring deletions and insertions in regions coding for epitopes from C3, V4 and C4-V5-C5.

The rates of nucleotide mutations in region coding for epitope from C1 are significantly lower than the rates of nucleotide mutations in V3 region. This should be the consequence of the increased mutability of V3 region and low mutability of the region coding for epitope from C1.

There are no amino acids coded by GC-rich codons (glycine, alanine, arginine and prolyne) in the consensus sequence of the epitope from C1 (see Figure 1). These "mutable" (in HIV proteins) amino acids are present in four other B-cell epitopes (see Figures 2–5). That is why the level of hot spots for mutation (guanine and cytosine situated in first and second codon positions) in the epitope from C1 is significantly decreased.

In Figure 10 the consensus sequence of the most stable predicted HIV1 gp120 B-cell epitope (from C1 region) is shown. There are 4 polymorphic sites (sites containing given amino acid substitution in more than 2 sequences) in

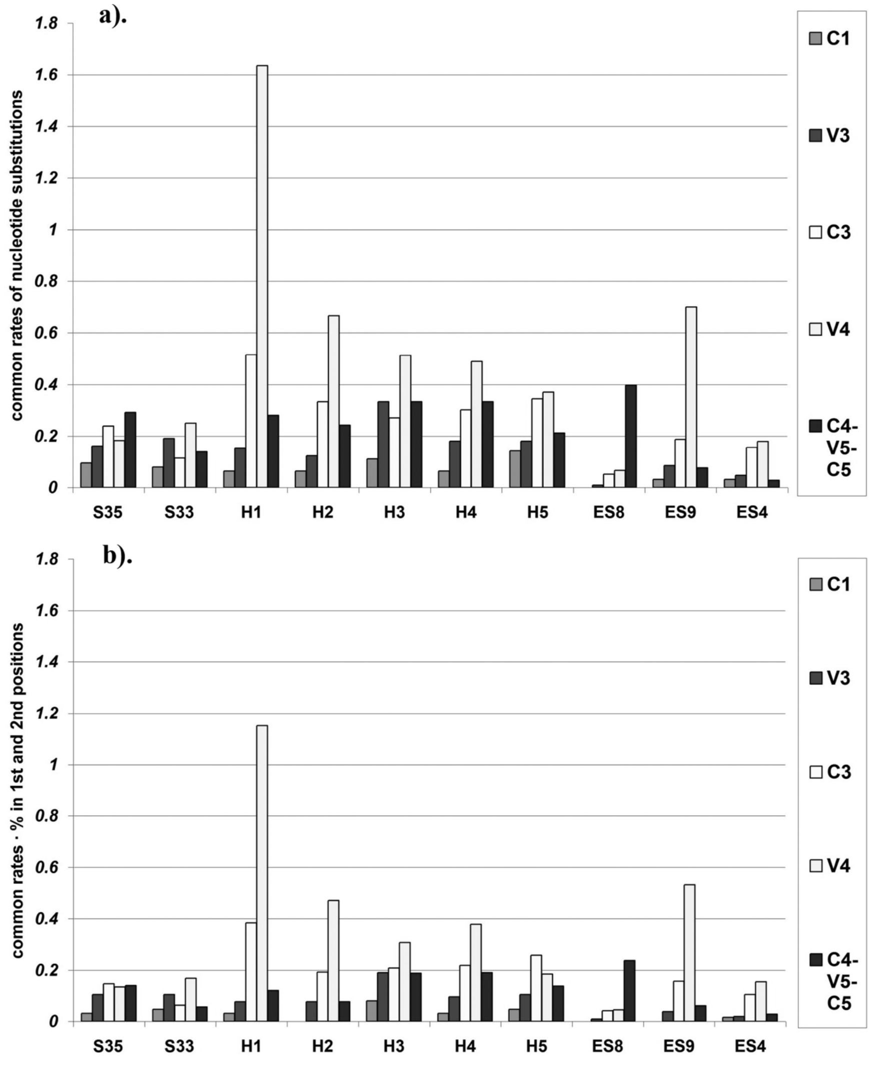


Figure 8: The rates of nucleotide mutations in ten groups of sequences coding for five predicted B-cell epitopes of HIV1 gp120. Figure 8a shows *the number of sites with different kinds of nucleotide mutations* in the alignment divided by the length of the region. Figure 8b shows the rates of nucleotide mutations only in first and second codon positions. Mutated sites have been counted from the individual consensus sequences.

the amino acid alignment of C1 region obtained during the periodical sampling from 10 patients. In the viral population infecting patient “S33” the first polymorphic site of the epitope from C1 region (see Figure 10) contains aspartic acid (D) instead of asparagine (N) which can be found in this site of the consensus sequence. In the viral population infecting patient “H5” the first polymorphic site may contain either aspartic acid (D) or glycine (G). There is

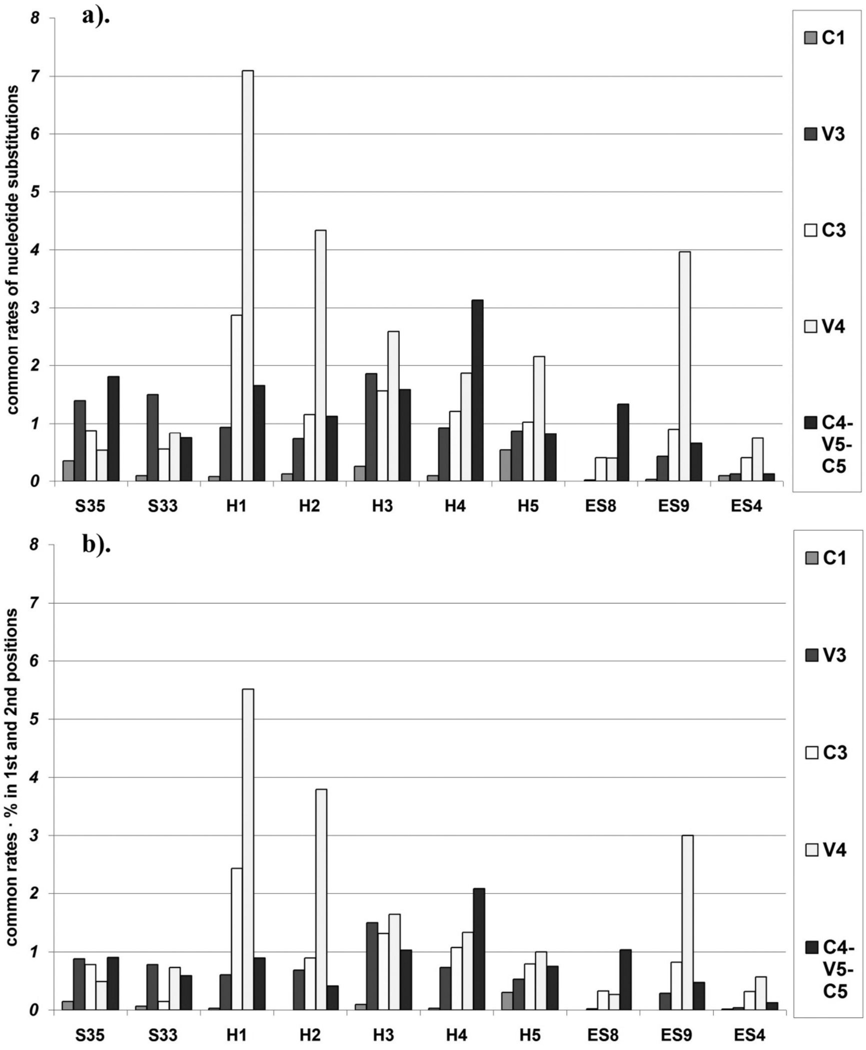


Figure 9: The rates of nucleotide mutations in 10 groups of sequences coding for five predicted B-cell epitopes of HIV1 gp120. Figure 9a shows *the number of mutated nucleotides* in the alignment divided by the length of the region. Figure 9b shows the rates of nucleotide mutations only in first and second codon positions. Mutated nucleotides have been counted from the individual consensus sequences.

aspartic acid (D) instead of glutamic acid (E) in the second polymorphic site of the epitope from C1 region of gp120 in the viral population infecting patients “H5” and “S33”, while in the viral population infecting patient “S35” there may be either aspartic (D) or glutamic acid (E) in this site. In the viral population infecting patient “S35” the third polymorphic site may contain either glutamic acid (E) or glycine (G), while the fourth polymorphic site contains

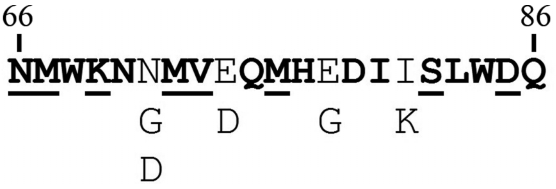


Figure 10: The consensus sequence of the predicted B-cell epitope situated in the first conserved region of HIV1 gp120. Invariable amino acid residues are written in **bold** type. Conserved amino acid residues (residues substituted in 1 or 2 sequences from 385 aligned ones) are written in **bold,** underlined type. Variable amino acid residues are written in regular type (residues found more than in 2 sequences from 385 aligned ones are written under the residues from consensus sequence).

lysine (K) instead of isoleucine (I). There are no polymorphic sites in the epitope from C1 region in the viral populations of 7 from 10 patients. In our opinion this epitope can be used in vaccine production because of its relative stability.

We hope that these data obtained from *in-silico* experiments will help to perform next steps of immunological investigations *in vitro* and *in vivo*.

DISCUSSION

The only thing we can be sure of is that the region of HIV1 gp120 protein mapped from 66th to 86th amino acid residue (in the reference strain) is significantly less variable than regions mapped from 268th to 303rd, from 308th to 340th, from 365th to 384th and from 429th to 452nd amino acid residue. But can we trust DiscoTope 1.2, Epitopia and EPCES predictions? Is the predicted B-cell epitope from C1 region of gp120 really immunogenic?

The immunogenic epitopes recognized by neutralizing antibodies (in mice immunized with recombinant vaccinia viruses expressing the wild-type of HIV1 gp160) were mapped primarily to the variable loops (V1-2) and to the conserved regions (C1 and C5) of gp120 (Kim et al., 2003). This fact is the evidence that the discontinuous B-cell epitope in C1 region of gp120 does exist.

Moreover, it is well known that antibodies specific for relatively conserved regions of *env* generally exhibit antibody-dependent cellular cytotoxicity activity against a broader range of HIV1 strains than those directed against variable epitopes (Alsmadi and Tilley, 1998).

According to the works of T.P. Hopp, the most acrophilic amino acid residues (residues situated on the surface of protein globule and so included in B-cell epitopes) are glycine and proline (Hopp, 1984; Hopp and Woods, 1983). These residues are coded by GC-rich codons. The level of GC-rich codons usage should increase with the growth of GC-content (Singer and Hickey, 2000). As we have found out, mutational pressure of AT to GC direction leads to the appearance of new linear B-cell epitopes or to the enlargement of previously existing ones at the probability of 25% (Khrustalev, 2009b). Does it mean that mutational pressure of GC to AT direction can lead to the disappearance of linear B-cell epitopes? The replacement of proline or glycine residues may change the tertiary structure of the protein: the epitope or its part may become “hidden” (buried) in the inner part of the protein.

Predicted B-cell epitope from C1 region of gp120 contains no glycine and proline, unlike other B-cell epitopes of gp120 (for example, there are four glycines and two prolines in V3 loop sequence from the reference HIV1 strain). It should make the epitope from C1 region less immunogenic (Hopp and Woods, 1983), but much more stable than any other epitope on the surface of gp120.

The method of epitope selection described in this article can be used to select the most stable T-cell epitopes of gp120, too. The principle of the selection is simple. Once you know the most “mutable nucleotides” (guanine and cytosine are the most mutable nucleotides in HIV genes), you can calculate the level of their usage in each region coding for an epitope. “Mutable nucleotides” situated in first and second codon positions are the “hot spots” for mutations, while these nucleotides in third codon positions are the kind of the “protective buffer” (Khrustalev, 2009a).

Performed in-silico test of variability confirmed that the less mutable region coding for predicted B-cell epitope from C1 of gp120 (this region contains the lowest amount of “hot spots” for mutations and the highest amount of “protective buffer”) is the most stable one among other regions coding for predicted 3D epitopes of HIV1 gp120.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

REFERENCES

Alsmadi, O., Tilley, S. A. Antibody-Dependent cellular cytotoxicity directed against cells expressing human immunodeficiency virus type 1 envelope of primary or laboratory-adapted strains by human and chimpanzee monoclonal antibodies of different epitope specificities. *J. Virol.* 72(1):286–293.

Andersen, P. H., Nielsen, M., Lund, O. Prediction of residues in discontinuous B cell epitopes using protein 3D structures. *Protein Science* 15:2558–2567.

Bailey, J. R., Lassen, K. G., Yang, H. C. Neutralizing antibodies do not mediate suppression of human immunodeficiency virus type 1 in elite suppressors or selection of plasma virus variants in patients on highly active antiretroviral therapy. *J Virol.* 80:4758–4770.

Bishop, K. N., Holmes, R. K., Sheehy, A. M., Malim, M. H. APOBEC-mediated editing of viral RNA. *Science* 305:645.

Berkhout, B., Paxton, W. A. HIV vaccine: it may take two to tango, but no party time yet. *Retrovirology* 6(1):88.

Bunnik, E. M., Pisas, L. van Nuenen, A. C., Schuitemaker, H. Autologous neutralizing humoral immunity and evolution of the viral envelope in the course of subtype B human immunodeficiency virus type 1 infection. *J Virol.* 82:7932–7941.

Fransen, S., Bridger, G., Whitcomb, J. M., Toma, J., Stawiski, E., Parkin, N., Petropoulos, C. J., Huang, W. Suppression of dualtropic human immunodeficiency virus type 1 by the CXCR4 antagonist AMD3100 is associated with efficiency of CXCR4 use and baseline virus composition. *Antimicrob Agents Chemother* 52:2608–2615.

Haynes, B. F., Montefiori, D. C. Aiming to induce broadly reactive neutralizing antibody responses with HIV-1 vaccine candidates. *Expert Rev. Vaccines* 5(3):347–363.

Hopp, T. P. Protein antigen conformation: folding patterns and predictive algorithms; selection of antigenic and immunogenic peptides. *Ann. Sclavo.* 2:47–60.

Hopp, T. P., Woods, K. R. A computer program for predicting protein antigenic determinants. *Mol. Immunol.* 20(4):483–489.

Hoxie, J. A. Toward an antibody-based HIV vaccine. *Annu. Rev. Med.* 61: Epub ahead of print.

Huang, C. C., Tang, M., Zhang, M. Y., Majeed, S., Montabana, E., Stanfield, R. L., Dimitrov, D. S., Korber, B., Sodroski, J., Wilson, I.A., Wyatt, R., Kwong, P. D. Structure of a V3-containing HIV-1 gp120 core. *Science* 310(5750):1025–1028.

Izumi, T., Shirakawa, K. Takaori-Kondo, A. Cytidine deaminases as a weapon against retroviruses and a new target for antiviral therapy. *Mini Rev. Med. Chem.* 8:231–238.

Kamath-Loeb, A. S., Hizi, A., Kasai, H. and Loeb, L. A. Incorporation of the guanosine triphosphate analogs 8-oxo-dGTP and 8-NH_2-dGTP by reverse transcriptases and mammalian DNA polymerases. *JBC* 272(9): 5892–5898.

Khrustalev, V. V., Barkovsky, E. V. An in-silico study of alphaherpesviruses ICP0 genes: Positive selection or strong mutational GC-pressure? *IUBMB Life* 60(7):456–460.

Khrustalev, V. V. HIV1 V3 loop hypermutability is enhanced by the guanine usage bias in the part of *env* gene coding for it. *In Silico Biology* 9:0022.

Khrustalev, V. V. Can mutational GC-pressure create new linear B-cell epitopes in herpes simplex virus type 1 glycoprotein B? *Immunol. Investig.* 38(7):613–623.

Kim, Y. B., Han, D. P., Cao, C., Cho, M. W. Immunogenicity and ability of variable loop-deleted human immunodeficiency virus type 1 envelope glycoproteins to elicit neutralizing antibodies. *Virology* 305(1):124–137.

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

Liang, S., Zheng, D., Zhang, C., Zacharias, M. Prediction of antigenic epitopes on protein surfaces by consensus scoring. *BMC Bioinformatics* 10:302.

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

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

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

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

Tamura, K., Dudley, J., Nei, M., Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24:1596–1599.

Yang, O. O. Candidate vaccine sequences to represent intra- and inter-clade HIV-1 variation. *PLoS One* 4(10): e7388.

Zhou, T., Xu, L., Dey, B., Hessell, A. J., Ryk, D.V., Xiang, S. H., Yang, X., Zhang, M. Y., Zwick, M. B., Arthos, J., Burton, D. R., Dimitrov, D. S. Sodroski, J., Wyatt, R., Nabel, G. J., Kwong, P. D. Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature* 445(7129):732–737.