JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE

A method for estimation of immunogenic determinants mutability: case studies of HIV1 gp120 and diphtheria toxin

Khrustalev Vladislav Victorovich*[1], Barkovsky Eugene Victorovich[1], Vasilevskaya Alla Evgenyevna[2], Skripko Svetlana Michailovna[2], Kolodkina Valentina Leonidovna[3], Ignatyev Georgiy Michailovich[3], Semizon Pavel Anatolyevich[4].

[1] Department of General Chemistry, Belarussian State Medical University, Belarus, Minsk, Dzerzinskogo, 83; [2] Laboratory of HIV/AIDS Diagnostic, Republican Center of Hygiene, Epidemiology and Public Health, Belarus, Minsk, Kazinca, 50; [3] Immunoprophylaxis Laboratory, Republican Research & Practical Centre for Epidemiology and Microbiology, Belarus, Minsk, Filimonova, 23; [4] Laboratory of Biotechnology and Immunodiagnostics of Highly Dangerous Infections, Republican Research & Practical Centre for Epidemiology and Microbiology, Belarus, Minsk, Filimonova, 23.

Received: 14 April 2011 **Accepted:** 10 June 2011 **Available Online**: 13 June 2011

ABSTRACT

There is a need for the method which helps to choose the less mutable immunogenic determinant for the design of recombinant or synthetic vaccines and ELISA test-systems. Our method based on the directional mutational pressure theory includes two steps: estimation of symmetric and asymmetric mutational pressure directions in a gene coding for a protein of interest; and selection of regions coding for its immunogenic determinants which are less prone to missense mutation occurrence and so, to immune escaping. Three original computer algorithms ("VVK Sliding Window", "VVK VarInvar" and "VVK Protective Buffer" available via www.barkovsky.hotmail.ru) have been created to perform all the necessary calculations and tests. "VVK Sliding Window" calculates nucleotide usage in fourfold and twofold degenerated sites, as well as usage of missense, nonsense and synonymous sites for each kind of nucleotide mutation along the length of a coding region, while "VVK Protective Buffer" calculates those indexes in a set of sequences. "VVK VarInvar" calculates percentage of variable sites in a set of aligned sequences, as well as nucleotide usage in invariable sites. Our method has been tested on HIV1 gp120 protein and on diphtheria toxin. The less mutable epitopes have been found for both proteins. Finally, it has been shown that antibodies recognizing the less mutable epitope of gp120 can be found in 80.22% of HIV1-infected persons.

Keywords: Mutational pressure; Sequence analysis; B-cell epitopes; HIV1 vaccine; gp120; Diphtheria toxin.

Abbreviations

G4f; C4f; A4f; T4f – usage of guanine, cytosine, adenine and thymine, respectively, in fourfold degenerated sites; **G2f3p; C2f3p; A2f3p; T2f3p –** usage of guanine, cytosine, adenine and thymine, respectively, in twofold degenerated sites from third codon positions.

1. Introduction

Recombinant and synthetic vaccines are thought to be less dangerous than life attenuated vaccines [1]. In case of immunization with the most conserved immunogenic determinant (determinants) protective immunity will be formed against the most of the strains from the whole population of the pathogenic microorganism (against those strains among which the immunogenic determinant is conserved). So, determination of the less variable and the less mutable immunogenic determinant (B-cell epitope or T-cell epitope) is one of the most important steps in recombinant or synthetic vaccine design [2].

It is possible to determine those parts of a coding region

***Corresponding author:** Vladislav Victorovich Khrustalev. Address: Belarus, Minsk, 220029, Communisticheskaya 7-24; Telephone: 80172845957; E-mail Address: vvkhrustalev@mail.ru

which are less prone to missense mutation occurrence then other ones with the help of our method based on the mutational pressure theory [3].

If the number of variable sites between sequences of the protein of interest is low, the boarders of conserved and variable regions cannot be determined, while levels of mutability can still be estimated [2]. The information on mutability of regions coding for immunogenic determinants is important even in case if the volume of data on amino acid substitutions in the protein of interest is sufficient. Immunization against the most conserved and the less mutable epitope (epitopes) of the pathogenic microorganism should be more efficient than immunization against the conserved and, at the same time, mutable epitope (epitopes) [2].

Envelope glycoprotein gp120 from Human immunodeficiency virus type 1 (HIV1) has been chosen to represent benefits of our method for the protein with sufficient amount of data on its variability. Diphtheria toxin has been chosen to represent benefits of our method for the protein with low number of available sequences.

3D B-cell epitopes have already been mapped on the surface of HIV1 gp120 in our previous work [2], while levels of their mutability have been determined by the less accurate method than that described in the present work. The less mutable and the most conserved predicted 3D B-cell epitope of HIV1 gp120 [2] was shown to be immunogenic in the present work. Antibodies to that predicted conformational B-cell epitope of gp120 can be found in 80.22% of HIV-infected persons using ELISA test-system with synthetic peptide (NQ21) corresponding to its consensus sequence. The fact that IgG from the serum of HIV1-positive person is able to cross-react with the NQ21 peptide has been confirmed by affinity purification.

Diphtheria toxin is the product of a gene (*tox*) which can be found in genomes of lysogenic corynephages infecting *Corynebacterium diphtheriae* and *Corynebacterium ulcerans* [4]. The toxin is lethal for susceptible persons in doses of 100 ng/kg [5]. In 1929, Ramon demonstrated that diphtheria toxin can be turned to its nontoxic, but antigenic equivalent (toxoid) by formaldehyde. The main disadvantage of diphtheria toxoid usage is the complicated procedure of its production [6].

Although diphtheria toxin is thought to be quite conserved protein [7], there are 54 variable amino acid sites in seventeen sequences studied. These data are not enough to separate conserved regions from variable ones for a sequence of 560 amino acids in length.

Our method requires the usage of three original computer algorithms available via our web page (www.barkovsky.hotmail.ru). These relatively simple algorithms included in MS Excel spreadsheets are described in our SciTopics Page [8].

The first step of the method is in the estimation of mutational pressure direction. The second step is in the comparison of nucleotide usage indexes in regions coding for immunogenic determinants with the aim to choose the less mutable one.

Symmetric mutational GC-pressure exists when the rates of AT to GC nucleotide mutations occurrence are higher then the rates of GC to AT nucleotide mutations occurrence [3]. Symmetric mutational AT-pressure exists when the rates of GC to AT mutations occurrence are higher then the rates of AT to GC mutations occurrence [3]. Those types of mutations which occur more frequently are fixed by random genetic drift in synonymous sites more frequently then those types of mutations which occur rarely [3]. As long as mutational GC-pressure leads to the almost complete saturation of synonymous sites with G and C, while mutational AT-pressure leads to the almost complete saturation of synonymous sites with A and T, it finally leaves no substrate for synonymous nucleotide mutations [9].

Directional mutational GC-pressure increases levels of usage of those amino acid residues which are encoded by GC-rich codons [9, 10]. Acrophilic (prone to be located on a surface of a protein [11]) proline and glycine are among those four amino acid residues [10]. Such strongly hydrophobic amino acid residues as isoleucine, phenylalanine, tyrosine and methionine [12] are encoded by GC-poor codons. Indeed, according to our computer simulations with BepiPred 1.0 algorithm [13], mutational GC-pressure frequently leads to formation of new linear B-cell epitopes and elongation of previously existing ones [9], while mutational AT-pressure frequently leads to disappearance of epitopes or their parts from the surface of proteins [14]. Percent of highly immunogenic amino acid residues forming linear and discontinuous B-cell epitopes is usually higher for homologous proteins encoded by GC-rich genes [15].

There may be significant symmetric mutational bias in twofold degenerated and fourfold degenerated sites even in coding genomes with 3GC levels close to 50% [9]. Asymmetric mutational pressure exists due to the differences in rates of occurrence of different types of mutations in leading and lagging strands of DNA, as well as due to the differences in rates of occurrence of different types of mutations in transcribed and nontranscribed strands of DNA [16, 17]. Even in genomes of prokaryotic organisms and viruses direction of symmetric mutational pressure may be different for different genes [9].

"VVK Sliding Window" and "VVK VarInvar" algorithms were designed for estimation of both symmetric and asymmetric components of mutational pressure, as well as for separation of biases in rates of transitions from biases in rates of transversions [8].

"VVK Protective Buffer" algorithm was designed for fast calculation of those indexes characterizing the usage of nucleotides prone to frequent mutations [8] which are used during selection of the region coding for the less mutable epitope.

2. Material and Methods

2.1 *Description of nucleotide sequences used*

Thirty four sets of *env* gene sequences coding for HIV1 gp120 protein have been used as a material. The total number of sequences is equal to 689. Each set of sequences has been obtained from a single HIV1-infected person. Codes of infected persons and GenBank accession numbers of *env* gene sequences are listed below. H1 – H5: **EU743973 – EU744175** [18]; S33 and S35: **EU604549 – EU604642** [19]; DM1 – DM9: **EF575363 – EF575486**; C61, C62, C93, C94, C96, C98, C109, ES2, ES4, ES7 – ES9: **DQ410040 – DQ410649** [20]; 9F, 605F, 605M, 32M, 183M and 120F: **EU852934 – EU853141** [21].

For a second example of the application of our method nine nucleotide sequences coding for *Corynebacterium diphtheria* toxin have been used. Three sequences were from integrated corynebacteriophages: one of them was from the reference genome of *Corynebacterium diphtheriae* (**NP_938615.1**), two of them were from other strains of *Corynebacterium diphtheriae* (**AJ576101.1**; **AY820132.1**). Six sequences were from different corynebacteriophages: four were from *Corynebacteriophage beta* (**K01722.1**; **EU069362.1**; **K01723.1**; **D78299**), one was from *Corynebacteriophage omega* (**V01536.1**) and the last one was from unclassified Corynebacteriophage (**X00703.1**). Eight sequences of *tox* gene from integrated *Corynebacterium ulcerans* phages (**AB304279.1**; **FJ858272.1**; **AB498872.1**; **AB304280.1**; **AB304278.1**; **AY141014.1**; **AY703827.1**; **AY141013.1**) have also been used in the study.

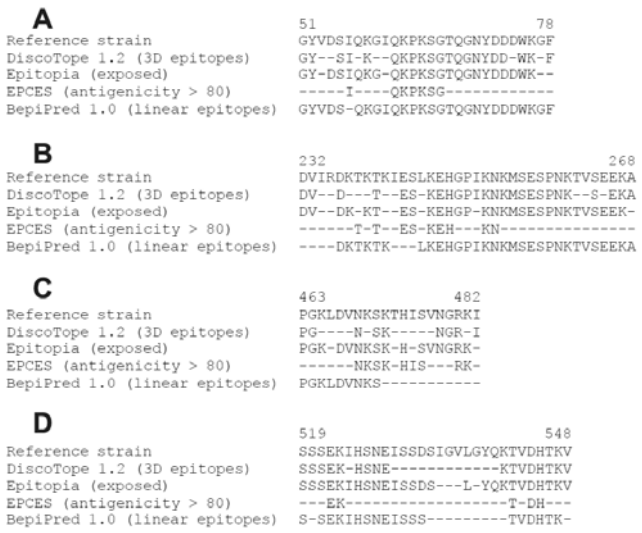

Figure 1. Amino acid sequences of four diphtheria toxin regions with predicted B-cell epitopes. Amino acid residues included in 3D B-cell epitopes by DiscoTope 1.2 algorithm, exposed amino acid residues (according to Epitopia prediction), highly antigenic amino acid residues (according to Epces algorithm prediction) and linear B-cell epitopes (according to BepiPred 1.0 prediction) are designated.

2.2 *Algorithms for 3D and linear B-cell epitopes prediction*

DiscoTope 1.2 algorithm [22] has been used to map 3D epitopes on the structure of diphtheria toxin deposited in PDB database (its accession number is "**1SGK**"). Then these results have been confirmed with the help of Epitopia [23] and Epces [24] algorithms. Sequences of the four most immunogenic regions of the diphtheria toxin are represented in Figure 1. For example, in Figure 1A amino acid sequence of the toxin from reference strain is written in the first line, amino acid residues included in 3D epitopes by DiscoTope 1.2 are written in the second line, residues which are exposed to solvent according to Epitopia prediction are written in the third line, residues with very high antigenicity score (according to Epces results) are written in the fourth line.

It has to be noted that there are always corresponding linear B-cell epitopes predicted by BepiPred 1.0 algorithm [13] for every 3D epitope from diphtheria toxin (see fifth lines in Figures 1A – 1D).

Seventeen sequences coding for *Corynebacterium diphtheria* and *Corynebacterium ulcerans* toxin have been aligned with sequences of its four most immunogenic regions (3D epitopes 1, 2, 3 and 4, respectively). Then four separate alignments of sequences (each of them is coding for one of the four epitopes) have been made to perform calculations in them.

Five most immunogenic regions of HIV1 gp120 protein have been predicted by us in the previous work [2] with the help of DiscoTope 1.2 [22], Epitopia [23] and Epces [24] algorithms.

2.3 *Original computer algorithms*

"VVK Siding Window" algorithm (which is a new version of "VVK in length" algorithm [25]) calculates nucleotide usage in fourfold and twofold degenerated sites along the length of a coding region in sliding windows, as well as probabilities to be synonymous, missense and nonsense for every type of nucleotide mutation [8]. These kinds of analyses help to determine main directions of mutational pressure and to perform screening test to find the less mutable regions. The algorithm requires a single nucleotide sequence as an input.

"VVK VarInvar" algorithm (which is a new version of "VVK Consensus" algorithm [25]) calculates the percentage of variable sites in a set of aligned sequences [8]. This algorithm requires alignment with at least 109 variable sites to give a reproducible percentage (see section 5.3). Relative frequencies of different types of nucleotide mutations can be estimated dealing with the percentage of variable sites. Average levels of nucleotide usage in third codon positions should be compared with nucleotide usage in invariable sites from third codon positions to confirm directions of those mutations [8]. The average usage of the "stable" nucleotide should be increased in invariable sites from third codon positions. The usage of "mutable" nucleotide should be decreased in invariable sites from third codon positions.

"VVK Protective buffer" algorithm (which is a new version of "VVK in group" algorithm [25]) calculates usage of synonymous, nonsense and missense sites for each kind of

nucleotide mutation in the set of sequences (up to 50 sequences can be used as an input) [8].

2.4 *Peptide synthesis*

Peptides of 21 amino acid residues in length have been synthesized for us by "Peptide 2.0 Inc." company. The "Symphony" solid-phase peptide synthesizer ("Protein Technologies, Inc.") has been used. The purity of both peptides was higher than 95% (according to the results of HPLC analyses). One of those peptides has been conjugated with biotin (via its N-terminal). Amino acid sequences of both peptides are the same as the consensus sequence for 3D epitope 1 from gp120 protein [2].

2.5 *Modification of ELISA test-system*

Commercial "ELISA-Recombinant-HIV 1,2" test-system produced by "PharmLand LLC" company (http://www.pharmland.by/en/products/elisa-recombinant-hiv-1) has been modified by us. This test-system includes a 96-well plaque with adsorbed recombinant gp160 and gp140 proteins, solution of recombinant gp160 (from HIV1) and gp140 (from HIV2) proteins conjugated with biotin and standard components, such as solutions of streptavidin conjugated with horseradish peroxidase, TMB, hydrogen dioxide and sulfuric acid, as well as wash solution. Solution of recombinant gp160 and gp140 conjugated with biotin from this test-system was replaced with the solution of the peptide NQ21 conjugated with biotin (0.04 mg/ml in 0.1M PBS with pH 7.4). The protocol of the analysis was as follows: 0.05 ml of the peptide biotin-NQ21 solution and 0.1 ml of serum have been simultaneously added to each well (0.002 mg of the peptide per well) and incubated for 30 minutes (37°C). During this period of time asymmetric sandwiches (epitope 1 of gp160 – antibody – biotin-NQ21) have been formed in certain wells. The rest of the protocol was standard (according to manufacturer's protocol): 7 washes; incubation with streptavidin conjugated with horseradish peroxidase for 30 minutes (37°C); 7 washes; incubation with TMB and hydrogen dioxide during 15 minutes (37°C); addition of sulfuric acid and measurement of optical density by the automatic 96-well plaque spectrophotometer (at the wavelength of 450 nm).

Serums from 91 persons with recently revealed HIV1-infection have been tested. Their diagnosis has been confirmed by two different ELISA test-systems and by immunobloting in the Laboratory of HIV/AIDS Diagnostic of the Republican Center of Hygiene, Epidemiology and Public Health (Minsk, Belarus). All of those persons did not receive antiretroviral therapy. Serums from 234 HIV-negative persons have also been tested.

To separate positive and negative results in the modified ELISA test-system the following calculations were performed. Average level of optical density (OD) for all the wells with serums from HIV-negative persons has been calculated for each of the four 96-well plaques, as well as the standard deviation. All the levels of OD higher than three standard deviations were considered to be positive (in a given plaque). In case of parallel analysis of a single serum (in two wells) an average OD for those two wells was calculated. For each of the four plaques an average level of OD for serums from HIV1-positive persons was significantly higher than that for serums from HIV-negative persons (according to the results of t-test).

2.6 *Affinity purification protocol*

The column from AminoLink Plus Immobilization Trial Kit ("Thermo scientific Inc.") was used to immobilize the peptide NQ21 via its NH_2 groups. There are two amine groups in the peptide NQ21: one is N-terminal and another one is from the side chain of lysine. Modified agarose contains aldehyde groups that react specifically with primary amines. After the spontaneous formation of semi-stable Schiff base bonds, reduction with sodium cyanoborohydride results in stable secondary amine bonds. Immobilization protocol (which can be found at http://www.piercenet.com/instructions/2160491.pdf) includes incubation with the peptide dissolved in pH 10 buffer during 4 hours; incubation with sodium cyanoborohydride dissolved in pH 7.4 PBS during 4 hours; incubation with quenching buffer (1M tris·HCl) during 30 minutes; incubation with sodium cyanoborohydride dissolved in pH 7.4 PBS during 30 minutes and washing the column with wash buffer (1M NaCl). About 2.3 mg of NQ21 have been finally immobilized.

1.5 ml of the serum from HIV1-infected person dissolved in 0.5 ml of PBS pH 7.4 has been added to the column with immobilized peptide NQ21 and incubated during 60 minutes. Then the column has been washed 11 times with PBS pH 7.4 (2 ml of PBS for each wash). Acetic acid (1M) has been used as the eluent. This procedure has been repeated for 1.5 ml of serum from HIV-negative person dissolved in 0.5 ml of PBS pH 7.4. Concentration of proteins in eluates has been estimated with the help of Hitachi 650-60 spectrofluorometer (excitation wavelength was equal to 296 nm, emission wavelength was equal to 345 nm).

2.7 *SDS-PAGE analysis*

Eluates 1, 3 and 7 obtained during affinity purification of the serum from HIV1-infected person were analyzed in SDS-PAGE. Composition of the sample buffer was as follows: 1M Tris·HCl, pH 6.8; 4% SDS; 2% 2-mercaptoethanol; 20% glycerol; 0.02% bromophenol blue. Samples were heated for 2 minutes at 95°C in a water bath. Three dilutions of each eluate have been prepared (1:1; 1:2 and 1:4). PageRuler™ Unstained Protein Ladder molecular mass markers from "Fermentas Life Sciences" were used.

3. Results

3.1 *Nucleotide usage biases in fourfold and twofold degenerated sites from third codon positions along the length of a coding region*

The usage of thymine in twofold degenerated sites from third codon positions (T2f3p) is always higher than the usage of cytosine in them (C2f3p) along the length of the region of HIV1 *env* gene coding for gp120 (see Figure 2A). The usage of adenine in twofold degenerated sites from third codon positions (A2f3p) is higher than the usage of guanine (G2f3p) in the most of "sliding windows" from the above-mentioned coding region.

The difference between adenine usage in fourfold degenerated sites (A4f) and cytosine usage in them (C4f) is very high (see Figure 2C), especially in comparison with the difference between thymine usage (T4f) and guanine usage (G4f) in those sites of the region of *env* gene coding for gp120 (see Figure 2D).

The level of C2f3p is significantly lower than the level of T2f3f along the most of the length of *tox* gene (see Figure 3A). The level of G2f3p is significantly lower than the level of A2f3p along the most of the length of *tox* gene (see Figure 3B).

The level of A4f is much higher than that of C4f in the most of "sliding windows" along the length of *tox* gene (see Figure 3C). The level of T4f is much higher than that of G4f (see Figure 3D). Moreover, T4f usage is significantly higher than A4f usage (average paired difference in t-test is equal to 0.158±0.006, P<0.001). This should be the evidence that bias between G to T and T to G transversions have been stronger than the bias between C to A and A to C transversions in *tox* gene, unlike those in the region of HIV1 *env* gene coding for gp120.

Taking together, data represented in Figures 2 – 3 approve that mutational AT-pressure in both *env* and *tox* genes is caused by both GC to AT transitions and GC to AT transversions. However, the rates of C to A transversions in HIV1 *env* gene should be higher than the rates of G to T transversions. In the gene coding for diphtheria toxin the rates of C to A transversions should be lower than the rates of G to T transversions.

According to the results of codon-based Z-test for positive selection, nonsynonymous distance (calculated by Kumar method) [26] is significantly higher than synonymous distance (dN>dS) between two sequences of *tox* gene from *Corynebacterium ulcerans* phages (for those with the following GenBank identifiers: **AY703827.1** and **FJ858272.1**). Indeed, there are three amino acid replacements between them, while the number of synonymous nucleotide mutations is equal to zero. Two of amino acid substitutions are specific to **AY703827.1** sequence. One of them (T262I) took place in Epitope 2. This substitution caused by a single C to T transition resulted in the replacement of relatively acrophilic and hydrophilic threonine [11] with strongly hydro-

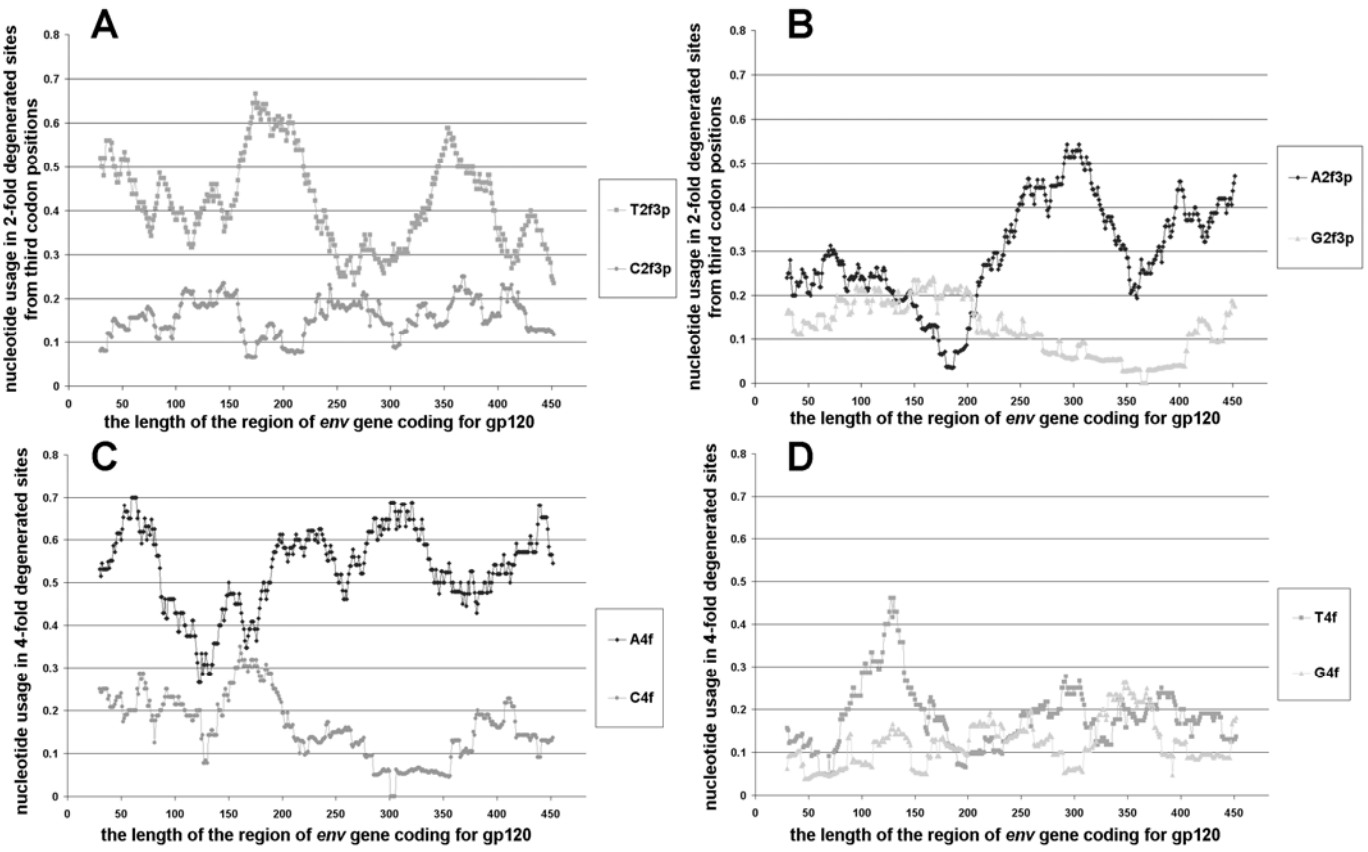

Figure 2. Thymine and cytosine (A) and adenine and guanine (B) usage in twofold degenerated sites from third codon positions along the length of the region of *env* gene coding for gp120. Adenine and cytosine (C) and thymine and guanine (D) usage in fourfold degenerated sites along the length of the *tox* gene. The length of sliding window is equal to 70 codons.

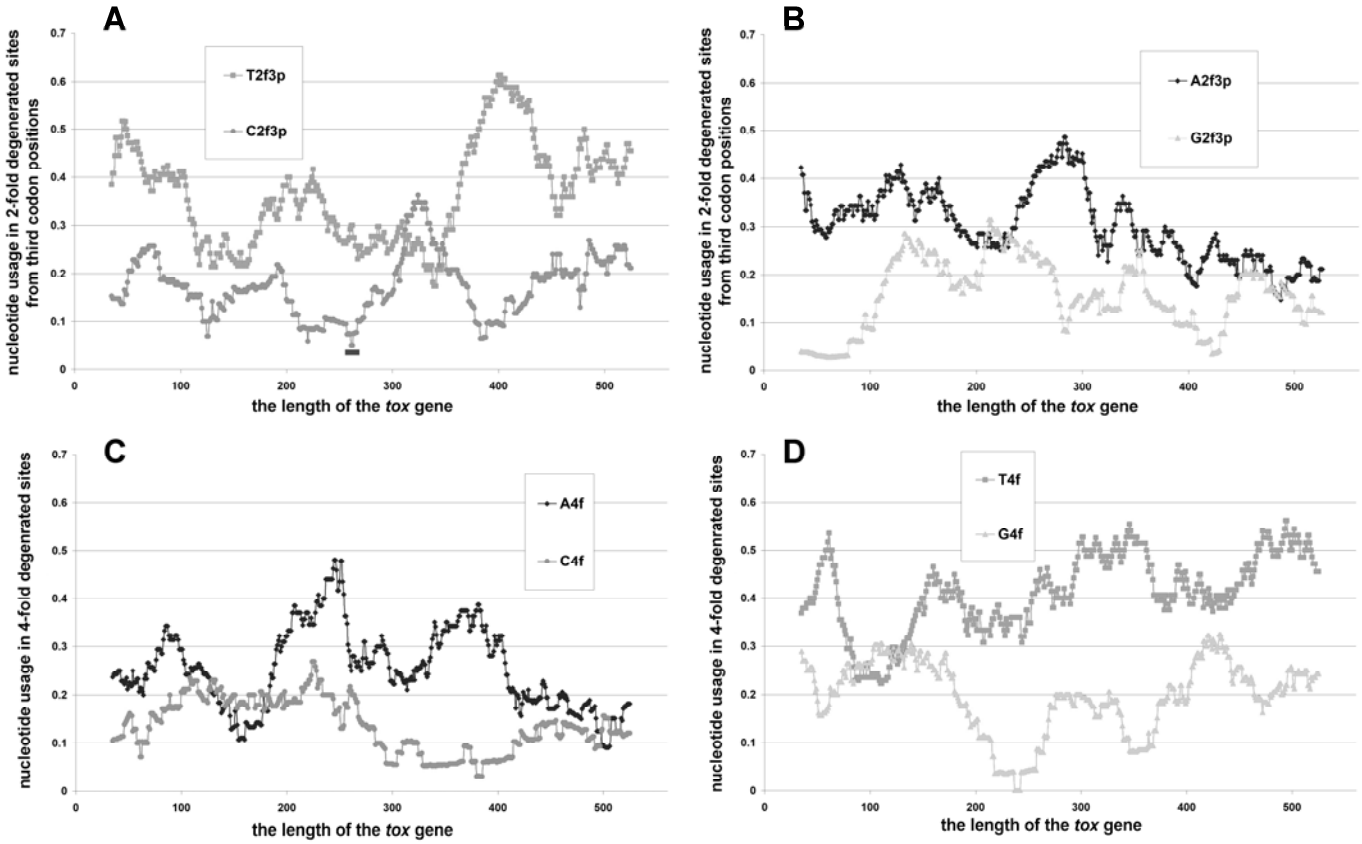

Figure 3. Thymine and cytosine (A) and adenine and guanine (B) usage in twofold degenerated sites from third codon positions along the length of the *tox* gene. Adenine and cytosine (C) and thymine and guanine (D) usage in fourfold degenerated sites along the length of the *tox* gene. The length of sliding window is equal to 70 codons. The region with the minimum of C2f3p usage is designated.

phobic isoleucine [12]. This substitution might lead to the loss of affinity of previously synthesized antibodies to this epitope, since scores of thirteen amino acid residues surrounding T262I to be included in linear B-cell epitope according to BepiPred 1.0 prediction, became lower than those in chain B of the reference strain, while they are still above the threshold.

Interestingly, the region of 73 codons in length surrounding the codon number 262 has the lowest level of C2f3p (see Figure 3A). This fact can be interpreted as yet another reason to estimate probabilities of synonymous mutation occurrence in regions coding for B-cell epitopes. Moreover, this fact is one more evidence that low probability of synonymous mutation occurrence (due to the strong mutational pressure) may sometimes lead to the situation when dN becomes significantly higher than dS.

3.2 *Distribution of synonymous, missense and nonsense sites for different types of mutations along the length of a gene*

Nonsynonymous sites can be divided into two groups: missense sites and nonsense sites. Nucleotide mutation in missense site (for a given type of nucleotide mutation) leads to the amino acid replacement. Nucleotide mutation in nonsense site leads to the formation of stop-codon.

There are three stop-codons in the universal genetic code (TAA, TAG and TGA). All of them are relatively GC-poor. That is why mutations of GC to AT direction have much higher probability to be nonsense than mutations of AT to GC direction [27]. Nonsense mutation leads to the translation of truncated protein. A protein may lose its C-terminal (if nonsense mutation occurred near the end of open reading frame), or N-terminal (in case if nonsense mutation occurred near the beginning of open reading frame, translation may start from the alternative initiation codon). Truncated protein may still perform its function. In this case nonsense mutations may lead to the immune escaping even more likely and more efficiently than missense mutations. In case if nonsense mutation leads to the complete loss of function for the given protein, one may not be afraid of immune escaping due to this type of mutation.

Distribution of synonymous, missense and nonsense sites along the length of the region of *env* gene coding for gp120 from the reference HIV1 strain is quite variable (see Figure 4). The lowest level of missense sites for C to T transitions is characteristic to the region coding for 3D epitope 1 of gp120 (see Figure 4A). The highest level of the "protective buffer" against G to A transitions (it includes both synonymous and nonsense sites) is also characteristic to that region (see Figure 4B). The lowest level of missense sites for G to A transitions is characteristic to the region coding for 3D epitope 3 of gp120, while highest levels are characteristic to regions

coding for 3D epitopes 5 and 2 (see Figure 4B). The region coding for 3D epitope 1 seems to be well protected from C to A transversions because of the low level of cytosine usage (see Figure 4C). Interestingly, the level of nonsense sites for G to T transversions is much higher than the level of nonsense sites for C to A transversions in both *env* (see Figures 4C and 4D) and *tox* genes (see Figures 5C and 5D), as well as in the most of bacterial coding regions [17].

Taking into account that the most frequent types of nucleotide mutations in HIV1 *env* gene are G to A transitions, as well as C to T transitions and C to A transversions, the first 3D epitope (from C1 region) was considered to be the less mutable 3D epitope of gp120 [2]. This statement has been made after the calculation of cytosine and guanine content in first, second and third codon positions of regions coding for gp120 3D epitopes. The level of "protective buffer" in that study was measured as the nucleotide usage in third codon positions. In the present work that level was calculated as accurate as it possible (for each type of GC to AT nucleotide mutation).

The region coding for 3D epitope 1 from the diphtheria toxin is well protected from missense C to T transitions (see Figure 5A), while levels of "protective buffer" against G to A (see Figure 5B), C to A (see Figure 5C) and G to T (see Figure 5D) mutations are too low for that region. There are no synonymous sites for G to T transversions in the region coding for 3D epitope 2 of the diphtheria toxin (see Figure D). Regions coding for 3D epitopes 3 and 4 seem to be less mutable than those coding for 3D epitopes 1 and 2 of the diphtheria toxin.

3.3 *Percentage of different types of variable sites in sets of aligned sequences*

The information on the percentage of different types of variable sites in the alignment of homologous sequences helps to estimate relative frequencies of different types of mutations.

In Figure 6A average percentage of different variable sites for 34 sets of regions coding for gp120 is shown. The most frequent type of variable site among HIV1 *env* gene sequences contains adenine and guanine. The type of variable site containing thymine and cytosine is approximately two times less frequent among the sequences studied. The third place belongs to the type of variable site containing cytosine and adenine. The percent of sites containing cytosine and adenine is approximately two times higher than the percent of sites containing guanine and thymine.

In general, information given in Figure 6A is consistent with data from Figure 2. Indeed, C to A transversions should

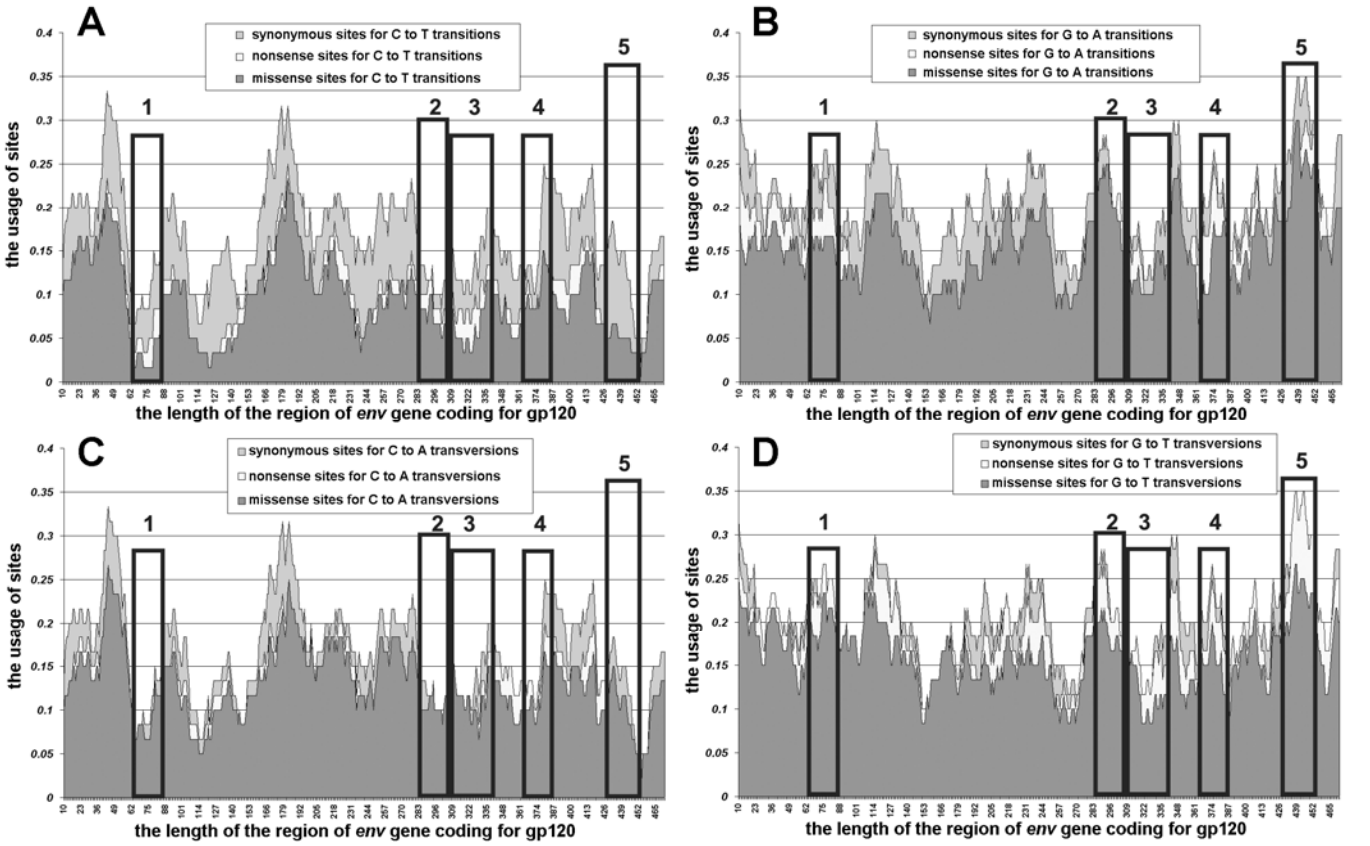

Figure 4. Usage of synonymous, nonsense and missense sites for C to T transitions (A), G to A transitions (B), C to A transversions (C) and G to T transversions (D) along the length of the region of *env* gene coding for gp120. The length of sliding window is equal to 70 codons. Regions coding for 3D B-cell epitopes are designated by numbered boxes.

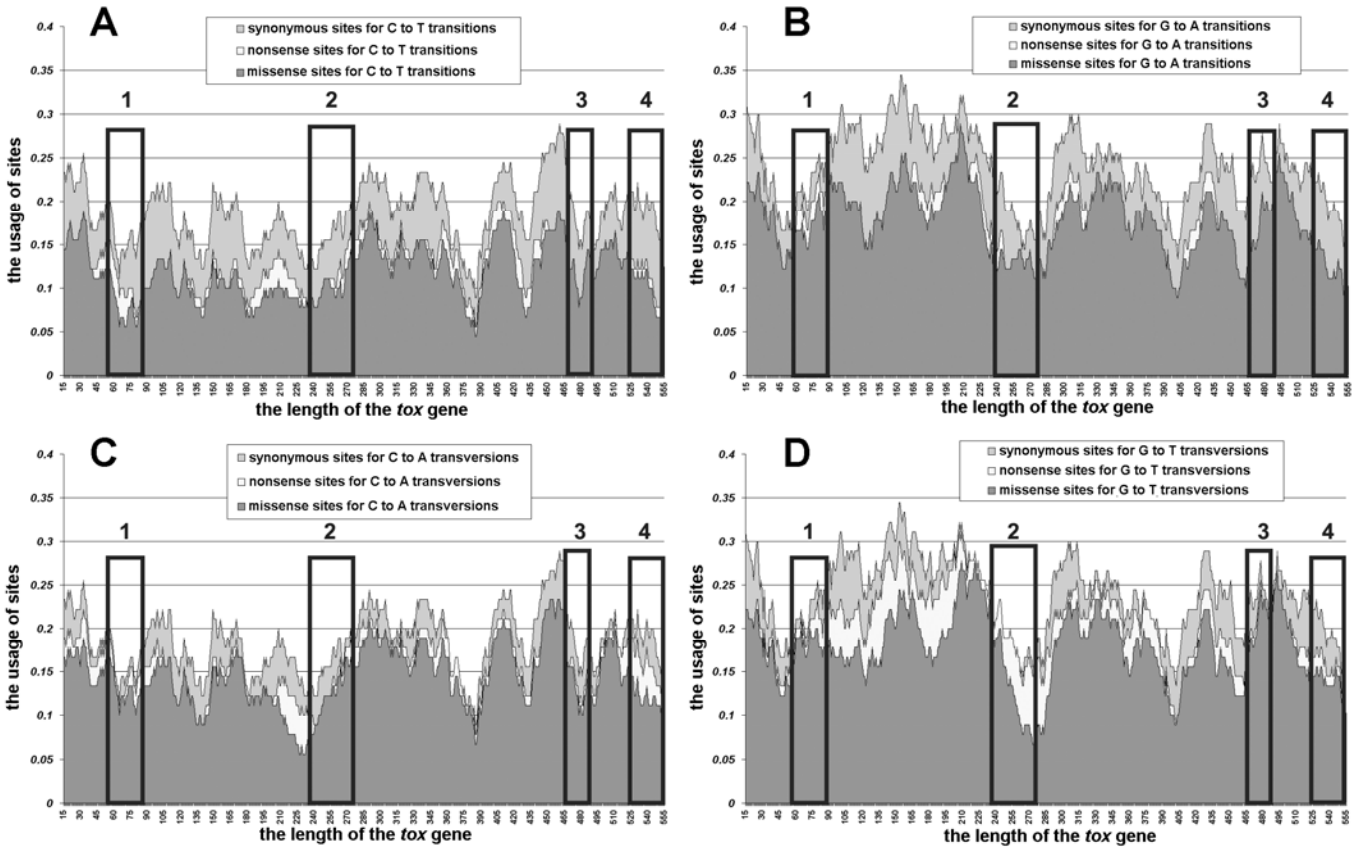

Figure 5. Usage of synonymous, nonsense and missense sites for C to T transitions (A), G to A transitions (B), C to A transversions (C) and G to T transversions (D) along the length of the *tox* gene. The length of sliding window is equal to 70 codons. Regions coding for 3D B-cell epitopes are designated by numbered boxes.

occur more frequently than G to T transversions. However, it is impossible to suggest that G to A transitions occur more frequently than C to T transitions during analysis of nucleotide usage biases in twofold degenerated sites (see Figures 2A and 2B).

The percentage given in Figure 6A can be reproduced from the most of the small groups of sequences. In all the 34 groups of sequences the percent of G↔A sites is higher than the percent of C↔T sites. In 33 from 34 groups of sequences the percent of C↔A sites is higher than the percent of G↔T sites. In the group of HIV1 *env* gene sequences from the patient designated as "ES2" (see Supplementary Material, Table 1) the percent of G↔T sites is higher than the percent of C↔A sites. This deviation is the consequence of the low number of variable sites (64) among sequences from that group. The number of transversions in that set is just 11.

The main requirement for appropriate work of "VVK VarInvar" algorithm is not the sufficient number of sequences but the sufficient number of variable sites between them.

There are 6 from 34 groups of sequences in which the percent of variable C↔A sites is higher than the percent of variable C↔T sites. This kind of deviation should not be the consequence of the low number of variable sites (see Supplementary Material, Table 1).

About 75% of variable sites (see Figure 6B) from the alignment of seventeen sequences coding for diphtheria toxin contain transitions (the total number of variable sites is equal to 141 in this alignment). The number of sites with A↔G transitions is approximately equal to the number of sites with T↔C transitions. Only 25% of variable sites contain transversions. Numbers of sites containing different types of trasversions are close to each other. This information may be interpreted as an evidence that transitions occur approximately three times more frequently in *tox* gene than transversions. More attention should be paid to the mutability of regions coding for diphtheria toxin epitopes under the pressure of GC to AT transitions than to their mutability under the pressure of GC to AT transversions.

3.4 *Nucleotide usage in invariable sites from sets of aligned sequences*

According to the mutational pressure theory [3], adenine and thymine should be more stable nucleotides in both *env* and *tox* genes than guanine and cytosine. Mutations of guanine and cytosine should happen more frequently than mutations of adenine and thymine.

Indeed, the usage of guanine in invariable sites from third codon positions is significantly lower than its average usage in those positions (see Figure 7). The same situation is char-

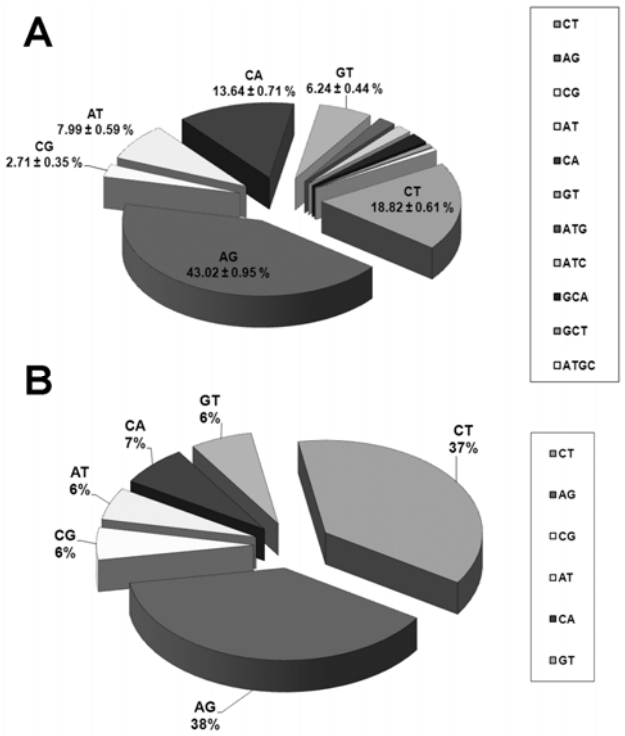

Figure 6. Average percentage of different types of variable sites for 34 alignments of sequences coding for gp120 (A); percentage of different types of variable sites between seventeen sequences coding for diphtheria toxin (B).

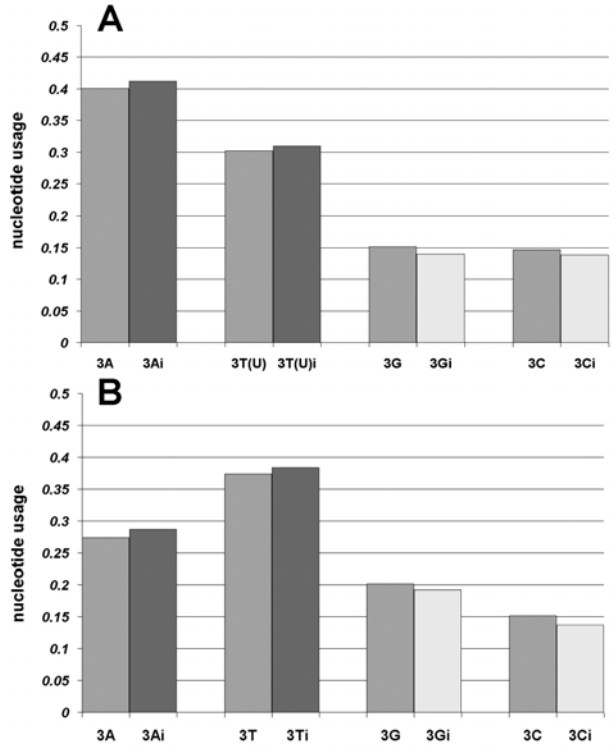

Figure 7. Average levels of nucleotide usage in third codon positions and nucleotide usage in invariable sites from third codon positions of sequences coding for gp120 (A) and diphtheria toxin (B).

acteristic for cytosine usage (see Figure 7). Levels of adenine and thymine in invariable sites from third codon positions are significantly higher than their average levels of usage in third codon positions. These facts approve that adenine and thymine are much more stable than guanine and cytosine in both *env* and *tox* genes. Average levels of nucleotide usage for 34 groups of sequences are given in Figure 7A. Differences between average nucleotide usage in third codon positions and nucleotide usage in invariable sites are significant, although there are a few exceptions from common tendencies in certain groups (see Supplementary Material, Table 2).

Levels of thymine and adenine are high in third codon positions. However, the number of invariable sites containing adenine and thymine is higher than the number of invariable sites containing guanine and cytosine. This situation is possible only in case of mutational AT-pressure existence. The change of the symmetric mutational pressure direction will lead to the opposite situation (adenine and thymine will become quite instable and the number of invariable sites containing them will decrease). As it has been shown previously [9], nucleotide usage indexes will remain the same for some time after the change of mutational pressure direction. So, the calculation of nucleotide content in invariable sites from third codon position is the most sensitive indicator of the recent change in mutational pressure direction.

3.5 *Probabilities to be missense, synonymous and nonsense for mutations of GC to AT direction, as well as amount of substrate for them, in regions coding for 3D B-cell epitopes of HIV1 gp120 and diphtheria toxin*

Mutability levels of regions coding for five most immunogenic regions of HIV1 gp120, as well as for four most immunogenic regions of diphtheria toxin, have been compared in this section. There are three criterions to compare (for each type of GC to AT mutation): i) amount of the substrate for nonsynonymous (or missense) mutation; ii) amount of the substrate for synonymous mutation and iii) probability to be synonymous (or synonymous or nonsense). The region with the lowest amount of the substrate for nonsynonymous (or missense) mutation, with the highest amount of the substrate for synonymous (or synonymous or nonsense) mutation and with the highest probability of synonymous (or synonymous or nonsense) mutation has the lowest level of mutability under the pressure of certain mutations.

Amount of the substrate for synonymous (or synonymous or nonsense) mutation is a kind of "protective buffer" against amino acid replacements caused by certain type of nucleotide mutations.

In case if protein loses its function completely due to nonsense mutation, amount of the substrate for nonsense mutations should also form "protective buffer" together with amount of the substrate for synonymous mutations.

Average (for 689 sequences) usage of missense sites for C to T transitions in the region coding for 3D epitope 1 of HIV1 gp120 is very low (see Figure 8A). The highest proba-

bility to be synonymous or nonsense for G to A transition is also characteristic for that region (see Figure 8B). Results of the comparisons between five epitopes are combined in the Table 1. The first criterion for comparison is the probability to be synonymous or nonsense for the given type of mutation; the second criterion is the usage of missense sites for the given type of mutation; and the third criterion is the usage of sites for synonymous and nonsense sites for the given type of mutation.

Regions coding for epitopes 1 and 3 seem to be less mutable than regions coding for epitopes 2, 4 and 5. However, the region coding for epitope 1 is better protected from missense transitions than from transversions, while the region coding for epitope 3 is better protected from missense transversions than from transitions (see Table 1).

The amount of guanine which is prone to missense mutation is much higher in a region coding for diphtheria toxin epitope 1 than in other three regions (see Figure 9B). The amount of guanine in synonymous and nonsense sites for G to A transitions is much lower in the region coding for Epitope 1 than in three other regions. Obviously, the lowest probability to be synonymous for G to A transition is characteristic for the region coding for Epitope 1.

As to the mutability under the pressure of C to T mutations, the region coding for epitope 1 is looking some better than three other regions (see Figure 9A and Table 2). The mutability of the region coding for epitope 2 is higher than that for the region coding for epitope 1: all three criterions show preference for the last one. However, regions coding for epitope 3 and epitope 4 have higher amount of "protective buffer" than the region coding for epitope 1.

The lowest amount of "protective buffer" against missense C to A transversions is characteristic to the region coding for epitope 1 (see Figure 9C). The highest probability to be synonymous or nonsense for C to A mutation is characteristic to the region coding for epitope 4, as well as the highest amount of the substrate for missense C to A mutations.

There is no substrate left for synonymous G to T transversions in the region coding for epitope 2 (see Figure 9D), and so its "protective buffer" is represented only by nonsense sites for G to T mutations. It is clear that the region coding for epitope 4 is the less mutable one under the pressure of G to T mutations (see Table 2), while regions coding for epitope 2 and epitope 3 are the most mutable ones.

The final conclusion is as follows: the region coding for epitope 4 is the less mutable one, since i) it is protected from missense G to T transversions better than other three ones, ii) it is protected from missense G to A transitions better

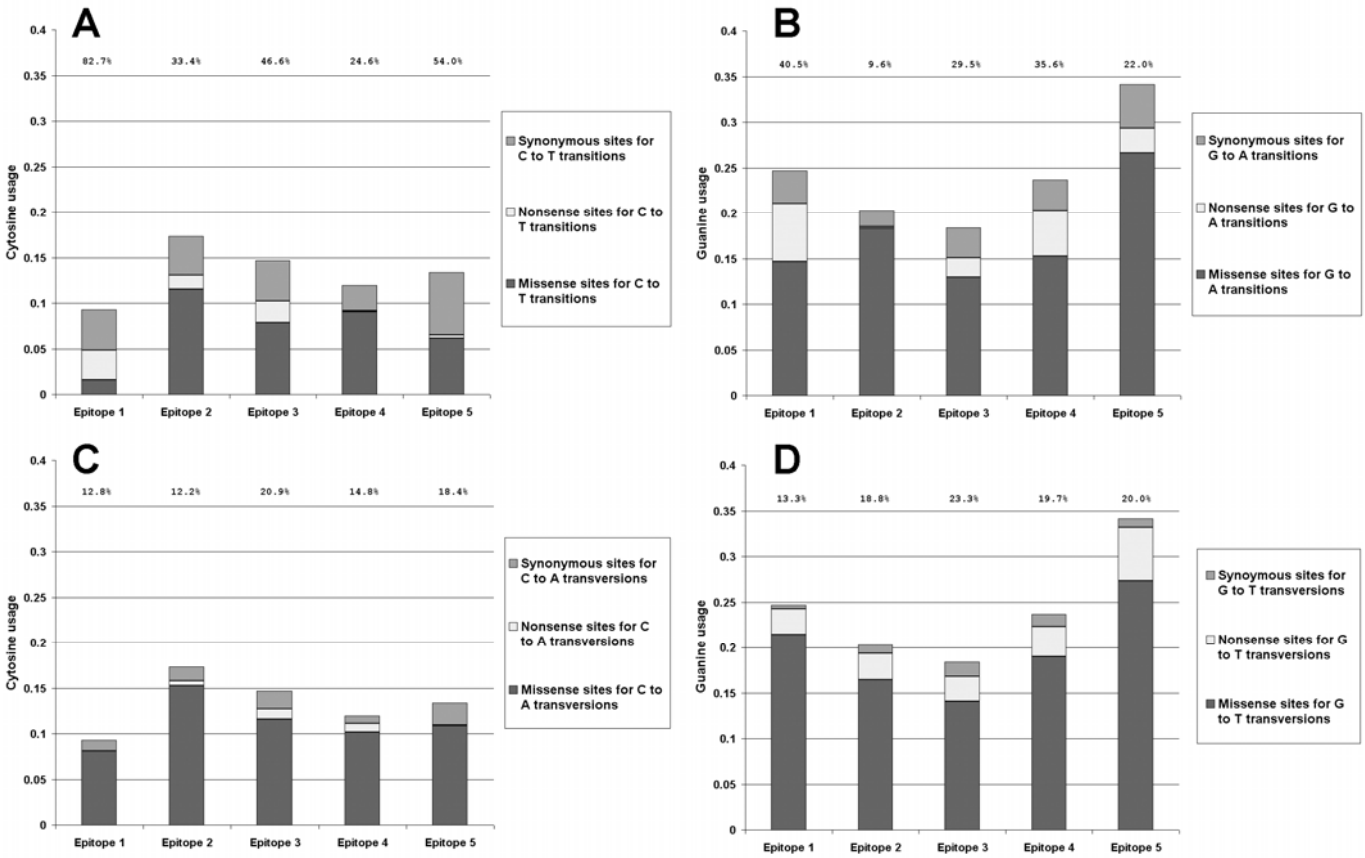

Figure 8. Average usage of synonymous, nonsense and missense sites for C to T transitions (A), G to A transitions (B), C to A transversions (C) and G to T transversions (D) in regions coding for five 3D B-cell epitopes of gp120. Probabilities to be synonymous or nonsense are written above the columns.

Table 1. Results of the application of our method to five regions of *env* gene coding for the most immunogenic 3D B-cell epitopes of gp120. For each type of mutation the region with the highest (according to the results of t-test) probability of synonymous or nonsense mutation is marked by the first "+", the region with the lowest level of missense sites is marked by the second "+", the region with the highest level of "protective buffer" (synonymous sites and nonsense sites) is marked by the third "+".

Regions coding for 3D epitopes	G to A		C to T			G to T		C to A	
epitope 1	+		+	+	+			+	
epitope 2									
epitope 3		+				+	+		+
epitope 4									
epitope 5							+		

than regions coding for epitope 1 and epitope 2, iii) it has the highest amount of "protective buffer" against nonsynonymous C to A transversions.

It has to be noted that there are no proline residues in consensus sequences of the epitope 1 of HIV1 gp120 and the epitope 4 of diphtheria toxin, while there are a few of them in other epitopes. Mutations of proline residues (especially those caused by C to T transitions) usually have a drastic effect on length of linear B-cell epitopes [14]. It means that epitope 1 of HIV1 gp120 and epitope 4 of diphtheria toxin

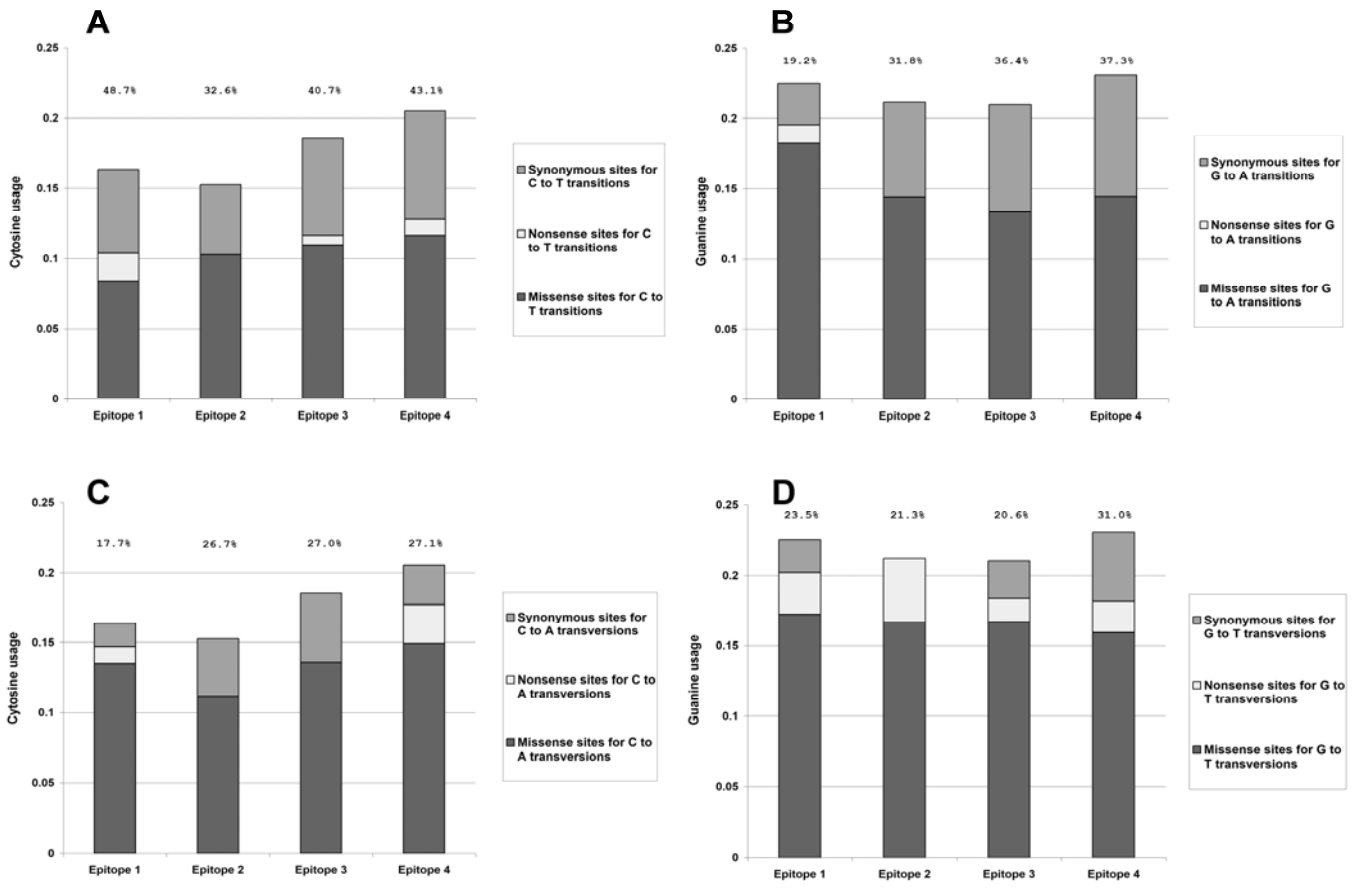

Figure 9. Average usage of synonymous, nonsense and missense sites for C to T transitions (A), G to A transitions (B), C to A transversions (C) and G to T transversions (D) in regions coding for four 3D B-cell epitopes of diphtheria toxin. Probabilities to be synonymous or nonsense are written above the columns.

Table 2. Results of the application of our method to four regions of *tox* gene coding for the most immunogenic 3D B-cell epitopes of diphtheria toxin. For each type of mutation the region with the highest (according to the results of t-test) probability of synonymous or nonsense mutation is marked by the first "+", the region with the lowest level of missense sites is marked by the second "+", the region with the highest level of "protective buffer" (synonymous sites and nonsense sites) is marked by the third "+".

Regions coding for 3D epitopes	G to A	C to T	G to T	C to A
epitope 1		+ +		
epitope 2				+
epitope 3	+		+	
epitope 4	+ +	+ + +	+ + +	+

are protected from those drastic effects simply because of the absence of proline.

3.6 *Variability of four 3D B-cell epitopes from HIV1 gp120*

689 amino acid sequences of HIV1 gp120 have been aligned with the help of PAM-matrix included in MEGA4 program [26]. 3D epitope 1 is quite conserved (see Table 3). There are nineteen amino acid residues (from twenty one) which can be found in a given site in more than 95% of sequences. Six of them are invariable. Either asparagine (81%) or aspartic acid (17%) can be found in the position 6; either glutamic acid (79%) or aspartic acid (18%) can be found in the position 9. These amino acid substitutions should have quite neutral consequences for the structure of the 3D epitope 1. Relatively radical amino acid substitution can be found only in the position 16. However, lysine instead of isoleucine has been found in this position only in one from 34 groups of sequences.

Analogous tables with percentage of amino acid substitutions have been created for four other 3D epitopes. Because of the high levels of variability those tables were included in Supplementary Material. There are fifteen amino acid residues (from thirty eight sites, including gaps) which can be found in a given site in more than 95% of 3D epitope 2 sequences (see Supplementary Material, Table 3). Actually, 3D epitope 2 is V3-loop, which is well-characterized immunogenic determinant of gp120 [2]. Relatively conserved part of V3-loop can be found in its N-terminal (CTRPNNNTR).

There are nine relatively conserved amino acid residues (from thirty six sites, including gaps) in 3D epitope 3. The most conserved motif from 3D epitope 3 (SSGGD) is situated in its C-terminal (see Supplementary Material, Table 4). This motif is actually a part of CD4 binding receptor of gp120 [28].

Interestingly, sequences corresponding to the epitope 4 cannot be aligned at all. There are no conserved amino acid residues in them. Gap can be found in each of the forty two

Table 3. Consensus sequence of the predicted 3D epitope 1 from gp120 of HIV1. Coordinates are given relatively to the length of gp160 and gp120.

94 / 66	*95 / 67*	96 / 68	*97 / 69*	*98 / 70*	*99 / 71*	*100 / 72*
N99,71% K0,14%; D0,14%;	**M97,54%** A2,17%; V0,29%;	**W100%**	**K96,96%** E1,59%; R1,16%; T0,29%;	**N100%**	N81,01% D16,52%; G2,17%; H0,29%;	**M99,71%** V0,29%;
101 / 73	*102 / 74*	103 / 75	*104 / 76*	*105 / 77*	*106 / 78*	*107 / 79*
V99,57% I0,29%; A0,14%;	E79,13% D17,54%; Q2,46%; N0,58%; K0,14%; G0,14%;	**Q100%**	**M99,86%** I0,14%;	**H97,54%** Q2,46%;	**E97,54%** G1,3%; T1,16%;	**D100%**
108 / 80	*109 / 81*	110 / 82	*111 / 83*	*112 / 84*	*113 / 85*	*114 / 86*
I96,52% V3,48%;	I93,91% K5,8%; V0,29%;	**S98,12%** N1,74%; G0,14%;	**L100%**	**W100%**	**D98,84%** E0,87%; N0,14%; G0,14%;	**Q97,25%** E2,75%;

sites of the alignment at least in one sequence (see Supplementary Material, Table 5).

There are thirteen conserved amino acid residues (from twenty nine sites, including gaps) in the alignment of 689 sequences corresponding to the 3D epitope 5 of gp120 (see Supplementary Material, Table 6). The most conserved (and relatively long) motif can be found in the C-terminal of that epitope (FRPGGGDMRDNWR).

In general, the less mutable epitope of HIV1 gp120 (3D epitope 1) is the less variable one. However, relatively conserved motif from the 3D epitope 5 seems to be highly mutable. This situation can be caused by the strong negative selection preventing fixation of amino acid substitutions in the C-terminal of the 3D epitope 5. Nonfunctional gp120 proteins containing amino acid mutations in the abovementioned part of the 3D epitope 5 should occur frequently. It means that mutability of the region coding for the 3D epitope 5 may still take part in the "deception" of the immune system.

3.7 *Modified ELISA test-system with the peptide NQ21 corresponding to the consensus sequence of 3D epitope 1 of gp120*

To confirm that the less mutable and the less variable 3D epitope of gp120 predicted by DiscoTope 1.2 is really immunogenic the peptide (NQ21) corresponding to its consensus sequence has been synthesized. Commercial ELISA test-system has been modified to check the presence of antibodies cross-reacting with the peptide NQ21 in serums of HIV1-infected persons. Peptide NQ21 conjugated with biotin via its N-terminal has been added to that ELISA test-system instead of biotinylated recombinant gp160 and gp140 proteins. Levels of optical density (OD) for wells with serums from 234 HIV-negative persons were not higher than 3 standard deviations from their average level for each from four plaques. Levels of OD higher than 3 standard deviations (relatively to the average level of OD for serums from HIV1-negative persons) have been found only in serums from HIV1-positive persons. The presence of antibodies recognizing 3D epitope 1 of the recombinant gp160 protein (adsorbed in plaques from the commercial ELISA test-system) which are able to cross-react with the peptide NQ21 have been confirmed in serums of 80,22% (in 73 from 91) HIV1-infected persons.

This data approves that predicted epitope 1 of gp120 is a real B-cell epitope. Moreover, this epitope is conserved enough to be included in polyvalent vaccines against HIV1.

3.8 *Affinity purification of antibodies cross-reacting with the peptide NQ21*

2.3 mg of the peptide NQ21 has been immobilized on the column from AminoLink Plus Immobilization Trial Kit ("Thermo scientific Inc."). Serum from HIV-negative person showed no cross-reactivity with immobilized peptide corresponding to the 3D epitope 1 of gp120 (see Figure 10A). There was a clear peak of fluorescence measured by Hitachi 650-60 spectrofluorometer in the third eluate for the serum from HIV1-positive person (see Figure 10B). According to the results of SDS-PAGE with 2-mercaptoethanol (see Figure 10C), there were four types of immunoglobulins in the third eluate. According to our results, IgG, IgM, IgA and IgE molecules can be synthesized against the epitope 1 of gp120.

4. Discussion

4.1 *Mutability and variability*

Our method has been created for estimation of mutability under the influence of mutational pressure. Some regions of a gene are usually more prone to missense mutations caused by mutational pressure than others. In our terminology this regions are "mutable" [2]. It means that missense nucleotide mutations will happen in them at a higher probability than in less mutable regions. However, mutable regions may theoretically be highly conserved [2]. In this case missense mutations will occur in them frequently, while all the amino acid replacements caused by them will be eliminated by natural selection.

In our previous study on HIV1 gp120 3D epitopes it has been found that its less mutable 3D epitope is the most conserved one [2]. Method which is able to predict level of mutability was used in that study [2], while in the present study improved and much more accurate method was introduced. It is impossible to check whether predicted 3D epitopes of diphtheria toxin are conserved or variable simply because of the small number of sequences and variable sites between them. That is why data obtained with the help of our method should be especially beneficial for studies on diphtheria toxin vaccine design.

4.2 *Experimental data on gp120 immunogenicity*

The statement that conserved regions of HIV1 gp120 are poorly immunogenic, while highly immunogenic regions of this protein are extremely variable can be found in many sources [29]. Several regions of gp120 that were shown to be recognized by monoclonal antibodies can be found in the HIV molecular immunology database (http://www.hiv.lanl.gov/content/immunology/tables/ab_summary.htm). Some of those regions include N-terminal, C-terminal or central part of the epitope 1 predicted by us, but never the whole immunogenic determinant. Some of those antibodies (recognizing the epitope which includes N-terminal of the NQ21) were even shown to be neutralizing [30]. However, antibodies to the C1 region of gp120 are thought to bind monomeric and not oligomeric protein [29]. From this point of view, epitopes from C1 region of gp120 were considered to be bad targets for protective immunity development [29].

Attention should be paid to the fact that the main part of 3D epitope 1 is presented by long alpha-helix (see Figure 11). This alpha-helix is amphiphilic: one half of it is hydrophobic, while another half is hydrophilic. Synthesis of smaller pep-

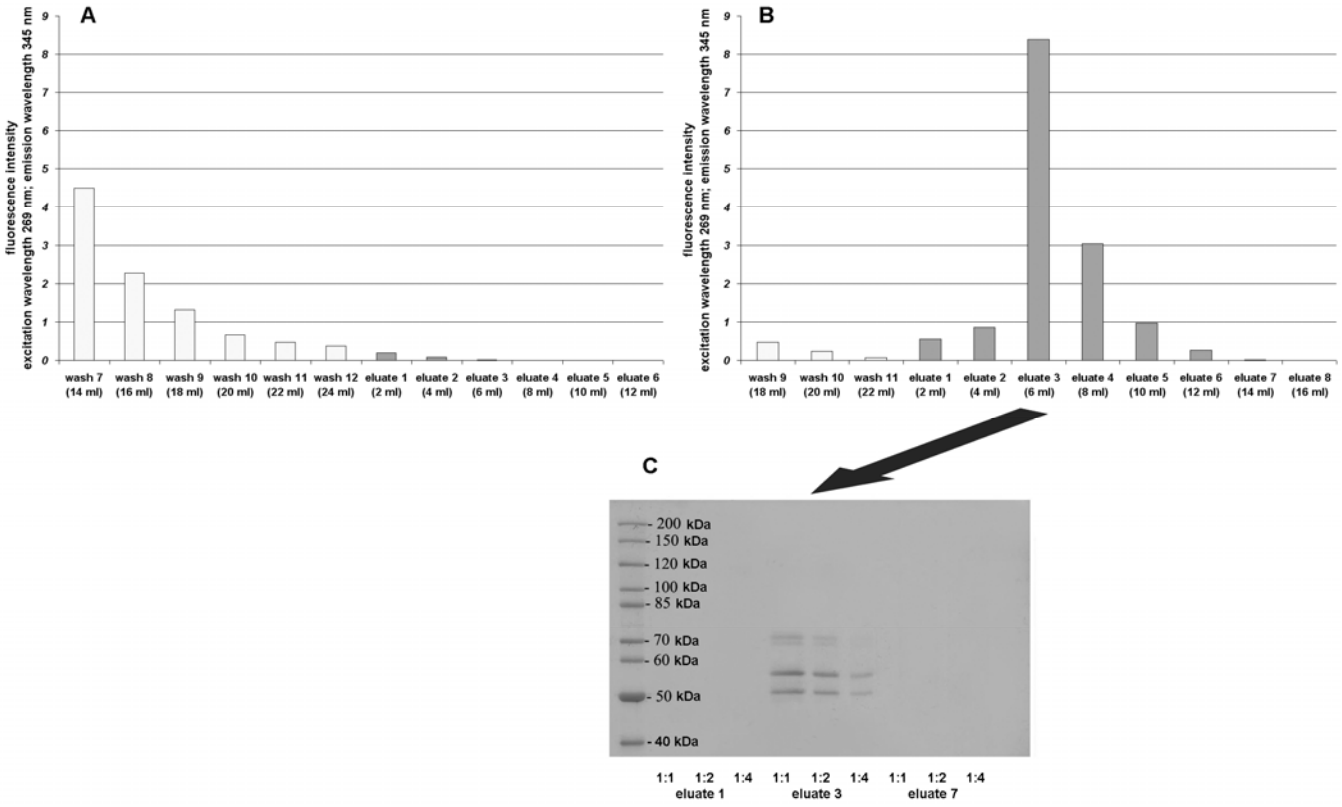

Figure 10. Intensity of fluorescence in washes and eluates collected during affinity purification of the serum from HIV1-negative person (A) and HIV1-positive person (B); results of SDS-PAGE analysis of eluates 1, 3 and 7 collected during affinity purification of the serum from HIV1-positive person (C).

tides might lead to the partial destruction of that alpha-helix and to the loss of their cross-reactivity with many types of antibodies recognizing the corresponding part of native molecule.

Neutralizing antibodies to gp120 (preventing interactions between gp120 and CD4 molecules) were shown to stimulate infection of macrophages [28]. Since macrophages are thought to be a good reservoir for HIV-infection, the idea of the creation of vaccine stimulating synthesis of neutralizing antibodies may seem to be compromised.

In the present study it has been shown that antibodies to the most immunogenic part of C1 region can be found in more than 80% of HIV1-infected persons. The fact that those antibodies were not found in serums of approximately 20% of HIV1-infected persons can be explained by at least three hypotheses. It is important to highlight that serums from persons with recently revealed HIV1-infection were tested in the present work. Probably, antibodies against gp120 3D epitope 1 in some of those persons have not been synthesized yet. In other words, those antibodies may not be synthesized in the yearly period of HIV1-infection in certain persons. On the other hand, antibodies against gp120 3D epitope 1 may not cross-react with the NQ21 peptide due to mutations in that region of viral protein. Even though 3D epitope 1 is the most conserved one, mutations disturbing cross-reactivity with NQ21 may still happen inside it. Once again it has to be noted that mutations should happen in 3D epitope 1 much less frequently than in other epitopes. Finally, antibodies to certain epitopes of gp160 may somehow disturb or totally prevent binding of antibodies against 3D epitope 1 to adsorbed molecules in our modified ELISA test-system.

In case if antibodies cross-reacting with NQ21 are synthesized mostly against monomeric (and not oligomeric) viral glycoprotein, they should be able to bind gp120 molecules situated on the membrane of infected cells. In one of the experimental works C1 region of gp120 was shown to be the best target for antibodies with antibody-dependent cellular cytotoxic activity (ADCC) among other regions of this protein [31, 32]. Indeed, NK-cells recognize antibodies bound to viral antigens situated on the cellular membrane and kill infected cells. In our opinion, immunization against the peptide NQ21 should cause production of antibodies with antibody-dependent cellular cytotoxic activity. High titer of antibodies with ADCC is one of the predictors of slow progression of HIV-infection [33, 34].

4.3 *Experimental data on diphtheria toxin immunogenicity*

Diphtheria toxin consists of two chains connected by a single disulfide bound [35]. Precursor of diphtheria toxin encoded by *tox* gene is cleaved by proteases into Chain A (N

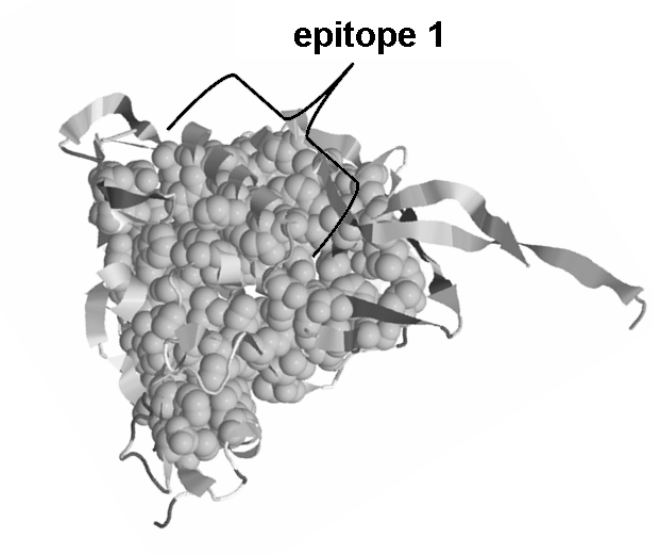

Figure 11. Epitopia output for the X-ray structure (PDB ID: 3JWO) of gp120. Buried amino acid residues are shown as balls.

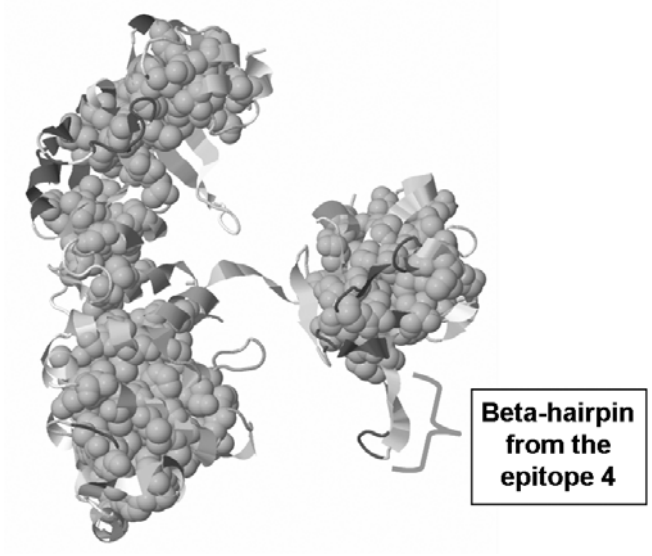

Figure 12. Epitopia output for the X-ray structure (PDB ID: 1SGK) of diphtheria toxin. Buried amino acid residues are shown as balls.

-terminal) and Chain B (C-terminal). There is also a short signal peptide encoded by the first 25 codons of *tox* open reading frame [36].

Chain A forms catalytic domain of the toxin. Catalytic domain blocks protein synthesis by causing an ADP-ribosylation of elongation factor 2, thus provoking cell death [35]. Epitope 1 is situated on the surface of chain A.

There are two distinct domains in chain B of the toxin (see Figure 12). One of them (N-terminal part of chain B) is predominantly alpha-helical one. This domain is called "translocation domain" or "transmembrane domain". Indeed, long alpha-helixes from this domain are involved in translocation of the catalytic domain from endosome to cytoplasm. Actually, acidic pH existing in endosome causes conformational changes in the structure of translocation domain leading to the penetration of endosome membrane by two long alpha-helixes [35]. Epitope 2 is situated on the surface of translocation domain.

C-terminal part of B Chain is predominantly beta-structural (see Figure 12). This domain is able to bind its specific receptor on the surface of cells [35]. Both epitope 3 and epitope 4 are situated on a surface of receptor-binding domain. Epitope 4 really consists of two immunogenic regions situated close to each other in the primary structure, but relatively far from each other in the tertiary structure of the toxin (see Figure 12).

Many researchers tried to use recombinant proteins representing certain parts of diphtheria toxin in vaccine design studies [1, 6, 37, 38]. It has been shown [6, 38] that recombinant Chain B of the toxin is able to induce toxin-neutralizing antibodies in laboratory animals (rabbits and guinea pigs). Moreover, recombinant receptor-binding domain of Chain B is also able to induce formation of toxin-neutralizing antibodies [1].

4.4 *Solutions for antigen design for future recombinant diphtheria toxin vaccines*

According to our results, the target for neutralizing antibodies (receptor-binding domain) possesses the less mutable 3D epitope. It means that the usage of recombinant receptor-binding domain in new vaccine against the diphtheria toxin is well justified. Abovementioned recombinant receptor-binding domain may also be a good antigen for production of ELISA diagnostic tests to control the level of neutralizing antibodies.

N-terminal part of the epitope 4 is situated on a surface of the receptor-binding domain, while C-terminal part of it forms a protruding structure (see Figure 12). This structure is formed by two beta strands connected together (in a beta structure) and a loop between them (see Figure 12). This loop containing a 3/10 helix is recognized as linear B-cell epitope by BepiPred 1.0 [13], and as 3D epitope by DiscoTope 1.2 [22] (see Figure 1). This loop is accessible to a solvent according to Epitopia [23] prediction (see Figure 12). In our opinion, antibodies against synthetic peptide with the amino acid sequence of this beta-hairpin (SIGVLGYQKTVDHTKVNSKLSLF) may be able to cross-react with the subsequent region of diphtheria toxin and *vice versa*.

There are a few amino acid substitutions between *C. diphtheria* and *C. ulcerans* toxins in epitope 1 and epitope 2 (see Figure 13). Epitope 3 is quite invariable at least in seventeen sequences studied. There are six amino acid substitutions between sequences of *C. diphtheria* and *C. ulcerans* toxins in epitope 4 [4]. All of them are concentrated in the center of this region, which is not recognized as highly antigenic by all the methods used (see Figure 1). However, there is a possibility that at least certain antibodies against this epitope of *C. diphtheria* toxin are not able to bind *C. ulcerans* toxin. So, antibodies to this region of the diphtheria toxin may be used

Figure 13. Alignment of amino acid sequences with predicted 3D-epitopes. Invariable amino acid residues are written in bold. Amino acid variations are written separately for toxins from *Corynebactrium diphtheria* and *Corynebacterium ulcerans* phages.

in ELISA assays to discriminate between *C. diphtheria* and *C. ulcerans* toxins. The beta-hairpin described above (at least its highly antigenic loop) seems to be conserved between *C. diphtheria* and *C. ulcerans* toxins.

5. Concluding Remarks

A method for estimation of mutability levels for immunogenic determinants includes two steps: i) estimation of mutational pressure direction and ii) selection of the less mutable immunogenic determinant (determinants). The second step is based on the estimation of the amount of mutable nucleotides in missense, synonymous and nonsense sites of regions coding for immunogenic determinants for each type of the most commonly occurring nucleotide mutations in the given gene. The region coding for the less mutable immunogenic determinant should satisfy three criterions: i) a probability to be synonymous (or synonymous or nonsense) for the most common types of nucleotide mutations should be the highest one inside it; ii) it should has the lowest amount of mutable nucleotides in nonsynonymous (or in missense) sites for the most common types of nucleotide mutations; ii) it should has the highest amount of mutable nucleotides in synonymous (or in synonymous and nonsense) sites for the most common types of nucleotide mutations.

Our *in silico* method showed a good performance on HIV1 gp120 protein. Predictions have been successfully confirmed *in vitro*.

6. Supplementary material

Supplementary material regarding this manuscript is online available in the web page of JIOMICS.

http://www.jiomics.com/index.php/jio/rt/suppFiles/64/0

Table 1. Percentage of variable sites among sequences coding for HIV1 gp120 from each of the 34 monophyletic sets.

Table 2. Average levels of nucleotide usage in third codon positions (3A; 3T(U); 3G and 3C) in comparison with nucleotide usage in invariable sites from third codon positions (3Ai; 3T(U)i; 3Gi and 3Ci) for each from 34 sets of monophyletic sequences coding for HIV1 gp120.

Table 3. Consensus sequence of the predicted 3D epitope 2 from gp120 of HIV1.

Table 4. Consensus sequence of the predicted 3D epitope 3 from gp120 of HIV1.

Table 5. Consensus sequence of the predicted 3D epitope 4 from gp120 of HIV1.

Table 6. Consensus sequence of the predicted 3D epitope 5 from gp120 of HIV1.

Acknowledgements

Authors thank Dr. Oleg Vladimirovich Stoma, the General Director of the MedVax Company (Minsk, Belarus) and Professor Alexander Stanislavovich Vladiko, the Head of the Laboratory of Biotechnology and Immunodiagnostics of Highly Dangerous Infections from Republican Research and Practical Centre for Epidemiology and Microbiology (Minsk, Belarus) for support and useful consultations.

References

1. K. Lobeck, P. Drevet, M. Léonetti, C. Fromen-Romano, F. Ducancel, E. Lajeunesse, C. Lemaire, A. Ménez, Infect. Immun. 66 (1998) 418–423.
2. V.V. Khrustalev, Immunol. Invest. 39 (2010) 551-569. DOI: 10.3109/08820131003706313
3. N. Sueoka, Proc. Natl. Acad. Sci. USA 85 (1988) 2653–2657.
4. A. Sing, M. Hogardt, S. Bierschenk, J. Heesemann, J. Clin. Microbiol. 41 (2003) 4848–4851.
5. A. M. Jr. Pappenheimer, Ann. Rev. Biochem.46 (1977) 69-94.
6. A. A. Kaberniuk, O. S. Oliïnyk, D. V. Kolybo, S. V. Komisarenko, Ukr. Biokhim. Zh. 81 (2009) 92-101.
7. H. Nakao, K. Mazurova, T. Glushkevich, T. Popovic, Res. Microbiol. 148 (1997) 45-54.
8. V. V. Khrustalev, E. V. Barkovsky. SciTopics (2011) Retrieved from http://scitopics.com/Original_computer_algorithms_for_studies_on_directional_mutational_pressure.html
9. V. V. Khrustalev, E. V. Barkovsky, Genomics Proteomics Bioinformatics 8 (2010) 22-32. DOI: 10.1016/S1672-0229(10)60003-4
10. V. V. Khrustalev, Immunol. Invest. 38 (2009) 613-623. DOI: 10.1080/08820130903062202
11. T. P. Hopp, K. R. Woods, Mol. Immunol. 20 (1983) 483–489.
12. J. Kyte, R. Doolittle, J. Mol. Biol. 157 (1982) 105-132.
13. J. E. P. Larsen, O. Lund, M. Nielsen, Immunome Res. 2 (2006) 2.
14. V. V. Khrustalev, Molecular Immunology 47 (2010) 1635-1639. doi:10.1016/j.molimm.2010.01.006
15. V. V. Khrustalev, E. V. Barkovsky. Journal of Theoretical Biology 282 (2011) 71-79. doi:10.1016/j.jtbi.2011.05.018
16. J. R. Lobry, N. Sueoka, Genome Biol. 3 (2002) 0058.
17. V. V. Khrustalev, E. V. Barkovsky, Genomics 96 (2010) 173-

180. doi:10.1016/j.ygeno.2010.06.002

18. E. M. Bunnik, L. Pisas, A. C. van Nuenen, H. Schuitemaker, J. Virol. 82 (2008) 7932-7941.
19. S. Fransen, G. Bridger, J. M. Whitcomb, J. Toma, E. Stawiski, N. Parkin, C. J. Petropoulos, W. Huang, Antimicrob. Agents Chemother. 52 (2008) 2608–2615.
20. J. R. Bailey, K. G. Lassen, H. C. Yang, T. C. Quinn, S. C. Ray, J. N. Blankson, R. F. Siliciano, J. Virol. 80 (2006) 4758–4770.
21. M. Sagar, O. Laeyendecker, S. Lee, J. Gamiel, M. J. Wawer, R. H. Gray, D. Serwadda, N. K. Sewankambo, J. C. Shepherd, J. Toma, W. Huang, T. C. Quinn, J. Infect. Dis 199 (2009) 580–589.
22. P. H. Andersen, M. Nielsen, O. Lund, Protein Science 15 (2006) 2558–2567.
23. N. D. Rubinstein, I. Mayrose, E. Martz, T. Pupko, BMC Bioinformatics 10 (2009) 287.
24. S. Liang, D. Zheng, C. Zhang, M. Zacharias, BMC Bioinformatics 10 (2009) 302.
25. V. V. Khrustalev, E. V. Barkovsky. International Journal of Bioinformatics Research and Applications 7 (2011) 82-100.
26. K. Tamura, J. Dudley, M. Nei, S. Kumar. Mol. Biol. Evol. 24 (2007) 1596-1599.
27. E. V. Barkovsky, V. V. Khrustalev, Mol. Gen. Microbiol. Virol. 24 (2009) 17-23.
28. R. L. Dunfee, E. R.Thomas, D. Gabuzda, Retrovirology 6 (2009) 69.
29. J. P. Moore, Q. J. Saytientau, R. Wyatt, J. Sodroski, J. Virol., 68 (1994) 469-484.
30. G. R. Nakamura, R. Byrn, K. Rosenthal, J. P. Porter, M. R. Hobbs, L. Riddle, D. J. Eastman, D. Dowbenko, T. Gregory, B. M. Fendly, AIDS Res Hum Retroviruses 8 (1992) 1875-1885.
31. A. Chung, I. Stratov, E. Rollman, S. J. Kent, 4th IAS Conference on HIV Pathogenesis, Treatment and Prevention, Sydney, Australia, (2007) abstract [TUPEA015].
32. A. Chung, E. Rollman, L. J. Center, S. J. Kent, I. Stratov, J. Immunol. 182 (2009) 1202-1210.
33. D. N. Forthal, G. Landucci, T. B. Phan, J. Becerra, J. Virol. 79 (2005) 2042–2049.
34. V. R. Gómez-Román, C. Cao, Y. Bai, H. Santamaría, G. Acero, K. Manoutcharian, D. B. Weiner, K. E. Ugen, G. Gevorkian, J. Acquired Immune Deficiency Syndromes. 31 (2002) 147-153.
35. C. E. Bell, D. Eisenberg, Biochemistry 36 (1997) 481-488.
36. V. Kolodkina, L. Titov, T. Sharapa, O. Drozhzhina, Mol. Gen. Mikrobiol. Virusol. 1 (2007) 22-29.
37. V. Y. Perera, M. J. Corbel, Epidemiol. Infect. 105 (1990) 457-468.
38. D. V. Nascimento, E. M. B. Lemes, J. L. S.Queiroz, J. G. J. Silva, H. J. Nascimento, E. D. Silva, R. J. Hirata, A. A. S. O Dias, C. S. Santos, G. M. B. Pereira, A. L. Mattos-Guaraldi, G. R. G. Armoa, Brazilian Journal of Medical and Biological Research 43

IN PRESS | JIOMICS 2011