DENNIS MALETICH JUNQUEIRA

HUMAN IMMUNODEFICIENCY VIRUS TYPE 1 SUBTYPE C EXTERNAL GLYCOPROTEINS EPITOPES: *IN SILICO* PREDICTIONS

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciências Biológicas – Ênfase Molecular, Celular e Funcional.
Universidade Federal do Rio Grande do Sul
Departamento de Microbiologia do Instituto de Ciências Básicas da Saúde.

DR. PAULO MICHEL ROEHE
Orientador

DRA. SABRINA ESTEVES DE MATOS ALMEIDA
DR. FERNANDO ROSADO SPILKI
Co-orientadores

PORTO ALEGRE

2008

Eu, Paulo Michel Roehe, orientador do aluno Dennis Maletich Junqueira, atesto para os devidos fins que o trabalho de conclusão de curso do aluno foi corrigido após a apresentação para a banca avaliadora.

Porto Alegre, 09 de dezembro de 2008.

Orientador

HUMAN IMMUNODEFICIENCY VIRUS TYPE 1 SUBTYPE C EXTERNAL GLYCOPROTEIN EPITOPES: *IN SILICO* PREDICTIONS

Dennis Maletich Junqueira[a], Rúbia Marília de Medeiros[a], Sabrina Esteves de Matos Almeida[a], Fernando Rosado Spilki[b], Paulo Michel Roehe[c]

[a] Scientific and Technologic Development Center (CDCT), State Foundation of Health's Research and Production (FEPPS), Ipiranga Avenue, 5400 - Porto Alegre/RS, Brazil
[b] Institute of Health Sciences, FEEVALE Universitary Center, RS-239 Road, 2755 - Novo Hamburgo/RS , Brazil
[c] Virology Laboratory, Department of Microbiology, Health's Basic Science Institute, Rio Grande do Sul Federal University, Sarmento Leite Street, 500 - Porto Alegre/RS, Brazil
E-mail: dennismaletich@hotmail.com

ABSTRACT

Subtype C human immunodeficiency virus type 1 (HIV-1) is rapidly diversifying among populations, which display extensive polymorphism of genes encoding class I human leukocyte antigen (HLA) proteins, as detected **in different regions of the world. Broadly conserved HIV-1 cytotoxic T cell (CTL) epitopes considering 128 subtype C external glycoprotein gp120 sequences selected from GenBank were identified to A*0201, A*0301, A*1101 and B*07 HLA alleles using Epijen software. NetChop allowed to predict proteasome cleavage followed by prediction of binding to transport associated with antigen processing on TapPred. Glycosylation and positively selected sites within epitope sequences were also observed. Furthermore, three-dimensional structures of subtype C gp120 were predicted from consensus sequences in PHYRE and the PYMOL software was used to verify positions occupied by conserved epitopes. Finally, we predicted discontinuous B cell epitopes in DiscoTope 1.2. There is a recognized evolutionary force of HIV-1 to escape from B cells and CTL responses mutating sites that can negatively select the viral population. These types of analyses could be useful to understand HIV-1 epidemiology associated with polymorphisms in HLA alleles frequent in a determined region. It is expected that such knowledge may provide additional support for vaccine development.**

Key Words: HIV-1, subtype C, bioinformatics, epitope, gp120, HLA

1. Introduction:

Human immunodeficiency virus type 1 (HIV-1) is the etiological agent of Acquired Immunodeficiency Syndrome (AIDS). Established HIV-1 infection causes selective depletion of CD4+ T cells, leading the host to a persistent immunodeficiency (Costin, 2007). HIV-1 genome is characterized by a high propensity to mutate, generating an extensive genetic variation (Rambaut et al., 2004). Such rapid evolution is the result of a combination of factors, including the lack of a proof-reading mechanism coupled with the high mismatch error rates (Korber et al., 2001) in viral RNA polymerase reverse transcriptase enzyme function. Moreover, events of genetic recombination during the viral cycle in co-infected host cells can generate new variants (Kandathil et al., 2005).

Subtype C is the most prevalent HIV-1 form outside Africa and it is responsible for more than 48% of infections worldwide and the main cause of infections in India, plus several countries in Africa (Novitsky et al., 1999; Soares et al., 2005). Despite the high prevalence of subtype B in South America (Russell et al., 2000), in Southern Brazil, subtype C also circulates as a predominant form in infections due to recent introduction of this viral genotype in the region (Bello et al., 2008). Recent data points to regional clusters within subtypes, showing significantly differences in subtype C among isolates from different parts of the world (WHO-UNAIDS, 2001). Difficulties on development of HIV vaccines are mainly due to this heterogeneity in viral populations.

Cellular immune response, mediated by T lymphocytes, is responsible for controlling both chronic and acute HIV infections (McMichael and Rowland-Jones, 2001). Cytotoxic T lymphocytes (CTLs) recognize epitopes presented on surface of infected cells by human leukocyte antigen (HLA) class I

molecules, inducing T-cell activation and clonal expansion of activated cells (Bernardin et al., 2005). To achieve effective formation of the epitope-HLA complex, the peptide must to be cleaved in the proteasome in citoplasm of antigen presenting cells (APCs), followed by transport of the products by the transporter associated with antigen processing (TAP) protein before binding to HLA class I molecules (De Groot et al., 2001). A central role in controlling the virus is dependent on allelic variation in genes coding for HLA on the host's genome (Stephens, 2005). These polymorphisms are variable both within and between different ethnic groups and affect the binding and presentation of peptides in the context of HLA class I molecules (Carrington and O'Brien, 2003; McMichael and Rowland-Jones, 2001; Stephens, 2005).

Viral mutations in specific sites provide an important mechanism to escape from CTL responses (Moore et al., 2002). When variation occurs in genome regions selected by host cellular mechanisms to generate epitopes, CTL response may be inhibited, especially when such mutations alter amino acids used as anchors for HLA binding (MacKinney et al., 2004). Sequence modifications could result in altered intracellular proteasome processing of peptides and consequent altered binding to HLA or interfere with binding of the HLA-peptide complex to the T cell receptor (De Groot et al., 2003). Consequently, CTL escape mutants would lead rapidly to immune collapse and rapid progression to AIDS (Kmieciak et al., 1998). However, conserved residues in functional or structural regions of a given protein must to be perpetuated to ensure viral replication, thus amino acid substitutions on these motifs are negatively selected (MacKinney et al., 2004).

Glycoprotein (gp) 120 is the surface unit protein of HIV envelope (env) spikes and is anchored by gp41, a transmembrane protein, to the viral envelope (Wyatt and Sodroski, 1998). External glycoproteins are important targets for the development of vaccines against AIDS, including specific CTL effectors elicitation approaches (Johnson et al., 1991; Koenig et al., 1988; Takahashi et al., 1988). Recent evidences in HIV-infected individuals of circulating CTL against envelope proteins was obtained in experimental procedures (Takahashi et al., 1988), revealing a potential target to viral epitope researches.

Epitope mapping on virus proteins can provide useful information to vaccine development. Epitope-driven development of vaccines is especially important when dealing with pathogens that display high levels of genetic diversity (Mota-Miranda et al., 2007; Martin et al., 2003). The use of experimental methods to discover relevant CTL epitopes - such as cloning and microarray analysis of overlapping peptides, may have low cost-effectiveness and the final results may be hampered by the presence of diversity among isolates for the epitopes found. To avoid these problems, a previous step putative epitope screening using bioinformatics tools, followed by confirmation in vitro, reduces the cost of development for candidate vaccines as well as may increases the efficacy of the experimental immunogens (Martin et al., 2003). Another possible advantage of this *in silico* approach is the possibility to inquire whether the diverse HLA alleles present in a given population are prone to direct a immune response to different pathogen genotypes. The present study aims to predict the presence of CTL HIV-1 subtype C-specific epitopes on the gp-120 envelope protein, when analyzed for putative to HLA alleles prevalent in world's population.

2. Material and Methods:

2.1. *Sequences targetting:*

All sequences analyzed in this study were selected from GenBank data base, between July and October 2008. We selected and analyzed only nucleotide sequences from gp120 region of HIV-1 subtype C, consisting of 320 samples. Fragments smaller than 1300 base pairs in length were rejected from the present study.

Using Bioedit (Hall, 1999), the resultant sequences were renamed following a numeric order of selection from GenBank. After screening and identification, fragments were submitted to Rega HIV Subtyping Tool version 2.0 (Oliveira et al., 2005) and to Stanford database (Rhee et al., 2003) to confirm

subtype and to verify the presence of premature stop codons respectively. Sequences not subtyped as C or/and presenting premature stop codons were also discarded.

Alignment of these fragments was conducted using ClustalX 2.0 software (Thompson et al., 1997). Based on reference sequences collected from Los Alamos Sequence Data Base, ClustalX allowed us to identify clusters of sequences presenting specific patterns of HIV-1 subtype C from different parts of the world. All sequences were edited and verified manually using Bioedit.

2.2. *Construction of phylogenetic trees:*

Neighbor-joining phylogenetic trees were constructed under HKY distance matrix implemented in PAUP 4.0 software (Swofford, 2001) to analyze the relationship among sequences. Constructed trees allowed us to identify and discard fragments presenting high values of homology.

2. 3. *T cell epitopes prediction:*

T cells epitope mapping were performed using Epijen online software (Doytchinova et al., 2006). It provides information on epitope sequences according to selected HLA allele and scores the potential ability to bind to HLA class I molecules, considering its concentration in cell.

Four HLA alleles were selected in Epijen to generate predicted epitopes: HLA A*0201, HLA A*0301, HLA A*1101 and HLA B*07. Peptides recognized by these four alleles give rise to a predicted epitope recognition in more than 90% of the global population, regardless ethnicity (Sette and Sidney, 1999). Threshold 0.0 to proteasome cleavage were used to B*07 allele, wich prefer Phe or Trp at the C-terminus (Doytchinova et al., 2006). The epitopes generated for 0201, 0301 and 1101 HLA alleles were predicted using a threshold of 0.1. TAP prediction threshold of 5 were selected to 0201, 0301 and 1101 alleles, while to B*07 allele a threshold of 3 were used as recommended by Doytchinova et al. (2006). The 5% threshold of available peptides sourced by one protein ranked by Epijen were use as a default value to all HLA class I alleles. NetChop online software allowed us to reiterate the proteasome cleavage and select epitopes correctly processed.

After prediction, we screened for conserved epitopes among all sequences considering the same HLA allele. Similarity and variation observations of the most conserved epitopes, assisted by Bioedit software, were conducted to perform an evolutionary analyze and to verify conservation of anchor sites within epitopes.

2.4. *TAP binding:*

Binding to TAP protein was predicted by TapPred online program (Bhasin et al., 2007). We selected to predict TAP binders using cascade support vector machine model. Results were separated in three groups by TapPred (low, intermediate and high) according to predicted TAP binding value.

2.5. Prediction of *Glycosylation sites:*

Glycosylation is required for proper folding and conformational stability of the envelope glycoprotein and when component of an epitope can prejudice the binding to HLA class I molecule and may inhibit the adhesion of ligants to a given protein fragment (Fenouillet et al., 1994). To verify potential sites with high specificity of N-glycosylation we used NetNGlyc 1.0 server software (Kukuruzinska and Lennon, 1998). NetOGlyc 3.1 server (Julenius et al., 2005) was accessed to predict O-glycosylation within epitopes sequences. We only considered as likely glycosylation sites those with scores greater than 0.5.

2.6. *Selective pressure test:*

To test the hypothesis that amino acids substitutions within the most conserved epitopes could be undergoing different selective pressures (Yang, 1997) we tested differential models at individual sites. The analyses were performed using the nucleotide sequence of the proteasome cleavage fragments of the

most common epitopes found in this study. To verify conservation of proteasome cleavage sites we considered 3 codons before and after the cleavage sites. Likelihood methods were implemented in the PAML package (Yang et al., 2000), version 3.14a, with several site-specific models : M0 (one ratio) - which assumes constant selective pressure across codon sites and over time - and M3 (discrete) - which assumes discrete distribution to model heterogeneous ratio among sites, M1a (neutral), M2a (selection), M7: beta and M8: beta&ω – with variable selective pressure across codon sites, were used to estimate selective pressure and test for positive selection. The selective pressure at the protein level was measured by ω, the ratio of non-synonymous to synonymous rates *d*N/*d*S, with $\omega < 1$, $= 1$, or > 1 indicating conserved, neutral or adaptive evolution, respectively. The posterior probability of each codon site belonging to the positively selected category was computed by the naive empirical Bayes method (NEB) and Bayes empirycal Bayes (BEB). Finally, likelihood ratio test (LTR) analysis was used to determine: (1) if site heterogeneity selection was present and (2) if there were to be positive selection sites.

2.7. *Three-dimensional structure (3D) prediction:*

To construct a 3D model of HIV-1 gp120, a nucleotide consensus sequence of each subtype C group was defined using JalView software (Clamp et al., 2004). PHYRE online program (Bennett-Lovsey et al., 2008a, 2008b) was used to predict the 3D protein conformation based on the consensus sequence of each group. Four best models provided by PHYRE were selected to be checked by PROCHECK (Laskowski et al., 1993; Morris et al., 1992). This software allowed the selection of the best model based on the stereochemical quality of protein structure coordinates. The models generated were additionally analysed using Whatcheck to reevaluate stereochemistry and Verify 3D to check statistical potentials of mean force (Bowie et al., 1991; Lüthy et al., 1992; Vriend, 1990). The model which best satisfied the program's criteria was selected.

PYMOL software (DeLano, 2002) was used to verify position occupied by the most conserved epitopes in the 3D predicted protein structure of each group of HIV-1 subtype C gp120.

2. 8. *B cell epitopes prediction:*

Prediction of discontinuous B cell epitopes from protein 3D structures of HIV-1 subtype C gp120 was performed using DiscoTope 1.2 Server online program (Andersen et al., 2006). This software utilizes calculations of surface accessibility and a novel epitope propensity amino acid score. The selected DiscoTope threshold score for epitope identification corresponded to a specificity of 75%.

After prediction, regions found as B cell epitopes were marked in the previous constructed gp120 3D model proteins to each subtype C in PYMOL software.

2.9. *Similarity to human sequences:*

We used Ensembl Genome Server (Xose et al., 2007) to verify similarity of the most conserved epitopes amino acid sequence to human genome sequence. This tool allowed us to identify possible epitopes that could induce autoimmunity due to similar characteristics of the human proteins to HIV-1 gp120 protein after intra cellular processing.

3. Results:

3.1. *Epitope selection:*

Following our criteria to screen the peptides, 128 different sequences representing gp120 envelope protein of HIV-1 subtype C were selected from Genbank to predict T cell epitopes. According to the constructed filogenetic trees, 72 sequences showed high similarity to Brazilian reference sequence, 44 were identified as Ethiopian subtype C and 12 as Indians.

We found 140 different epitopes; however, considering the repetitions, 545 epitopes were mapped in HIV-1 subtype C sequences as a total amount. All epitopes found in this study were composed

of nine amino acids and, at least one epitope was identified by one HLA allele in each sequence (Figure 1).

To HLA A*0201 allele we found significant number of conserved and related epitopes in analysed sequences, showing 51 different epitopes (Figure 2) in 317 total antigenic peptides sequences. Despite the recognized conservation of some epitopes restricted to 0201 allele, 70% of the predicted epitopes were unique, not being recognized in more than one sequence. Our results showed that 0201 was the only allele to recognize epitopes in all sequences. Three epitopes (QMHEDIISL, CTHGIKPVV and NLTNNVKTI) had a high frequency and showed varying degrees of conservation. The previously described QMHEDIISL epitope (Kmieciak et al., 1998) was the most conserved epitope in our sequences (51%); however, 13 other predicted epitopes showed a high similarity to this sequence (Table 1).

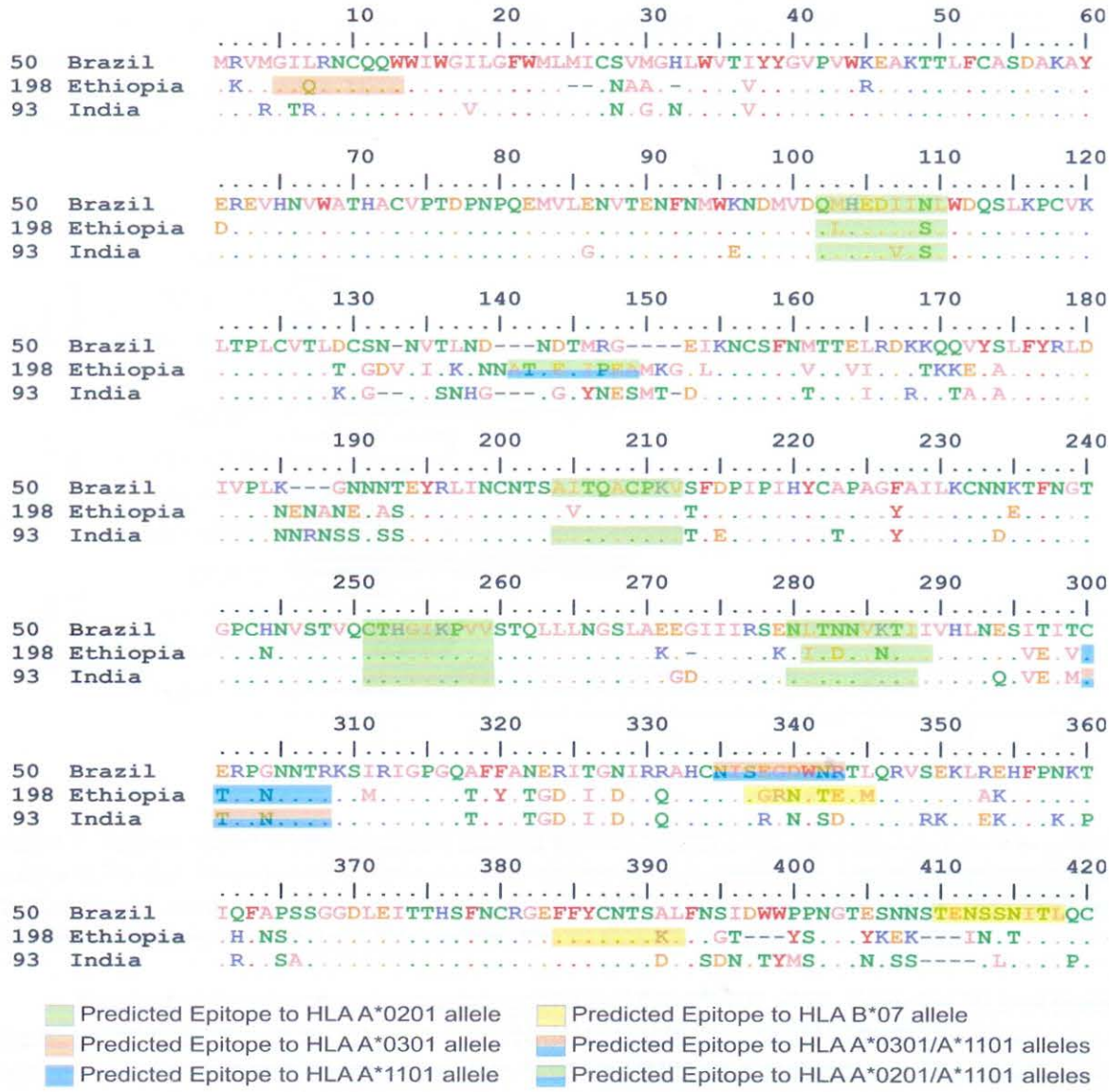

Figure 1. Three aligned amino acid sequences representing Brazilian, Ethiopian and Indian HIV-1 subtype C gp120 (not complet). Hachured peptides represents predicted epitopes to an HLA allele by Epijen online software. Green squares represents predicted epitopes resctrict to HLA A*0201 allele, orange squares represents peptides recognized to HLA A*0301 allele restrictly, blue squares shown predicted epitopes to HLA A*1101 allele and yellow squares represent epitopes restricted to HLA B*07 allele. Bicoloured squares show epitopes recognized by more than one HLA allele.

Considering all variants, QMHEDIISL epitope family was found in 98% of the sequences. The NLTNNVKTI family was found in 61% of the sequences, presenting a main similarity value of 80%. We observed a significant prevalence of this peptide in HIV-1 sequences classified as Brazilian and Indian subtype C. Finally, the CTHGIKPVV epitope was identified in 42% of the total sequences analysed in this work.

We found only one Ethiopian subtype C sequence showing an epitope (ATNETIPEA) capable to be recognized for both HLA A*0201 and A*1101 class I molecules (Figure 1). Five other sequences presented an epitope capable to bind in HLA class I molecules presented by HLA A*0201 and HLA B*07 alleles.

Based on *in silico* predictions, HLA A*0301 class I molecules recognized 37 possible antigenic peptides grouped in 15 different epitopes (Figure 1). The most conserved peptides appeared only in nine sequences and other six are fruit of amino acids substitutions. The RQAHCNISK predicted epitope exclusively bound to 0301 allele in two sequences recognized as brazilian subtype C and in one sequence identified as an Indian sequence. However, it could be recognized as an epitope for both HLA A*0301 e HLA A*1101 alleles in one sequence.

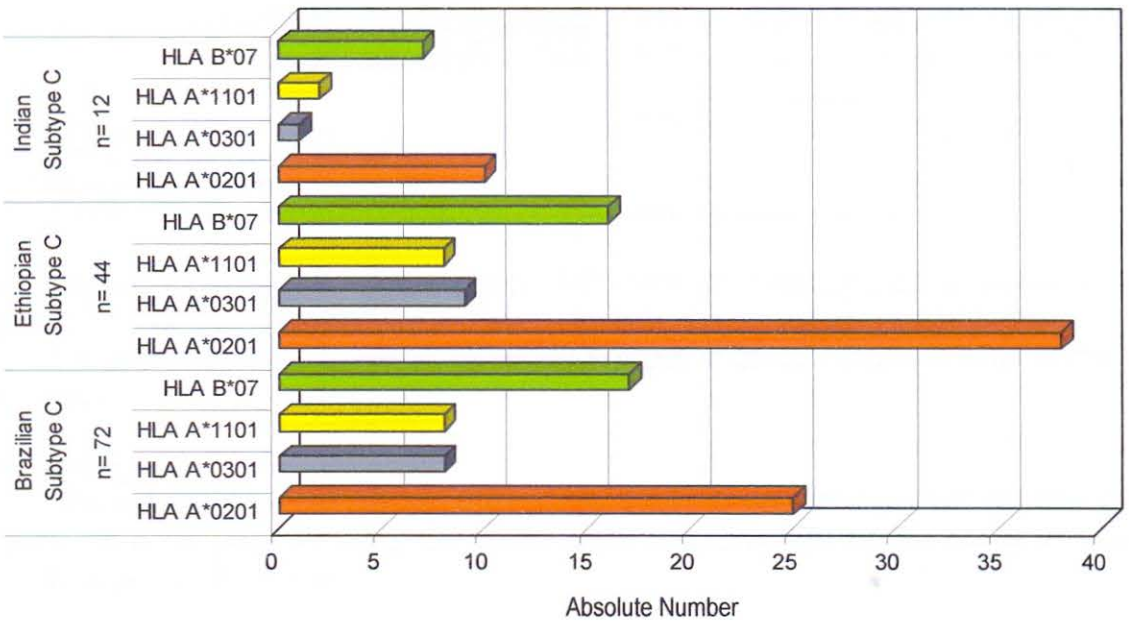

Figure 2. Absolute number of different epitopes predicted from HIV-1 gp120 amino acid sequences from three patterns of subtype C: Brazilian (72 sequences), Ethiopian (44 sequences) and Indian (12 sequences). Four HLA alleles were considered to the epitope prediction at Epijen software: HLA A*0201, HLA A*0301, HLA A*1101 and HLA B*07. This graphic do not consider equal epitopes identified in different sequences, here we show only the quantities of varied epitopes.

Fourteen different epitopes were recognized by HLA A*1101 allele. From the 36 total predicted antigenic peptides to this allele, 20 were CTRPNNNTR (Table 1). The CTRPNNNTR epitope also was recognized as a promiscuous epitope binding in 1101 and 0301 HLA molecules in 19 sequences. Other 17 epitopes were recognized for both molecules, mainly in Brazilian sequences. Our results demonstrate that CTRPNNNTR epitope can bind to both alleles, however appeared to be an epitope to HLA A*0301 only when it was an epitope to HLA A*1101, do not being recognized only by HLA A*0301.

To HLA B*07 from 39 different epitopes found in HIV-1 subtype C gp120 sequences only 3 appeared in more than one sequence. This allele showed a high prevalence of non conserved epitopes (93%). Binding affinity results from the peptides selected by Epijen did not show a conserved epitope

Table 1. Most conserved HIV-1 subtype C gp120 epitopes found among Brazilian, Ethiopian and Indian subtype C sequences and related amino acid sequences presented restrictly by HLA A*0201 class I molecules.

Absolute[a] Number	Epitope Sequence	%[b]	TAP Binding	Glycosylation	
				N	O
66	Q M H E D I I S L	1,00	Intermidate	No sites	No sites
15	. . . . . V . . .	0.89	Intermidate	No sites	01 site
25	. . . . . N .	0,89	Intermidate	No sites	No sites
9	 I	0,89	High	No sites	01 site
1	. . . . . G .	0,89	Intermidate	No sites	No sites
1	. . . Q	0,89	Intermidate	No sites	No sites
1	. . Q	0,89	High	No sites	01 site
1	. L	0,89	Intermidate	No sites	No sites
1	. . . . . R .	0,89	Intermidate	No sites	No sites
1	. . . G	0,89	Intermidate	No sites	No sites
1	H	0,89	Intermidate	No sites	01 site
2	. . . . . V . . I	0,78	High	No sites	01 site
1	. . Q . . . N .	0,78	High	No sites	No sites
1	. . . . . V . N .	0,78	High	No sites	No sites
126		0,80	-	-	-

[a] Number of sequences that showed the epitope;
[b] Percentage of similarity;
(.) Similar amino acid presented by epitope sequence that showed high absolute number

between Brazilian, Ethiopian and Indian subtypes to 07 allele. An Ethiopian subtype sequence showed one epitope (SVRIGPGQA) capable to be recognized for both HLA A*1101 and HLA B*07 alleles.

Six sequences were scored as conserved epitopes in gp120 subtype C gene to three HLA alleles (Table 2).

3.2. *TAP Binding:*

The majority of the epitopes showed an intermediate binding affinity to TAP (61.4%). Only 14 sequences had a low predicted value. Our results show an intermediate binding affinity to conserved aminoacids sequences found here.

3.3. *Glycosylation sites prediction:*

Predicted N-glycosylation demonstrated that 45.7% of the epitopes found in this study were glycosylated. HLA A*1101 was the only one to have more than 60% of putatively glycosylated Epitopes.

Sites of probable O-glycosilation were found in 80% of epitopes. To all epitopes predicted for 1101 we found, at least one site of O-glycosylation. Our data show that 93% of epitopes that are found to have O-glycosylation sites are also N-glycosylated.

QMHEDIISL epitope, most common in our sequences, showed no N-glycosylation nor O-glycosylation. Three of the most conserved epitope in this study had two sites of O-glycosylation and one site of N-glycosylation.

3.4 *Selective Pressure Test*:

ML methods were used to assess amino acid variation and differential pressure at individual sites. Analysing the fragment cleaved by proteasome, the peptide containing the CTHGIKPVV epitope

showed to be under negative selective pressure. The M0 (one-ratio) model, which allows no variations in the substitutions taxes for the selected sequence, best fit to this fragment peptide.

The QMHEDIISL and AITQACPKV (Figure 4A, 4D, respectively) containing fragments fit to M3 model (discrete). This model allows different taxes of variaton between sites. Finally, the fragments of NLTNNVKTI and CTRPNNNTR (Figure 4B, 4C) epitopes showed sites identified as targets of positive selection (M2a model). For more information, see Table 1S in supplementary data.

3.5. *Three-dimensional structure prediction:*

Three models of the 3D structure of gp120 protein were obtained from PHYRE representing Brazilian, Ethiopian and Indian C subtype (Figure 3). The models were constructed based on structures with high similarity to the 2b4c structure obtained from X-ray experiments (Kwong et al., 1998). Only models possessing less than 2% of residues in disallowed regions of the Ramachandran plot were used (Table 3).

Five conserved epitopes were identified in predicted proteins (Figure 3). The epitopes QMHEDIISL and CTHGIKPVV were arranged in an alpha helix and in a loop, respectively, in all three models. Epitopes NLTNNVKTI and AITQACPKV were chiefly arranged in conserved sites on a loop.

The CTRPNNNTR epitope was found in different positions among the three models. In Brazilian subtype C, the epitope was positioned in an alpha helix-loop-alpha helix motif, while in Indian subtype C structure it covered the end of an alpha helix and an entire loop, reaching the initial portion of an alpha helix. In the Ethiopian structure the CTRPNNNTR epitope was arranged starting at the end of an alpha helix, ending in a loop structure (Figure 3).

Table 2. Most conserved epitopes present by HLA A*0201, HLA A*0301 and HLA A*1101.

HLA	Epitope Sequence	Percentage of Sequences			Total Number of sequences	TAP binding
		Brazilian Subtype C (72 sequences)	Ethiopian Subtype C (44 sequences)	Indian Subtype C (12 sequences)		
A*0201	Q M H E D I I S L	50%	61%	25%	66	Intermediate
	C T H G I K P V V	34.7%	50%	58.3%	54	Intermediate
	N L T N N V K T I	36.1%	9.1%	50%	36	Intermediate
	A I T Q A C P K V	31.9%	9.1%	50%	33	Intermediate
	Q M H E D V I S L	5.5%	6.8%	66.6%	15	Intermediate
A*0301	C T R P N N N T R*	12.5%	13.6%	33.3%	19	Intermediate
A*1101	C T R P N N N T R*	29.2%	29.5%	41.6%	39	Intermediate

* The CTRPNNNTR were recognized both by HLA A*0301 and HLA A*1101

3.6. *B cell epitopes prediction:*

One hundred and thirty two B-Cell epitope residues out of 330 total residues were predicted to Brazilian subtype C gp120 consensus amino acid sequence. To Ethiopian sequence 130 out of 327 total residues were predicted as B cell epitopes and Indian subtype C consensus sequence revealed 122 predicted residues as targets to B cells.

Using PYMOL software predicted residues of each subtype C were mapped in the 3D structure of HIV-1 gp120 protein (Figures 3D, 3E, 3F). Only 46 sites were predicted as in commom B cell epitopes in the three sequences.

3.7. *Similarity to human sequences:*

Analyzing similarity in amino acids sequence from five most conserved epitopes found in this study to human genome we found two regions presenting high homology to our epitopes. The QMHEDIISL epitope was found to have a identity of 87.5% to a region within a gene coding non-characterized protein C6orf50 located in the short arm of chromossome 6. In the short arm of chromosome 9 we found another region of similarity (62.5%), within a gene coding C9orf105 uncharacterized protein.

The CTHGIKPVV epitope was found within gene MFI2 (CD228) coding a melanotransferrin precursor on long arm of chromosome 3 presenting an identity of 85.7%.

Table 3. Plot statistics obtained from PROCHECK online software to gp120 HIV-1 subtype C protein models. Amino acid sequences were arranged in tertiary structural models by comparative modeling using as template the PDB 2b4c structure.

Evaluation	2b4c model (%)	Subtype C		
		Brazilian (%)	Ethiopian (%)	Indian (%)
Residues in most favoured regions	80.4	79.3	79.5	81.4
Residues in additional allowed regions	18.6	18.6	18.1	16.6
Residues in disallowed regions	0.7	1.4	1.7	1.0
Number of non-glycine and non-proline residues	0.3	0.7	0.7	1.0

4. Discussion:

The growing number of recombinant circulating forms of HIV-1 is a complicating factor for the development of an effective vaccine. The high rates of evolution in HIV-1 allow a rapid adaptation to the selective pressure exerted by host's immune response, both in humoral and cellular domains, becoming the main mechanism responsible for the viral escape from CTL response (Chackerian et al., 1997). In the present study, we have identified a set of multiple HLA A and HLA B binding CD4 T-cell epitopes derived from computer analysis performed at Epijen tool and other complementary predictions.

The diversity of responses to HIV considering different HLA alleles is evident in this study. HLA A*0201 allele was showed to display the potential to bind to several different putative epitope sequences, mainly in Ethiopian subtype C HIV-1, which presented 38 epitopes (Figure 2). The Indian HIV-1 subtype C showed a small number of different epitopes in Epijen sftware. However, this data can be attributable to the small number of sequences that do not represent the real diversity of this subtype. In fact there are only a few Indian sequences of HIV-1representing these genomic regions available in GenBank so far, but at least all of information here showed about epitope prediction and protein analysis are new and could give support to further studies.

Our results show a high number of epitopes recognized by 0201 allele when comparing the four alleles here analyzed. Previous reports (De Groot et al., 2003 ; Shankar et al., 1996) noticed a few number of HLA A2 epitopes and suggested that the HIV-1 genome has evolved to escape presentation by the highly prevalent HLA allele in world populations. However, these researches used neither non-specific subtype nor specific-viral gene-based approaches, reflecting in results all the diversity of several genes of HIV-1. This methodology of epitope searching could, in this way, underestimate the rates of epitope binding to HLA class I molecules when compared to a specific subtype. Moreover, the results found here to 0201 allele show 70% of non-conserved epitopes, reflecting an effort of HIV-1 to escape from CTL response restricted to the most commonly found and evolutionary conserved HLA allele (De Groot et al.,

2003). Our findings are associated specifically with gp120 subtype C; nevertheless, the methodology presented here is based only on *in silico* predictions and may not reproduce faithfully the whole biological

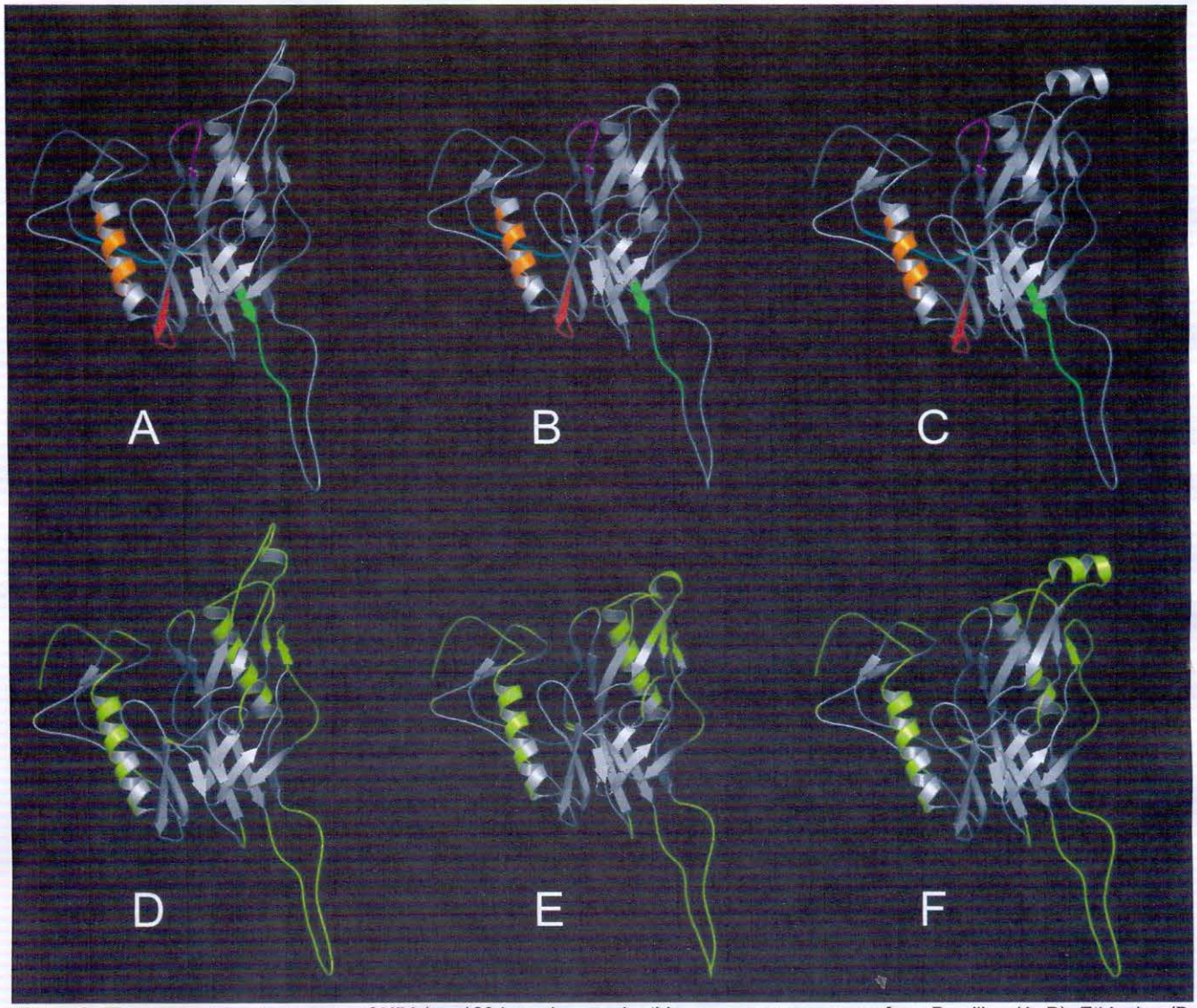

Figure 3. Three-dimensional structure of HIV-1 gp120 based on nucleotide consensus sequence from Brazilian (A, D), Ethiopian (B, E) and Indian (C,F) subtype C showing the positions occupied by conserved epitopes. The diagrams A, B and C shows QMHEDIISL epitope position in orange, the epitope CTHGIKPVV in blue, the NLTNNVKTI epitope in magenta, the AITQACPKV epitope in red and the CTRPNNTR epitope in green located in the V3 loop. Epitopes QMHEDIINL and QMHEDVISL occupy the same position (shown in orange) due to the high similarity in nucleotide sequence. Models D, E an F represent predicted sites of discontinuous B cell epitopes from DiscoTope Server based on consensus sequence of Brazilian, Ethiopian and Indian subtype C, respectively.

activity of the host immune system. Thus, experimental methods should be conducted to confirm the results obtained here.

Macdonald et al. (2000) found a positive correlation of A2 supertype, which includes 0201 allele, to protection against infection and hypothesized that this supertype may mediate protection through the presentation of conserved epitopes that are shared by alleles within supertype. The conserved QMHEDIISL epitope family was recognized in 98% of sequences by HLA molecules of 0201 alleles (Table 1). This high level of conservation is in agreement with Kmieciak et al. (1998) and may be attributable to the structural importance of this site to the protein function; any mutation in this sequence can affect the viral fitness and cause a negative selection by the host's immune system. The selective

pressure analysis in the proteasome cleavage fragment of this epitope showed that there is a negative pressure (ω values varying between 0.10227 a 0.55911) acting in the peptide. Observed mutations in these similar sequences are selected to maintain the amino acid group in the positions P2 and P9 of binding to MHC class I molecules (Table 1).

In a previous study of CTL response to HIV envelope glycoprotein (Kmieciak et al., 1998) the peptide KLTLPLCVTL was characterized by a remarkable degree of conservation and was capable of stimulating env-specific CTL responses in a large proportion of the HIV-infected individuals. Here, this epitope was identified once in a Brazilian subtype C sequence presenting a deletion in the fourth amino acid. This result emphasize the importance of subtype-specific analysis of epitopes to future vaccine development.

HLA A*0301 and HLA A*1101 belong to the A3-like superfamily, which is defined by a shared preference for epitopes carrying A, L, I, V, M, S or T at position 2 and R or K at the C-terminus (Sidney et al., 1996). Due to this shared peptide motifs, HLA class I molecules from a same superfamily can present identical viral epitopes (promiscuous epitopes) as showed by 19 epitopes, mainly in Brazilian subtype C sequences. Previous studies showed that promiscuous epitope recognition among HLA class I allele belonging to HLA A3 superfamily is a common event (Threkeld et al., 1997). However, recently, Lichterfeld et al. (2006) found that promiscuous recognition of HIV-1 epitopes presented by HLA A*03 or HLA A*11 occurs very rarely during natural infection (Lichterfeld et al., 2006). These contradictory data may be due to differential degrees of immunogenicity of each peptide to HLA class I molecule.

Several studies have demonstrated that the principal neutralizing domain of HIV-1 lies within the third variable (V3) region arranged in a loop of gp120 (Hwang et al., 1991; Page et al., 1992). Most of the amino acids of the V3 loop are highly variable between different subtypes of HIV-1; a Gly-Pro-Gly-Arg sequence at the tip of the loop is highly conserved among isolates derived from Europe and North America (Page et al., 1992) and is speculated as the major determinant of cell tropism in HIV-1 (Suberna et al., 1994). We provide relevant information of subtype differentiation, showing that Brazilian, Ethiopian and Indian subtype C, on the contrary of European and North American subtypes, do not present the conserved sequence GPGR. The predicted CTRPNNNTR peptide, arranged within V3 loop, is both promiscuous, recognized by 0301 and 1101 HLA alleles, and restrict only to 1101 allele, as showed in previous study (De Groot et al., 2003). This epitope seems to be arranged in a positive selective pressure site, however in the exact sequence of the epitope (CTRPNNNTR) we found a relaxed selective pressure (ω =1). This discordance may be attributed to the worldwide distribution of HLA A*0301 allele. Thus, HIV-1 could be adapted to the CTL response associated with 0301 HLA due to its wide circulation and selection (De Groot et al., 2003; MacDonald et al., 2000; Moore et al., 2002). Our results showed a low capacity of HLA A*0301 to recognize epitopes in HIV-1 subtype C sequences.

The HLA B7 supertype has been shown to be associated with high viral loads and rapid progression to disease (Kiepiela et al., 2004). De Groot et al. (2008) in an experimental method found that the HLA B7 restricted epitopes were conserved in as many as 93% of comparison sequences considering several subtypes and viral genes. Our results show a low tax of epitope conservation to this supertype and may be explained by the specificity predictions to HIV-1 subtype C gp120 protein.

The transport antigen protein (TAP) is an important component of the intracellular immune response. Human TAP genes carry limited polymorphism, however, when these variants are analyzed, no diferences in the spectrum of transported epitopes was shown (Daniel et al., 1997). Previous association report some interaction between specific HLA class I alleles and specific TAP variants (Henderson et al., 1992). Some peptides are able to access the endoplasmatic reticulum (ER) via other TAP independent pathways. Thus, TAP binding affinity values prediction to HLA A*0201 and HLA A*1101 molecules do not totally define the cellular capacity for transport peptides from citosol to ER before binding to HLA molecule (Henderson et al., 1992; Smith and Lutz, 1996).

HIV genome encodes no gene products capable of synthesizing carbohydrates, it utilizes the cellular machinery to perform N and O-linked glycosylation, however, the locations of glycosylations are

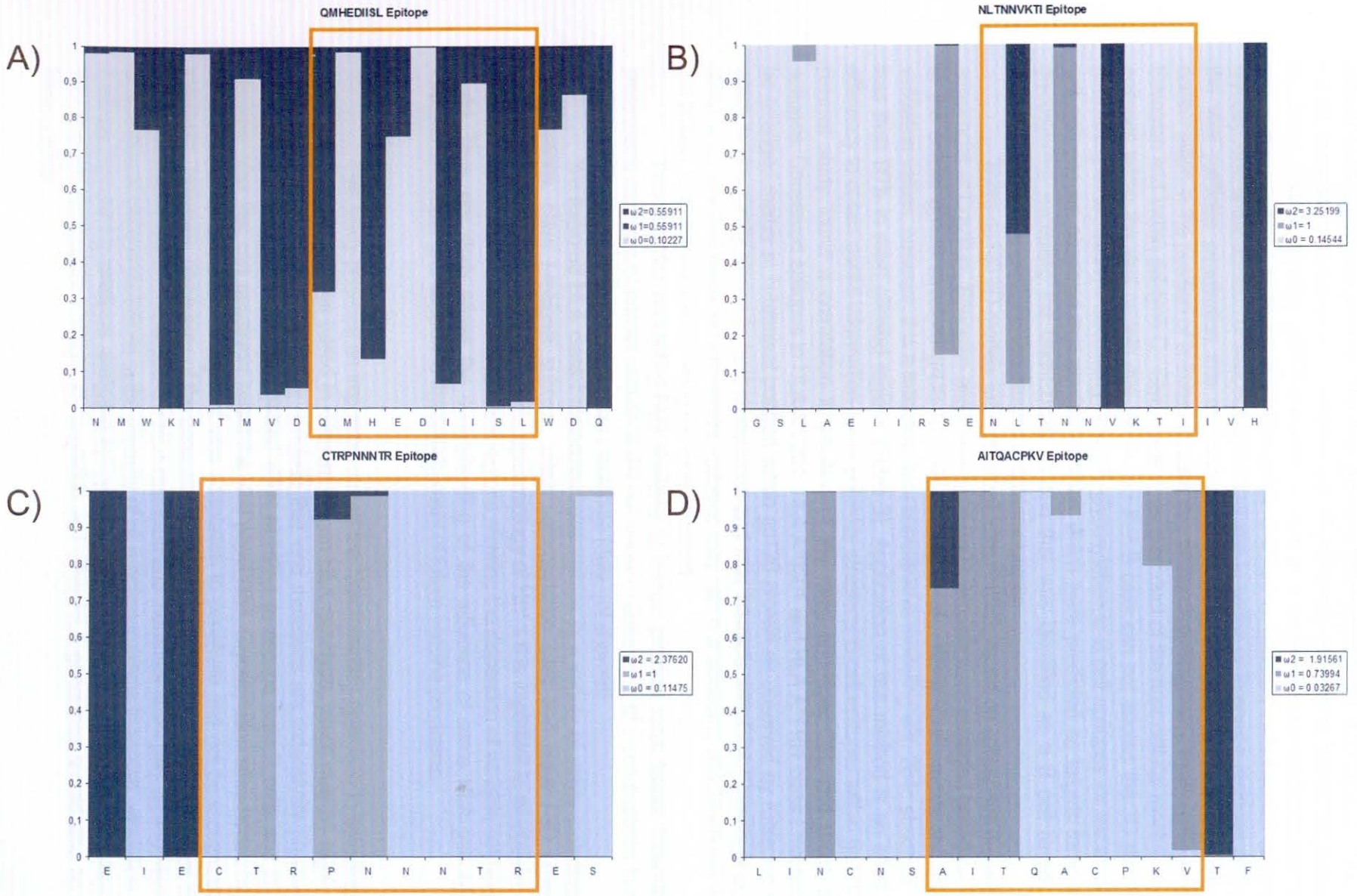

Figure 4. Posterior probabilities that each codon of the epitopes from the site classes under several models, calculated using the naïve empirical Bayes (NEB), the Bayes empirycal Bayes (BEB) procedure produced virtually identical posterior probabilities; (A) QMHEDIISL epitope - M3 model (discrete); (B) NLTNNVKTI epitope - M2a (selection) model; (C) CTRPNNNTR epitope - M2a (selection) model. (D) AITQACPKV epitope - M3 model (discrete); The CTHGIKPVV epitope (M0 - one-ratio model) did not show any variation.

directly encoded by viral genome. The folding of viral glycoproteins, the transmission of the virus and the nature of the immune response to infection are all profoundly affected by this glycosylation (Scanlan et al., 2007). Chackerian et al. (1997) using simian virus as a model showed that the escape from antibody recognition appeared to be influenced by either O-linked or N-linked carbohydrate additions. O-linked glycosylation is the addition of N-acetyl-galactosamine to serine or threonine residues and act facilitating the viral particles penetration into mucosal secretions (Calvete and Sanz, 2008). We found some conservation in the position of N-linked glycans due to the close filogenetic relationship among Brazilian, Ethiopian and Indian subtype C. Glycosylation, as a product of the host cellular mechanism, can protect epitopes within the viral protein due to the fact that B cells can not access the 3D structure of gp120 (McCaffrey et al., 2004). Associating O and N-glycosylation to B cell epitopes for the most conserved T cell epitopes we observed that several peptides which were predicted as targets for B cells are also glycosylated, excepting the QMHEDIISL epitope that was identified as a B cell target, but was not predicted as glycosylated. Further studies should investigate

Three gp120 3D-models were obtained from PHYRE based on consensus deduced amino acid sequences of Brazilian, Ethiopian and Indian subtype C. Models were arranged based on high similarity to the solved gp120 structure (PDB ID 2b4c; Kwong et al., 1998). Our models showed sub-optimal validation values, however, it is a reflection from comparative modeling with 2b4c structure, which shows less than 90% in most allowed regions (Table 3). Application of this method is an effort to increases efficiency to obtain useful information about potential structural B-cell epitopes within gp120 protein. The present study found a correlation in position of the most abundant CTL and B-cell discontinuous epitopes. Conserved epitopes to HLA A*0201, HLA A*0301 and HLA A*1101 class I molecules seems to be arranged mainly in N-terminal position within gp120 protein as observed by Kmieciak et al. (1998), which is structurally similar to host chemokines and may interacts with CD4 (Sharon et al., 2003). This result suggests that predicted conserved epitopes may retain a structural and functional importance to the protein, thus are maintained along the virus evolution.

Two peptides presenting high similarity to human proteins were found; however analysis of proteasome cleavage of human sequence did not reveal maintenance of these peptides to correctly bind to TAP and to be presented to HLA class I molecule (data not showed).

We explored HLA class I molecules relationships in a variety of ways, including class I-TAP combinations, class I-regional subtype, class I-glycosylation sites and class I-3D protein structure. Despite the recognized regional-specific epidemiology of HIV-1 it is important to emphasize the constant circulation of viral forms between different regions allowing the dispersion of variable subtypes around the world. Thus, our study allows the understanding of the influence of some HLA genotypes to HIV-1 infection and could give support to the development of future vaccine programs.

5. Conclusions:

This study provides an important role of the bioinformatics approach to selecting CTL and B-cell discontinuous epitopes based in different human HLA genotype, which may be highly influential in the immune response to HIV-1. We conclude that viral mutations are extensively pressured by host's immune response. There is a recognized effort of HIV-1 to escape from B cells and CTL responses mutating sites that can negatively select the viral population. Glycosylation sites added by host cellular mechanisms are a striking example of viral adaptation to penetrate and not be recognized by immune response. However, some sites of the genome have to be conserved due to the important function of the encoded protein. So, CTL epitopes are HLA-restricted and individual hosts play an important mechanism of selection to escape mutations in the diversity of HLA polymorphisms in human population. Additionally, it will be important to continue analyzing epitope variability in other viral proteins and subtypes of HIV-1. Vital tropism, co-receptors usage, glycosylation and other HLA genotypes should also be evaluated in fuure studies.

6. Acknowledgements:

This manuscript benefited greatly from the comments of Vanessa R. Paixão-Côrtes. This study was supported by Scientific and Technologic Development Center (CDCT)/ State Foundation of Health's Research and Production (FEPPS), RS, Brazil.

7. Abreviations:

AIDS	acquired immunodeficiency syndrome
BEB	Bayes empirical Bayes
CTL	cytotoxic T lymphocyte
env	HIV envelope glycoprotein
gp	glycoprotein
HIV-1	human immunodeficiency virus type 1
HLA	human leukocyte antigen
LTR	likelihood ratio test
NEB	naïve empirical Bayes
TAP	transport associated with antigen processing
V3	third variable region within HIV-1 gp120 nucleotide sequence
ω	Substitutions taxes (non-synonymous/synonymous substitutions)
3D	three-dimensional

8. References:

Andersen, P.H., Nielsen, M., Lund, O., 2006. Prediction of residues in discontinuous B cell epitopes using protein 3D structures. Protein Science 15, 2558-2567.

Bello, G., Passaes, P.B., Guimaraes, M.L., Lorete, R., Almeida, S.E.M., Medeiros, R.M., Alencastro, P.R., Morgado, M.G., 2008 Origin and evolutionary history of HIV-1 subtype C in Brazil. AIDS 22, 1993-2000.

Bennett-Lovsey, R.M., Herbert, A.D., Sternberg, M.J., Kelley, L.A., 2008a. Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins 70 (3),611-625.

Bennett-Lovsey, R.M., Herbert, A.D., Sternberg, M.J., Kelley, L.A., 2008b. Proteins: structure, function, bioinformatics. Vol. 70, 3, 611-625.

Bernardin, F., Kong, D., Peddada, L., Baxter-Lowe, L.A., Delwart, E., 2005. Human Immunodeficiency Virus Mutations during the First Month of Infection Are Preferentially Found in Known Cytotoxic T-Lymphocyte Epitopes. J. of Virol. 79, 11523-11528.

Bhasin, M., Lata, S., Raghava, G.P., 2007. TAPPred prediction of TAP-binding peptides in antigens. Methods Mol Biol. 409,381-386.

Bowie, J.U., Lüthy, R., Eisenberg, D.A., 1991. Method to identify protein sequences that fold into a known three-dimensional structure. Science. 12, 253, 164-170.

Calvete, J.J., Sanz, L., 2008. Analysis of O-glycosylation. Methods Mol Biol. 446,281-292.

Carrington, M., O'Brien, S.J., 2003. The Influence of HLA Genotype on AIDS. Annu. Rev. Med. 54, 535-551.

Chackerian, B., Rudensey, L.M., Overbaugh, J., 1997. Specific N-linked and O-linked glyco-sylation modifications in the envelope VI domain of simian immunodeficiency virus that evolve in the host alter recognition by neutralizing antibodies. J. Virol. 71, 7719–7727.

Clamp, M., Cuff, J., Searle, S.M., Barton, G.J. 2004. The Jalview Java alignment editor. Bioinformatics. 20,426-427.

Costin, J.M., 2007. Cytophatic Mechanisms of HIV-1. Virol. J. 4, 100.

Daniel, S., Caillat-Zucman, S., Hammer, J., Bach, J.F., van Endert, P.M., 1997. Absence of functional relevance of human transporter associated with antigen processing polymorphism for peptide selection. J Immunol 1997; 159: 2350–2357.

De Groot, A.S., Sbai, H., Frost, J., Saint-Aubin, C., Chinai, N., Martin, W., Bosma, A., Skowron, G., Mayer, K.H., 2001. Designing HIV-1 vaccines to reflect viral diversity and the global context of HIV/AIDS. AIDScience 1, 2, 24-40.

De Groot, A.S., Jesdale, B., Martin, W., Aubin, C.S., Sbai, H., Bosma, A., Lieberman, J., Skowron, G., Mansourati, F., Mayer, K.H., 2003. Mapping cross-clade HIV-1 vaccine epitopes using a bioinformatics approach Vaccine 21, 4486-4504.

De Groot, A.S., Rivera, D.S., McMurry, J.A., Buus, S., Martin, W., 2008. Identification of immunogenic HLA-B7 "Achilles' heel" epitopes within highly conserved regions of HIV. Vaccine 26, 3059—3071.

DeLano, W.L., 2002. The PyMOL Molecular Graphics System DeLano Scientific, Palo Alto, CA, USA. http://www.pymol.org.

Doytchinova, I.A., Guan, P., Flower, D.R., 2006. EpiJen: a server for multi-step T cell epitope prediction. BMC Bioinformatics. 7, 131.

Fenouillet, E., Gluckman, J.C., Jones, I.M., 1994. Functions of HIV envelope glycans. Trends Biochem. Sci. 19,65–70.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids Symp. Ser. 41, 95-98.

Henderson, R.A., Michel, H., Sakaguchi, K., Shabanowitz, J., Appella, E., Hunt, D.F., Engelhard, V.H., 1992. HLA-A2.1-associated peptides from a mutant cell line: a second pathway of antigen presentation. Science 255, 1264-1266.

Hwang, S.S., Boyle, T.J., Luerly, H.K., CULEN, B.R., 1991. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. Science 253, 71-74.

Johnson, R.P., Trocha, A.,Yang, L., Mazzara, G., Panicali, D., Buchanan, T., Walker, B.D., 1991. HIV-1 gag-specific cytotoxic T Lymphocytes recognize multiple highly conserved epitopes: fine specificity of the gag-specific response by using unstimulated peripheral blood mononuclear cells and cloned effector cells. J. Immunol. 147, 1512-1521.

Julenius, K., Molgaard, A., Gupta, R., Brunak, S., 2005. Prediction, conservation analysis and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiology. 15,153-164.

Kandathil, A.J., Ramalingam, S., Kannangai, S., Shoba, D., Sridharan, D., 2005. Molecular epidemiology of HIV. Indian J. Med. Res. 121; 333-344.

Kiepiela, P., Leslie, A.J., Honeyborne, I., Ramduth, D., Thobakgale, C., Chetty, S., et al. 2004. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. Nature 432, 769—75.

Kmieciak, D., Bednarek, I., Takiguchi, M., Wasik, T.J., Bratosiewicz, J., Wierzbicki, A., Teppler, H., Pientka, J., Hsu, S.H., Kaneko, Y., Kozbor, D., 1998. The effect of epitope variation on the profile of cytotoxic T lymphocyte responses to the HIV envelope glycoprotein. Intern. Immun. 10 vol.12, 1789-1799.

Koenig, S., Earl, P., Powell, D., Pantaleo, G., Merli, S., Moss, B., Fauci, A.S., 1988. Group-specific, MHC class I restricted cytotoxic responses to HIV-1 envelope proteins by cloned peripheral blood T cells from an HIV-1-infected individual. Proc. Natl. Acad. Sci. 85, 8638-8642.

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., Detours, V., 2001. Evolutionary and immunological implications of contemporary hiv-1 variation. Br. Med. Bull. 58, 19-42.

Kukuruzinska, M.A, Lennon, K., 1998. Protein N-glycosylation: molecular genetics and functional significance. Crit. Rev. in Oral Biol. Med. 9,415-448.

Kwong, P.D., Wyatt, R., Robinson, J., Sweet, R.W., Sodroski, J., Hendrickson, W.A., 1998. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature 393, 648-659.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M., 1993. PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst. 26, 283-291.

Lichterfeld, M., Williams, K.L., Mui, S.K., Shah, S.S., Mothe, B.R., Sette, A., Kim, A., Johnston, M.N., Burgett, N., Frahm, N., Cohen, D., Brander, C., Rosenberg, E.S., Walker, B.D., Altfeld, M., Yu, X.G., 2006. T cell receptor cross-recognition of an HIV-1 CD8+ T cell epitope presented by closely related alleles from the HLA-A3 superfamily. International Immunology 18, n°7, 1179-1182.

Lüthy, R., Bowie, J.U., Eisenberg, D., 1992. Assessment of protein models with three-dimensional profiles. Nature. 5, 356, 83-55.

MacDonald,K.S., Fowke, K.R., Kimani, J., Dunand, V.A., Nagelkerke, N.J.D., Ball, T.B., Oyugi, J., Njagi,E., Gaur, L.K., Brunham, R.C., Wade, J., Luscher, M.A., Krausa, P., Rowland-Jones, S., Ngugi, E., Bwayo, J.J., Plummer, F.A., 2000. Influence of HLA Supertypes on susceptibility and Reistance to Human Immunodeficiency Virus Type 1 infection. The Journal of Infectious Diseases 181, 1581-1589.

MacKinney, D.M., Skvoretz, R., Livingston, B.D., Wilson, C.C., Anders, M., Chesnut, R.W., Sette, A., Essex, M., NOvitsky, V., Newman, M.J., 2004. Recognition of Variant HIV-1 Epitopes from Diverse Viral Subtypes by Vaccine-Induced CTL. The Journal of Immunology 173, 1941-1950.

Martin, W., Sbai, S., De Groot, A.S., 2003 Bioinformatics tools for identifying class I-restricted epitopes. Methods 29, 289–298.

McCaffrey, R.A., Saunders, C., Hensel, M., Stamatatos, L., 2004. N-linked glycosylation of the V3 loop and the immunologically silent face of gp120 protects human immunodeficiency virus type 1 SF162 from neutralization by anti-gp120 and anti-gp41 antibodies. J. Virol. 78, 3279–3295.

McMichael A.J., Rowland-Jones S.L.I., 2001. Cellular Immune Responses to HIV. Nature. 410, 980–87.

Moore, C.B., John, M., James, I.R., Christiansen, F.T., Witt, C.S., Mallal, S.A. 2002. Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level. Science. 296, 1439-1443.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G., Thornton, J.M., 1992. Stereochemical quality of protein structure coordinates. Proteins 12, 345-364.

Mota-Miranda, A.C., De Oliveira, T., Moreau, D.R., Bomfim, C., Galvão-Castro, B., Alcantara, L.C.J., 2007 Mapping the molecular characteristics of Brazilian human T-cell lymphotropic virus type 1 Env (gp46) and Pol amino acid sequences for vaccine design. Mem. Inst. Oswaldo Cruz 102(6), 741-749.

Novitsky, V.A., Montano, M.A., McLane, M.F., Renjifo, B., Vannberg, F., Foley, B.T., Foley, B.T., Ndung'u, T.P., Rahman, M., Makhema, M.J., Marlink, R., Essex, M., 1999. Molecular cloning and phylogenetic analysis of human immunodeficiency virus type I subtype C: a set of 23 full-length clones from Botswana. J. Virol. 73, 4427–4432.

Oliveira, T., Deforche, K., Cassol, S., Salminemm, M., Paraskevis, D., Seebregts, C., Snoeck, J., van Rensburg, E.J., Wensing, A.M.J., van de Vijver, D.A., Boucher, C.A., Camacho, R., Vandamme, A.M., 2005. An Automated Genotyping System for Analysis of HIV-1 and other Microbial Sequences. Bioinfomatics; 21 (19), 3797-3800.

Page, K.A., Stearns, S.M. Littman, D.R., 1992. Analysis of mutations in the V3 domain of gpl60 that affect fusion and infectivity. Journal of Virology 66, 524-533.

Rambaut, A., Posada, D., Crandall, K.A., Holmes, E.C., 2004. The Causes and Consequences of HIV Evolution. Nature Rev. 5, 52-61.

Rhee, S.Y., Gonzales, M.J., Kantor, R., Betts, B.J., Ravela, J., Shafer, R.W., 2003. Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res 31,298-303.

Russell, K.L., Carcamo, C., Watts, D.M., Sanchez, J., Gotuzzo,E., Euler,A., Blanco,J.C., Galeano,A., Alava, A., Mullins,J.I., Holmes, K.K., Carr, J.K., 2000. Emerging genetic diversity of HIV-1 in South America. AIDS. 14,1785-1791.

Scanlan, C.N., Offer, J., Zitzmann, N., Dwek, R.A., 2007. Exploiting the defensive sugars of HIV-1 for drug and vaccine design NATURE 446, 1038-1045.

Sette, A., Sidney, J., 1999. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. Immunogenetics 50:201.

Shankar, P., Fabry, J.A., Fong, D.M., Lieberman, J., 1996. Three regions of HIV-1 gp160 contain clusters of immunodominant CTL epitopes. Immunol. Lett. 52,23-30.

Sharon, M., Kessler, N., Levy, R., Zolla-Pazner, S., Gorlach, M., Anglister, J., 2003. Alternative conformations of HIV-1 V3 loops mimic beta hairpins in chemokines, suggesting a mechanism for coreceptor selectivity. Structure 11, 225–236.

Sidney, J., Grey, H.M., Southwood, S., Celis, E., Wentworth, P.A., Del Guercio, M.F., Kubo, R.T., Chesnut, R.W., Sette, A., 1996. Definition on a HLA-A3-like supermotif demonstrates the overlapping peptide-binding repertoires of common HLA molecules. Hum. Immunol. 45, 79-93.

Smith, K.D., Lutz, C.T., 1996. Peptide-dependent expression of HLA-B7 on antigen processing-deficient T2 cells. J Immunol 156, 3755-3764.

Soares, E.A.J.M., Martínez, A.M.B., Souza, T.M.S., Santos, A.F.A., Da Hora, V., Silveira, J., Bastos, F.I., Tanuri, A., Soares, M.A., 2005. HIV-1 subtype C dissemination in southern Brazil. AIDS. 19 (suppl 4), S81–S86.

Stephens, H.A.F., 2005. HIV-1 diversity versus HLA class I polymorphism. TRENDS in Immunol. 26, 41-47.

Suberna C., Chavda, A., Griffin, P., Han-Liu, I.Z., Keys, B., Vekony, M.A., Cann, 1994 Molecular determinants of the V3 loop of human immunodeficiency virus type 1 glycoprotein gpl20 responsible for controlling cell tropism. A.J. Journal of General Virology 75, 3249-3253.

Swofford D. L., 2001. PAUP*: phylogenetic analysis using parsimony (*and other methods) Sinauer Associates, Sunderland, Mass.

Takahashi, H., Cohen, J., Hosmalin, A., Cease, K.B., Houghten, R., Cornette, J.L., DeLisi, C., Moss, B., Germain, R.N., Berzofsky, J.A., 1988. An immunodominant epitope of the human immunodeficiency virus envelope glycoprotein gp160 recognized by class I major histocompatibility complex molecule-restricted murine cytotoxic T lymphocytes. Immunology 85, 3105-3109.

Thompson,J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research. 24,4876-4882.

Threkeld, S.C., Wentworth, P.A., Kalams, S.A, Wilkes, B.M., Ruhl, D.J., Keogh, E., Sidney, J., Southwood, S., Walker, B.D., Sette, A., 1997. Degenerate and promiscuous recognition by CTL of peptides presented by the MHC class I A3-like superfamily: implications for vaccine development. J. Immunol. 159, 1648.

Vriend, G., 1990. WHAT IF: a molecular modeling and drug design program. J Mol Graph. 8, 52–56.

WHO-UNAIDS Report from a meeting Vaccine Advisory Comitee Geneva. 2001. Approaches to the development of broadly protective HIV vaccines: challenges posed by the genetic, biological and antigenic variability of HIV-1. AIDS 15, W1-W25

Wyatt, R., Sodroski, J., 1998. The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. Science 280,1884–88.

Xose, M., Suarez, F., Schuster, M.K., 2007. Using the Ensembl Genome Server to Browse Genomic Sequence Data. UNIT 1.15 in Current Protocols in Bioinformatics, Supplement 16.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.

Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.M., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449.

9. Web References:

DiscoTope 1.2 Server:
http://www.cbs.dtu.dk/services/DiscoTope/

Ensembl Genome Server:
http://www.ensembl.org/Multi/blastview

Epijen:
http://www.jenner.ac.uk/EpiJen/

GenBank data base:
http://www.ncbi.nlm.nih.gov/Genbank/

Los Alamos Sequence Data Base:
http://www.hiv.lanl.gov/content/index

NetChop:
http://www.cbs.dtu.dk/services/NetChop/

NetNGlyc 1.0 server:
http://www.cbs.dtu.dk/services/NetNGlyc/

NetOGlyc 3.1 server:
http://www.cbs.dtu.dk/services/NetOGlyc/

PHYRE:
http://www.sbg.bio.ic.ac.uk/phyre/html/index.html

PROCHECK:
http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html

Rega HIV Subtyping Tool version 2.0:
http://www.bioafrica.net/subtypetool/html/

Stanford database:
http://hivdb.stanford.edu/pages/algs/HIVdb.html

TapPred:
http://www.imtech.res.in/raghava/tappred/

Verify 3D:
http://www.doe-mbi.ucla.edu/Services/Verify3D.html

Whatcheck:
http://swift.cmbi.kun.nl/swift/whatcheck/

10. Supplementary Data

Table 1S. Parameter estimates and likelihood scores under models of variable ω ratios among sites.

Nested model pairs [a]	*dN/dS* [b]	Parameter estimates [c]	ℓ	*LTR p-value*
QMHEDISL				
M0: one-ratio(1)	0.3029	ω = 0.30287	-652.319849	0.0216941
M3: discrete (5)	0.3303	*p*0 =0.50091, *p*1 =0.21112, (*p*2 =0.28797) ω0= 0.10227, ω1 =0.55911, **ω2 = 0.55911**	-646.581287	
M1a: neutral (1)	0.4527	*p*0 = 0.68873, (*p*1 = 0.31127) (ω0 = 0.20535), (ω1 =1)	-649.458657	0.13374
M2a:selection(3)	0.4524	*p*0=0.68901, *p*1 =0.21947, (*p*2=0.09151) (ω0 = 0.20522), (ω1 =1), **ω2 = 1**	-647.446801	
M7: beta (2)	0.3327	p= 0.94853 q= 1.88844	-645.518354	0.999
M8: beta&ω (4)	0.0950	p0= 0.97097, p= 1.00919, q= 2.14954 , (p1= 0.02903) ω = 1	-647.877106	
NLTNNVKTI				
M0: one-ratio(1)	0.2910	ω = 0.29100	-1168.363631	> 0.00001
M3: discrete (5)	0.6329	*p*0 = 0.73881, *p*1 = 0.06860, (*p*2 = 0.19258) ω 0 = 0.14995, ω1 = 0.31213, **ω2 = 2.59991**	-1090.288617	
M1a: neutral (1)	0.3082	*p*0 = 0.78013, (*p*1 = 0.21987) , (ω0 = 0.11321), (ω1 =1)	-1102.773805	> 0.00001
M2a:selection(3)	0.5926	*p*0= 0.78031, *p*1 = 0.10448, (*p*2= 0.11521) (ω0 = 0.14544), (ω1 =1), **ω2 = 3.25199**	-1088.967201	
M7: beta (2)	0.3705	p= 0.44989 q= 0.76352	-1115.724392	> 0.00001
M8: beta&ω (4)	0.6497	p0= 0.80706 p= 18.97489 q= 96.34941 (p1= 0.19294) ω = 2.68016	-1091.611148	
CTHQIKPVV				
M0: one-ratio(1)	0.04201	ω = 0.04201	-314.917919	0.754562
M3: discrete (5)	0.0420	*p*0 =0.27863, *p*1 =0.38460, (*p*2 =0. 0.33678) ω0= 0.04204, ω1 =0.04204, **ω2 = 0.04204**	-313.969063	
M1a: neutral (1)	0.0419	*p*0 = 1, (*p*1 = 0) (ω0 = 0.04188), (ω1 =1)	-313.873131	0.400441
M2a:selection(3)	0.0421	*p*0=1, *p*1 =0, (*p*2=0) (ω0 = 0.04213), (ω1 =1), **ω2 = 7.48064**	-312.957942	
M7: beta (2)	0.0433	p= 4.51742 q= 99.00000	-313.894358	0.999
M8: beta&ω (4)	0.0950	p0= 1 p= 4.50930 q= 99 (p1= 0) ω = 1.85724	-315.370443	
CTRPNNNTR				
M0: one-ratio(1)	0.4271	ω = 0.42705	-1165.717128	> 0.00001
M3: discrete (5)	0.5983	*p*0 = 0.19845, *p*1 = 0.37337, (*p*2 = 0.42818) ω 0 =0 , ω1 = 0.15783, **ω2 = 1**	-1088.639749	
M1a: neutral (1)	0.4824	*p*0 = 0.57148, (*p*1 = 0.42852) , (ω0 = 0.09429), (ω1 =1)	-1095.054852	0.00311332
M2a:selection(3)	0.7004	*p*0= 0.57044, *p*1 = 0.28033, (*p*2= 0.14923) (ω0 = 0.11475), (ω1 =1), **ω2 = 2.37620**	-1089.282788	
M7: beta (2)	0.3684	p= 0.17259 q= 0.29584	-1091.597101	0.00170127
M8: beta&ω (4)	0.6354	p0= 0.84961 p= 0.29483 q= 0.51309 (p1= 0.15039) ω = 2.16370	-1085.220722	
AITQACPKV				
M0: one-ratio(1)	0.3549	ω = 0.35489	-682.062480	> 0.00001
M3: discrete (5)	0.3810	*p*0 = 0.63140, *p*1 = 0.29405, (*p*2 = 0.07455) ω 0 = 0.03267 , ω1 = 0.73994, **ω2 = 1.91561**	-640.436294	
M1a: neutral (1)	0.3810	*p*0 = 0.64132, (*p*1 = 0. 0.3586), (ω0 = 0.03482), (ω1 =1)	-643.944089	0.260249
M2a:selection(3)	0.4629	*p*0= 0.63934, *p*1 = 0.29502 (*p*2= 0.06564) (ω0 = 0.03831), (ω1 =1), **ω2 = 2.18443**	-642.597972	
M7: beta (2)	0.3328	p= 0.18308 q= 0.36713	-644.448591	0.282471
M8: beta&ω (4)	0.3997	p0= 0.92703 p= 0.20966 q= 0.54463 (p1= 0.07297)) ω = 1.94651	-643.181441	

a The number after the model code, in parentheses, is the number of free parameters in the ω distribution. ***b*** This *dN/dS* ratio is an average over all sites in the epitopes alignment. ***c*** Parameters in parentheses are not free parameters; ω dN/dS ratio; *p*, proportion of sites in ω; M0-M3: χ^2 df=4; M1a-M2a: χ^2 df=2; M7-