Molecular and structural characterization of HIV-1 subtype B Brazilian isolates with GWGR tetramer at the tip of the V3-loop

Élcio Leal *, Wilson P. Silva, Maria C. Sucupira, L. Mário Janini, Ricardo S. Diaz

Federal University of São Paulo, Rua Pedro de Toledo, 781-16° andar, Vila Clementino, São Paulo–SP, CEP 04039-032, Brazil

ARTICLE INFO

Article history:
Received 6 June 2008
Returned to author for revision 30 June 2008
Accepted 19 August 2008
Available online 23 September 2008

Keywords:
HIV-1
V3 loop
GWGR
Coreceptor
CCR5
CXCR4
3D-structure

ABSTRACT

One of the most intriguing features of the HIV-1 subtype B epidemic in Brazil is the high frequency of isolates exhibiting tryptophan (W) in the tetramer (GWGR) at the tip of the V3 loop. We observed that the frequencies of glutamic and aspartic acids at site 25 of the V3 loop are quite distinct in GWGR isolates compared with viruses with other tetramers. The basic amino acids at sites 11 and 25 are strongly linked with CCR5-to-CXCR4 coreceptor shift. We therefore predicted phenotype usage and found that GWGR isolates are exclusively CCR5-using. Further evidence of this came from intrahost sequences, where basic amino acid substitutions at sites 11 and 25 emerged only in isolates presenting a tryptophan-to-glycine replacement at the tetramer of the V3. In addition, modeled 3D-structures of the V3 loop of GWGR and GGGR in intrahost viruses differ essentially in the binding region of the coreceptor.

© 2008 Elsevier Inc. All rights reserved.

Introduction

Infection with HIV-1 is primarily seen in CD4+ T cells and macrophages, although other cell types can also be infected. The cellular tropism is essentially determined by a series of interactions between the viral envelope glycoprotein (Env) and cell receptors. To enter target cells, HIV-1 typically needs to recognize and bind two membrane receptors. These events are perhaps the most crucial for HIV-1 infection, initially involving the interaction of the envelope with the main receptor (CD4) and then with chemokine coreceptors (Berger et al., 1999; Boyd et al., 1993; Wu et al., 1996). The envelope is a noncovalently associated trimeric glycoprotein composed of surface (gp120) and transmembrane (gp41) subunits (Roux and Taylor, 2007). The envelope interaction with the cell surface causes conformational changes in the viral glycoprotein. It is thought that, following the binding of gp120 to CD4, conformational modifications enable the gp41 subunit to interact with the cell membrane, thereby allowing the virus to enter the host cell (Wei et al., 2003). The V3 region of gp120 plays a fundamental role in various characteristics of HIV. For example, upon HIV infection, there is an immunodominant antibody response directed primarily against the V3 region (Frost et al., 2005). In addition to virus-cell binding, the V3 region also determines which of two chemokine coreceptors (CCR5 or CXCR4) will be exploited for entry (Choe et al., 1996; Coakley et al., 2005; Rizzuto et al., 1998). In roughly 50% of all infected individuals, the virus changes its chemokine receptor usage during the progression of HIV-1 infection (Boyd et al., 1993; Schuitemaker et al., 1992). In the early phase of HIV-1 infection, the viruses isolated are typically the R5 variants, which present tropism for the CCR5 receptors, whereas the X4 variants, which preferentially use the CX4 receptors, emerge later in the infection. In the late phase of HIV-1 infection, the intrahost viral population can present a variety of compositions; it can be composed exclusively of X4 variants, it can consist of variants capable of using both coreceptors (R5X4 variants), or it can present equal numbers of the R5 and X4 variants (Kuiken, 1999; Scarlatti et al., 1997; Shioda et al., 1991).

In European and North American countries, HIV-1 infection has been well characterized and involves a predominance of subtype B strains. In Brazil, subtype B also prevails in the HIV-1 epidemic, although nearly half of the infections caused by the subtype B strain are due to viruses with the unusual GWGR motif in the V3 loop of the env gene. Although GW viruses have sporadically been observed in other countries, such as China, France, the Czech Republic, the Philippines, and Cuba, high frequencies of GW isolates are seen only in Brazil. This suggests that the presence of tryptophan (W) at the tip of the V3 loop, although rare in other countries, is not deleterious to HIV-1. Previous studies have demonstrated that the GW and GP viruses are not phylogenetically distinguishable (Kuiken, 1999; Meng et al., 2002; Morgado et al., 1994; Potts et al., 1993; Santoro-Lopes et al., 2000), whereas the HIV-1 subtype B found in Thailand differs from the subtype B isolates of other regions in terms of their tetramers (Kuiken, 1999). Nevertheless, the evolutionary mechanisms responsible for the fact

* Corresponding author. Fax: +55 11 5081 5394/5571 2130/5579 8226.
E-mail address: e.leal@unifesp.br (É. Leal).

0042-6822/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2008.08.029
that GW and GP viruses are found in equal proportions in Brazilian epidemics have yet to be explored. Therefore, we focused on diverse aspects of the evolution of the GW and GP isolates of subtype B viruses co-circulating in Brazil during a specific period (1983–2000).

We found that the frequencies of amino acids at sites of the V3 loop linked with CCR5-to-CXCR4 shift are quite distinct between the GW and GP viruses. Additionally, GW viruses may use the CCR5 coreceptor exclusively, according to the predicted phenotype based on the V3 sequences. Lastly, the modeled 3-D structure of the V3 loop of the GW and GG variants from intrahost samples differs mostly in the binding region of the coreceptor.

**Results**

**Amino acid composition at the V3 loop**

In order to characterize HIV-1 variants, we analyzed the amino acid substitutions in the V3 loop on a site-by-site basis. Fig. 1 depicts the consensus sequences and the alternative amino acids obtained from the alignments of GWGR viruses and subtype B viruses with other tetramers at the tip of V3 loop (e.g., GPGR, GFGR, GGGR, etc). In addition to the second V3 tetramer site, consensuses also differed at sites 14 and 25 (arrows in Fig. 1). Although a limited number of sequences were analyzed, both viruses presented an extensive number of amino acid substitutions throughout the V3 region. Most of the sites were highly variable, although a great number of the substitutions observed were due to singleton mutations. Notably, some amino acids presented at high frequencies, whereas others were unique substitutions. For example, at site 29, the aspartic acid-to-asparagine substitution reached a high frequency in both viruses, while other substitutions, such as aspartic acid-to-glutamic acid, were less common. This is probably due to the genetic barrier of the ancestral sequence, since replacement of aspartic acid to asparagine involves only one G-to-A transition while the replacement to glutamic acid requires at least one transversion (i.e., GAT-to-GCG or GAC-to-GCA).

![Fig. 1. Amino acid frequencies in the V3 loop of HIV-1. Consensus sequences of GW viruses (panel A) and viruses with other motifs (i.e., GP, GF, GL, etc) at the V3 loop (panel B) are within the dark grey area, and the V3 loop tetramer is underlined. Below each nonconserved site, the specific frequencies of the most common residues are depicted. The alternative amino acids are indicated above the sites of consensus sequences; the number above each residue represents its occurrence in the data set. The numbers inside the parentheses indicate sites 11 and 25, which are important for the R5-to-X4 coreceptor shift. Arrows indicate sites where the two consensuses differ.](image)

**Proportions of aspartic and glutamic acid at site 25 of the V3 loop**

The most noticeable dissimilarity between the GWGR and GXGX viruses was in the proportions of aspartic (D) and glutamic (E) acids at site 25 of the V3 loop of HIV-1. The frequencies of aspartic and glutamic acids at site 25 are shown in Table 1. The proportions were calculated using Fisher’s exact test, which is more appropriate for small sample sizes. The results indicate a significant difference in the proportions of aspartic and glutamic acids at site 25, with a p-value of 0.001, suggesting a strong association between the coreceptor type and the amino acid composition at site 25.

**Table 1**

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Dataset</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>GWGR (n=41)</td>
<td>7</td>
</tr>
<tr>
<td>E</td>
<td>GXGX (n=78)</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>X4 (n=56)</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>R5 (n=368)</td>
<td>182</td>
</tr>
</tbody>
</table>

*Fisher’s two-tailed probability calculated in 2×2 contingency tables.

**Notes**

1. These sequences together with their coreceptor preference status were obtained from the HIV Sequence Database (http://hiv-web.lanl.gov).
PSSM score for each sequence of the GWGR and GXGR viruses using scoring matrices (PSSM) method. To that end, we determined the context of coreceptor usage, as predicted by the position-specific scoring matrices (PSSM) method. This distinction between the two viruses was also observed in the coreceptor shift. We were then interested in determining whether GXGR viruses since these sites are linked with CCR5-to-CXCR4 acids at sites 11 and 25 of the V3 loop region between the GWGR and X4 viruses (da Silva, 2006; Hung et al., 1999). Consequently, for comparative purposes we also analyzed 268 sequences of R5-using viruses and 56 sequences of X4-using viruses with known coreceptor status obtained from the Los Alamos HIV database (http://hiv-web.lanl.gov). Likewise, a great variety of amino acids were observed at site 25 of the HIV-1 V3 loop, and, in most viruses, aspartic acid and glutamic acid was present in high proportions. Equivalent quantities of the ratio of aspartic acid to glutamic acid (D:E) were observed between the GXGX and X4 viruses \((p = 0.629, \text{Table 1})\). However, the overall quantity of aspartic acid plus glutamic acid (D+E) in the GXGX viruses was roughly 73%, whereas in the X4 viruses it comprised only 41% of the amino acids found at site 25. The fact that the quantities of aspartic acid and glutamic acid observed in the GXGX virus and the X4 variant were similar (1:1) might indicate that the two acids have a comparable impact on the fitness of HIV-1. On the other hand, in the GWGR and R5 viruses there is an imbalance in the proportion of aspartic and glutamic acids at site 25 of the V3 loop. For example, the quantity of glutamic acid was disproportionately high \(23(41)\) in the GWGR viruses compared to the proportion of glutamic \(26(78)\) in the GXGX viruses \((p = 0.007, \text{Table 1})\), indicating that the alternative amino acids might be detrimental to the GWGR viruses. Conversely, in the R5 viruses the quantity of aspartic acid at site 25 was high \(182(268)\) in comparison with the proportion of aspartic acid \(11(56)\) observed in the X4 viruses \((p = 0.001, \text{Table 1})\).

Predicted coreceptor usage

We noticed some differences regarding the frequencies of amino acids at sites 11 and 25 of the V3 loop region between the GWGR and GXGR viruses since these sites are linked with CCR5-to-CXCR4 coreceptor shift. We were then interested in determining whether this distinction between the two viruses was also observed in the context of coreceptor usage, as predicted by the position-specific scoring matrices (PSSM) method. To that end, we determined the PSSM score for each sequence of the GWGR and GXGR viruses using the x4r5 matrix (Jensen et al., 2003). We also determined the PSSM scores of 268 sequences of R5 viruses and 56 sequences of X4 viruses. The overall PSSM scores of the GWGR, GXGX, R5, and X4 data sets are plotted in Fig. 2. The medians, 25–75% intervals, and range of non-outliers for the GWGR, GXGX, and R5 viruses have negative values that did not differ significantly between these data sets. On the other hand, the X4 viruses have positive values that are significantly distinct from the other data sets. This is expected because high PSSM scores are indicative of CXCR4-using viruses. In this context, we detected six isolates \(6(78)\) of the GXGX data set with high PSSM scores (open diamonds in Fig. 2) that are within the non-outlier range for the X4 viruses, therefore suggesting that these isolates are CXCR4-using viruses.

Intrahost analysis of the V3 loop

The above analyses were performed on sequences from unlinked individuals (interhost) and therefore reveal the evolution of GW viruses at the population level. To further our understanding of the evolution of GW viruses during intrahost infection, we studied sequences of a region of the env gene from an HIV-1-infected patient. These sequences were from molecular clones sampled during antiretroviral therapy, and again at 2 weeks and 12 weeks after structured therapy interruption. In a previous study, we noted that complete discontinuation of therapy-induced recrudescence of wild-type drug-sensitive viruses (Silva et al., 2006). Specifically, drug-resistant viruses with a GGGR tetramer at the tip of the V3 loop predominated during antiretroviral therapy, whereas at 12 weeks after the interruption of therapy, the intrahost population was completely replaced by viruses presenting the GWGR tetramer at the tip of the V3 loop. However, the less glycosylated pattern of X4 variants has been associated with structural modification of the V3 loop region during coreceptor shift (Ogert et al., 2001). Our intrahost analysis indicated that the GW sequences presented a lower proportion of N-linked glycosylation sites than did the GG sequences. In the GW sequences, the proportions of N-linked glycosylation, specifically at two sites prior to the V3 loop, were 60% and 50%.
respectively, compared with 100% at both sites in the GG sequences. In addition, the GG sequences also presented basic amino acids at sites 11 and 25 of the V3 loop that typically confer the status of CXCR4-using to HIV-1. In contrast, the GW sequences had an uncharged (serine) and negative amino acid at V3 loop sites 11 and 25, respectively.

Intrahost V3 loop evolution

For the intrahost HIV-1 population, the phylogenetic analysis revealed a clearer scenario concerning the dynamics of the GW and GG variants. As can be seen in Fig. 3, the maximum a posteriori (MAP) phylogenetic tree shows that the GW sequences formed two clusters located near the root, whereas the GG sequences formed a unique and well-defined cluster distant from the root. We also used the maximum likelihood approach to reconstruct the ancestral sequences of this intrahost population. This reconstruction (based on a midpoint rooted MAP tree) showed that the most recent common ancestral (MRCA) virus of the intrahost population most likely had a GWGR tetramer at the tip of the V3 loop (Fig. 3). In addition, the reconstructed ancestral states showed an extensive number of amino acid substitutions, including basic amino acids, at sites 11 and 25 (sites 30 and 44, respectively, in our alignment), that emerged only in the lineage of GG viruses. Notably, most of the changes in the GG lineage, including an increase in the number of positive charged amino acids, occurred within the V3 loop.

Modeled structure of GW and GG V3 loop

To gain additional insights about the changes associated with the GWGR to GGGR shift, we evaluated the characteristics of the three dimensional (3D) structure of the V3 loop predicted from these variants. To construct the 3D models, we used the ancestral sequences previously reconstructed with ML analysis. Initially, a blast search was performed in order to identify the best template structure. The best-scored structure was from Gp120 of the HIV-1 JR-FL variant (PDB=2b4c). Next, we extracted the V3 loop region from the 2b4c structure and used it as a backbone template to model the side chains of the GW and GG ancestral sequences. Our modeled structures were quite similar to the 2b4c template, and the quality of these models was evaluated by comparing the Ramachandran plot of the V3 loop of the template with plots of the GW and GG structures (data not shown). The V3 loop structure of the GW and GG variants indicated that they were similar to each other, and differences in the structure were due to side chains of some amino acids, such as the tetramer at the tip of the V3 tetramer (grey region in Fig. 4). Additionally, the most striking difference between the V3 structures of the GW and GG variants was

Fig. 3. Phylogenetic tree of intrahost viruses. Maximum a posteriori tree of HIV-1 from sequences of an individual infected by GWGR lineage that later was replaced by a GGGR virus. The numbers above the branches indicate the posteriori probability support for each cluster of the tree. Grey areas indicate the main clusters. (A) Maximum likelihood ancestral reconstruction of amino acid substitutions that occurred in the lineages are indicated by one-letter amino acid code, along with the position in the alignment and the charge residues. The dotted area indicates amino acid changes with residues inside the V3 loop region. The sites 11 and 25 of V3 are underlined. The arrow indicates the tryptophan (W)-to-glycine (G) replacement in the V3 loop tetramer. The sequence B.U5.98.1058_11 (AY331295) was used as outgroup.
seen in the coreceptor-binding region of the V3 loop (stem region of the V3 loop in Fig. 4). This region has a hydrophilic base formed by uncharged asparagines at sites 6 and 7 of the V3 loop (depicted in green in Fig. 4), where CCR5 binds and interacts predominantly with the asparagine at site 7 (N7) (Huang et al., 2007). In our case, the side chains of amino acids found in the V3 loop of the GG variants had a great impact on the structure of the pocket region. For example, the side chain of glutamic acid (E) at site 24 of the V3 loop found in the GG variants partially occluded the binding pocket (cyan in panel B of Fig. 4). Interestingly, a replacement from asparagine (N) to lysine (K) at site 7 of the V3 loop was observed in the GG variants (magenta in panel B of Fig. 4), which could completely abrogate CCR5 binding.

**Discussion**

Our study indicates that the GWGR and GXGX viruses were quite distinct in terms of the frequencies of amino acids at V3 sites 14 and 25 (Fig. 1). The amino acid most frequently identified at site 14 of the GWGR viruses was methionine (61%), whereas, in the GXGX viruses, that distinction was held by isoleucine (54%). This disparity between the GWGR and GXGX viruses in terms of the proportions of methionine and isoleucine might be related to the founder effect, and the impact that such a disparity has on the structure of the pocket region. For example, the side chain of glutamic acid (E) at site 24 of the V3 loop found in the GG variants partially occluded the binding pocket (cyan in panel B of Fig. 4). Interestingly, a replacement from asparagine (N) to lysine (K) at site 7 of the V3 loop was observed in the GG variants (magenta in panel B of Fig. 4), which could completely abrogate CCR5 binding.

![Fig. 4. Structures of the V3 loop of GWGR and GGGR variants. Models of 3-D structures of the V3 loop of ancestral sequences of GWGR and GGGR variants from intrahost samples are depicted (panels A and B, respectively). The models were based on the Gp120 structure of a HIV-1 subtype B isolate (PDB=2b4c). The tetramer at the tip of the V3 loop is depicted in grey, and the hydrophilic base of coreceptor-binding region (N6N7) is depicted in green in the structures. The replacement from asparagine (N) to lysine (K) at site 7 of V3 is also depicted in magenta (panel B). Likewise, replacement of Glycine (G) to Glutamic (E) at site 24 of the V3 loop in the GG variants is depicted in cyan (Panel B) in the binding pocket of the V3 structure. Sites 11 and 25 of V3 usually linked with CCR5-to-CXCR4 coreceptor shift are also depicted (yellow) in the structures.](image-url)
aspartic acid plus glutamic acid (D+E) in the GXGX viruses was 73% (57/78) of the total amount of amino acids observed at site 25 of the V3 loop, but in the X4 viruses, the proportion was 41% (23/56). The lower proportion of aspartic and glutamic acids in the X4 variant is in agreement with the observation that the R5-to-X4 shift is accompanied by basic mutations at sites 11 and 25 of the V3 loop, the so-called 11/25 rule. However, if the glutamic acid–aspartic acid equilibrium is principally driven by selection based on coreceptor usage, then both acids are equally advantageous at site 25. Therefore, the different replicative fitness reported by da Silva (2006) might be neglected in HIV-1, since glutamic and aspartic acid are maintained at high frequencies in GXGX viruses and X4 variants. In addition, the imbalance in favor of glutamic acid at site 25 in GWGR viruses might indicate distinct preferences in coreceptor usage by viruses with W at the tip of V3 loop. Alternatively, immune pressure exerted on GWGR viruses could differ drastically from that observed in GXGX viruses; however, since the GWGR and GXGX viruses are co-circulating in the same host population, it is unlikely that these viruses would elicit distinctive host responses to the infection.

The intrahost samples showed that HIV-1 variants containing basic mutations at sites 11 and 25 also replaced the larger tryptophan with the tiny glycine. Although the GG viruses were not empirically characterized by their coreceptor usage, the presence of basic amino acids at both sites typically confers HIV-1 X4 phenotype status (De Jong et al., 1992). An additional feature that is considered a strong selective pressure is the variation outside epitopes at N-linked glycosylation sites mediated by humoral immune escape antibody attack: the glycan shield mechanism (Wei et al., 2003). During the initial phase of HIV-1 infection, the virus needs to escape the antibody attack, and this is achieved by acquisition of mutations at N-linked glycosylation sites, resulting in structural modifications at the surface of sugar moieties on the viral envelope. These mutations permit the HIV to avoid the neutralizing antibodies in order to bind to the viral surface proteins, thus allowing the virus to escape from the humoral immune surveillance (Dacheux et al., 2004). Although this mechanism is not directly associated with viral diversification, it is strongly linked to host immune pressure. In addition, we became interested in this issue because the reduced glycosylation in the X4 variants had been associated with the structural modification of the V3 loop during the coreceptor shift (Ogert et al., 2001). The proportion of N-linked glycosylation sites also changed with the emergence of GG viruses. It has been demonstrated that this sort of modification within or near variable regions of the env gene affects HIV-1 coreceptor usage and cell tropism (Ogert et al., 2001). Consequently, the amino acid modifications observed in our analysis might be linked to the R5-to-X4 shift.

Binding of gp120 to CD4 also causes conformational modifications that project the V3 loop to interact with the cell coreceptor, and this protrusion renders the V3 loop flexible in order to interact with the coreceptor (Huang et al., 1999; Kwong et al., 1998). It has been shown that intrahost evolution of the V3 loop structure is highly diverse and structurally flexible throughout the infection (Watabe et al., 2006). Therefore, structural modifications seem to be related to immune pressure and interactions with the cell coreceptor (Sander et al., 2007). Indeed, monoclonal antibodies directed against the D19 epitope within the V3 region neutralize only the X4 variants (Lusso et al., 2005). Consequently, inaccessibility of this antibody to R5 might indicate that there are significant V3 loop conformational differences between these two variants (Sander et al., 2007). In fact, it has been demonstrated that intricate electrostatic interactions occur among amino acid residues in order to preserve the proper tertiary structure of the V3 loop (Rosen et al., 2006). This sort of structural modification was also observed in our intrahost analysis, where the 3D-structures of V3 in the GW and GG variants differ primarily in the coreceptor-binding pocket region. Furthermore, it has been demonstrated that CCR5 interacts with the conserved sequence (94N56N67) of the V3 loop, and that binding of this coreceptor is blocked when N7 is replaced by charged amino acid (Huang et al., 2007). Similarly, our intrahost analysis also showed that the polar asparagine (N) at site 7 of the V3 loop was replaced by a positive lysine (K) in the GG variants.

The results of the present study show that the composition of charged amino at sites 11 and 25 of the V3 loop is quite distinct between the GWGR and GXGX viruses. Consequently, the coreceptor preference of these viruses might be equally affected. In addition, our intrahost analysis showed that basic residues were present at sites 11 and 25 of the V3 loop only when tryptophan at the GWGR tetramer (GW viruses) was replaced by glycine in the GG viruses. GWGR isolates were characterized by a shortage of basic amino acids at sites 11/25 of the V3 loop and by the absence of X4 variants based on the phenotype prediction. In addition, 3D models of intrahost viruses showed structural changes in the CCR5 binding pocket region when the tryptophan at the tip of V3 was replaced by a glycine in variants within the host. Although we have analyzed a limited number of sequences and the tropism replacement was observed in only one individual, our findings suggest that there are constraints on the envelope of HIV-1 to use the CXC4 coreceptor in strains with the GWGR tetramer at the tip of the V3 loop. Consequently, one way to maintain the proper tertiary conformation and to use the CXC4 coreceptor would be the replacement of the tryptophan.

Finally, our results belie the concept that GWGR viruses are less pathogenic, based on the rate of disease progression, than are other subtype B strains in Brazil (Santoro-Lopes et al., 2000). Since X4 variants are generally associated with a rapid decline in CD4+ cell counts and accelerated progression to AIDS (Schulmeister et al., 1992; Shankarappa et al., 1999), any comparisons between the GWGR and GPGR viruses should incorporate the profile of coreceptor usage. The potential pitfalls resides in the fact that GWGR viruses might only have R5 variants, whereas GPGR viruses might have R5 and X4 variants indiscriminately. Consequently, there will be a bias promoting the incorrect notion that infection with a GPGR virus results in lower CD4+ cell counts or more rapid progression to AIDS than does infection with a GWGR virus.

In summary, we thoroughly analyzed sequences of HIV-1 subtype B with the GWGR tetramer at the tip of the V3 loop from viruses circulating in Brazil. We find a considerable difference between the GWGR and GXGX viruses in terms of the amino acid profile in the V3 region, particularly at sites 11 and 25, which are associated with the R5-to-X4 coreceptor shift. Notably, the scarcity of basic amino acids at sites 11 and 25 within the V3 loop of GWGR viruses might indicate limited usage of the alternative coreceptor. Further evidence of the constrained use of the CXC4 coreceptor by the GWGR viruses came from our intrahost analysis, in which the tryptophan in the tetramer was found to be replaced by glycine only in isolates with X4 amino acid profiles (i.e., basic amino acids at sites 11 and 25 of the V3 loop). In view of these facts, we hypothesize that the V3 loop of GWGR viruses might prevent the electrostatic interaction of charged amino acids necessary to maintain the proper V3 loop structure after coreceptor binding.

**Materials and methods**

**Samples**

The informed consent for collecting blood samples and the protocol of this study was approved by the Federal University of São Paulo. We generated 41 sequences from GWGR viruses and 78 sequences from GXGX viruses, all taken from a 219-bp fragment of the V3 region of the env gene (gp120) of HIV-1. These sequences were all taken from Brazilian individuals clinically asymptomatic, naïve to drug therapy, and with CD4 levels above 200 cells/mm³. The samples were collected between 1983 and 2000. The remaining sequences were obtained from the HIV Sequence Database (http://hiv-web.lanl.gov). We also analyzed 32 sequences; these were obtained over time from a single
individual under therapy with CD4 levels below 200 cells/mm$^3$, from a 270-bp fragment of the V3 region of HIV-1 subtype B, sampled during antiretroviral therapy and again at 2 weeks after and 12 weeks after structured therapy interruption. These fragments of the V3 region of the env gene (gp120) of HIV-1 were amplified and sequenced using primers and conditions previously described (Silva et al., 2006).

Sequence alignment

The sequences were initially aligned using the ClustalX program (Thompson et al., 1997). All sites with deletions and insertions were then excluded in order to preserve the reading frames of the genes. After this editing process, the sequences were manually aligned using the SE-AL program, version 2.0 (http://evolve.zoo.ox.ac.uk/software/).

Prediction of coreceptor usage

In addition to the presence of basic amino acids at sites 11 and 25 of the V3 loop (11/25 rule), we also used the position-specific scoring matrices (PSSM) method to analyze our sequences (Jensen et al., 2003). The PSSM detects nonrandom distributions of amino acids in a set of sequences that may be associated with some empirically determined property (for example, coreceptor usage) and constructs a matrix. The matrix is then used to compare the amino acids of query sequences and PSSM assigns scores to each query sequence. Consequently, these scores can be used to predict the likelihood of a query sequence having the property of interest. We predicted the coreceptor usage of our sequences using a PSSM matrix ($\times$4+5) constructed with the V3 region of HIV-1 subtype B with a known X4 or R5 phenotype (http://mullinslab.microbiol.washington.edu/computing/pssms/). In this context, high scoring sequences are more likely to use the CCR4 coreceptor, whereas low scoring sequences are more likely to use the CCR5 coreceptor.

Proportion of N-linked glycosylation sites

The N-linked glycosylation sites were determined by identifying the Nx[S/T]Y pattern (where X can be any amino acid). Amino acids, such as proline, in the X and Y position can be important determinants of N-linked glycosylation efficiency (Zhang et al., 2004). The analyses were performed using the web interface available at http://www.hiv.lanl.gov/content/hiv-db/GLYCOSITE/glycosite.html.

Phylogenetic inference

To infer the intrahost phylogenetic tree, we used Bayesian inference, assuming GTR and a gamma correction model, as is implemented in the MrBayes program, version 3.1.2 (Huelsenbeck & Ronquist, 2001). For this analysis, two independent Markov chain Monte Carlo (MCMC) runs were used for 10$^6$ generations with the initial 10% of each run discarded as burn-in. The convergence was evaluated using the TRACER software, version 1.2 (http://beast.bio.ed.ac.uk/), and runs were accepted when all parameters presented the effective sample size number (ESS) greater than 100. The ancestral reconstruction of the V3 loop sequence used the model 2 (M2: selection) that allows different proportions of conserved sites ($d_{60}/d_{61} \approx 0$), neutral sites ($\omega = 1$), and an additional class of sites, with its $\omega$ ratio (which can be $\geq 1$) estimated directly from the data. We then mapped the inferred sites on the phylogenetic intrahost tree. This reconstruction was performed using the CODEML program from the PAML v.3.14 package (Yang et al., 2000).

3-D structure prediction

Ancestral reconstructed sequences from intrahost samples were used to build the V3 structures of the GWGR and GGGR viruses. In order to identify structural homologues in the protein data bank (http://www.pdb.org), we used sequences of 35 amino acids (V3) and performed psi-BLAST searches using the Meta-Set at Centre de Biochimie Structurale (http://bioserv.cbs.cnrs.fr/HTML_BIO/frame_meta.html). Based on these searches, the structure 2bc4, corresponding to the protein Gp120 of a HIV-1 CCR5-using variant, was retrieved, and from this structure we then extracted the V3 loop region in the G chain and used it as a template to build the models. The C-\alpha backbone of the 2bc4 V3 loop structure was used to model the side chain coordinates of the ancestral sequences of the GWGR and GGGR variants from intrahost samples of HIV-1 using the SCRWL software (Canutescu et al., 2003). Visualization and editing of the structures were done using the PyMOL software (http://www.pymol.org).

Acknowledgments

We would like to thank Hirohisa Kishino for his valuable comments and suggestions on this work. We also would like to thank two anonymous reviewers for their criticisms and comments that have greatly improved this work. EL is a postdoctoral researcher supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo, Foundation for the Support of Research in the State of São Paulo; grants no. 04/10372-3, 07/52841-8).

References


