BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net *Hypothesis*

Volume 6(2)

HIV-1 envelope accessible surface and polarity: clade, blood, and brain

Gopichandran Sowmya[1, 2], Gunasagaran Shamini[1, 2], Sathyanarayanan Anita[1], Meena Sakharkar[3], Venkat Mathura[4, 5], Hector Rodriguez[6], Andrew J Levine[7, 8], Elyse Singer[7, 8], Deborah Commins[7, 9], Charurut Somboonwit[10, 11], John T Sinnott[10, 11], Harcharan S Sidhu[2], Ganapathy Rajaseger[12], Peter Natesan Pushparaj[13], Pandajarasamme Kangueane[1, 2], Paul Shapshak[11, 14] *

[1]Biomedical Informatics, Pondicherry 607402, India; [2]Aimst University, 08100 Semeling, Malaysia; [3]Graduate School of Life and Environmental Sciences, University of Tsukuba, Japan; [4]Archer Pharmaceuticals, Sarasota, Florida, USA; [5]Roskamp Institute, 2040 Whitfield Avenue, Sarasota, FL 34243, US; [6]Department of Biology, University of Miami, Coral Gables, FL 33146; [7]National Neurological AIDS Bank, UCLA School of Medicine, Westwood, CA 90095; [8]Department of Neurology, UCLA School of Medicine, Westwood, CA 90095; [9]Department of Neuropathology, USC Keck School of Medicine, Los Angeles, CA90089; [10]Clinical Research Unit, Hillsborough Health Department, Tampa, Florida 33602; [11]Division of Infectious Disease and International Medicine, Tampa General Hospital, USF Health, Tampa, FL 33601; [12]Defense Science Organization National Laboratories, Singapore; [13]Glasgow Biomedical Research Centre, Faculty of Medicine, Univ. of Glasgow, Glasgow, UK; [14]Department of Psychiatry & Behavioral Medicine, University of South Florida, College of Medicine, Tampa, FL 33613; Paul Shapshak - Email: pshapshak@gmail.com; Phone: 843-754-0702; Fax: 813-844-8013; *Corresponding author

Received March 9, 2011; Accepted March 11, 2011; Published March 22, 2011

Abstract:
The human immunodeficiency virus type-1 (HIV-1) gp160 (gp120-gp41 complex) trimer envelope (ENV) protein is a potential vaccine candidate for HIV/AIDS. HIV-1 vaccine development has been problematic and charge polarity as well as sequence variation across clades may relate to the difficulties. Further obstacles are caused by sequence variation between blood and brain-derived sequences, since the brain is a separate compartment for HIV-1 infection. We utilize a three-dimensional residue measure of solvent exposure, accessible surface area (ASA), which shows that major segments of gp120 and gp41 known structures are solvent exposed across clades. We demonstrate a large percent sequence polarity for solvent exposed residues in gp120 and gp41. The range of sequence polarity varies across clades, blood, and brain from different geographical locations. Regression analysis shows that blood and brain gp120 and gp41 percent sequence polarity range correlate with mean Shannon entropy. These results point to the use of protein modifications to enhance HIV-1 ENV vaccines across multiple clades, blood, and brain. It should be noted that we do not address the issue of protein glycosylation here; however, this is an important issue for vaccine design and development.

Keywords: HIV-1, clades, blood, brain, accessible surface area, compositional polarity, Shannon entropy, gp120, gp41, gp120-gp41 complex, gp160, ENV, trimer, vaccine.

Abbreviations: HIV-1 - human immunodeficiency virus type 1; AIDS - acquired immunodeficiency syndrome; ENV - envelope; gp160 - 160,000d glycoprotein; gp120 - 120,000d glycoprotein; gp41 - 41,000d glycoprotein; LANL - Los Alamos National Laboratories; PDB- Protein Data Bank; HVTN - STEP HIV vaccine trial; AA - amino acids; MSA - multiple sequence alignment; ASA - accessible surface area; SNPs - single nucleotide polymorphisms; HAART - Highly Active Antiretroviral Therapy; CCR5 - C-C chemokine receptor type 5; CNS - central nervous system; HIVE - HIV encephalitis; P - polarity; NP - non-polarity; CTL - cytotoxic T lymphocyte. NIAID - National Institute of Allergy and Infectious Diseases

Background:
The design of an effective HIV-1/AIDS vaccine is still in development and its progress is debated **[1]**. The recent results of the HVTN 502/Merck V520-023 study, using three recombinant adenovirus-5 (rAD5) vectors expressing Ad5-gag, Ad5-pol and Ad5-Nef, developed jointly by NIAID and Merck Corporation **[2]** and the use of priming injections of recombinant canary pox vector (ALVAC-HIV) with booster injections of gp120 subunit (AIDSVAX-B/E) in a Thailand trial **[3]** have stimulated further research and strategies in the pursuit of HIV-1 vaccines. Studies of back-to-basics sequence structure relationships coupled with immunological characterization are needed and the HIV-1 ENV remains a key potential target for vaccine development **[4-7]**.

The study of charged AA in the ENV protein is of importance because of protein-protein surface interactions. These include gp120 and gp41 self- and

inter-subunit interactions, inner (core)-outer (solvent) molecular interactions, receptor and co-receptor attachment, fusion required for viral entry, and antibody and CTL immunity. Another key issue is that although AA sequences may vary widely within and among HIV-1 clades, there may be immunologically conserved three-dimensional structures that provide foci for improved vaccine development **[6-10]**.

AA positive selection is a consequence of the effects of immunity on HIV-1 ENV evolution. In 2007, the LANL HIV sequence database was examined and AA positive selection sites in the ENV protein were identified. Asian isolates had a higher positive selection level than North American isolates. The C3, C4, and C5 conserved domains had most of the positive selection sites detected and C1 and C2 were primarily positive selection-free **[11]**. In other studies, ClustalX was used to align 300 sequences of gp160 and located AAs with a high degree of conservation that were in proximity to one another in three-dimensional maps. For example, in HIV-1 clade A, conserved AA occurred at positions including 32, 137, 441, and 915. Several AA were classified according to their polarity or non-polarity at such conserved sites **[12]**.

Each component of the gp120-gp41 complex has specific functions. For example, anchoring the complex occurs via the gp41, a transmembrane protein **[13]**. The gp120 V3 variable region has long been considered crucial for ENV function. The V3 variable region binds to CCR5 or CXCR4 cell surface coreceptors and contains conserved regions including a band, arch, and hydrophobic core **[14]**. HIV-1 gp41 N- and C-domains mediate virus-membrane fusion. AA sequence residues 512-681 from 862 isolates were analyzed in HIV-1 clades A, B, C, D, E, F, G, H, I, J, and O. A highly conserved segment GIVQQQ on the C-terminal of the C-domain was identified that is involved in the formation of the three interfaces between neighboring helices in the trimer **[15]**. The HIV-1 gp41 amino-terminal region is a pre-transmembrane domain. It contains an amphipathic-at-interface sequence that is non-polar (aromatic AA-rich), and is conserved among several viral strains. The amphipathic-at-interface sequence also includes a beta-turn structure with nonhelical extended region. Interaction of the amphipathic-at-interface sequence with the fusion peptide region reduces its fusion ability **[16]**. Additional studies from 357 HIV-1 clades A, B, C, and D also indicated that the gp41 C-terminal tail loop and three beta-sheet membrane-spanning domains are involved in membrane fusion **[17]**.

In addition to ENV AA sequence and charge studies, Shannon entropy is a measure of diversity of AA sequences; the higher the entropy the greater is sequence diversity **[18]**. For example, during investigations of HIV-1 vaccine development, Shannon entropy was used to assess the intra- and inter-clade sequence variation of proteomes of HIV-1 clades A1, B, C, and D. Mean entropies were compared for strings of AA sequences and used to identify protein regions with little diversity that harbored epitopes **[19, 20]**. Entropy was also used to pinpoint clade sequence differences. For example, between clades B and C for HIV-1 gp120, amphipathicity was maintained whereas there was elevated entropy at the polar face of the C3 region alpha2-helix. In clade B there was increased hydrophobicity and in clade C, V4 loops were shorter **[21]**. Entropy analysis also helped identify protein regions with little AA variation that harbored CTL epitopes; these regions frequently occurred in alpha helices **[19, 22]**. Shannon entropy increased for V3 regions from patients treated with CCR5 antagonists vs. baseline. This may be due to treatment resistant viruses being able to produce a wider range of sequence variation than baseline viral strains **[23]**.

The loss of effectiveness of immunological and drug therapy against HIV-1 and difficulties with producing an effective vaccine are due in part to viral immune escape and protein sequence diversity. However, a study of protein structure, selection, and sequence diversity of HIV-1 proteins demonstrated a ceiling to the diversity reachable by HIV-1 despite its high mutation rate **[24, 25]**. Moreover, concomitant with the sequence variability ceiling, variability tended to occur at restricted locations in HIV-1 proteins including in ENV. Entropy studies for both clades B and C demonstrated that increased entropy occurred at AA sites with less constraint, and low entropy at sites with greater constraint. Entropies of AA sites in the protein core were lower than for sites on the protein surface. This is consistent with a paradigm in which loops have greater solvent accessibility than the more constrained core AAs. In this context, protein-protein interaction regions are considered solvent inaccessible (hydrophobic) **[24, 25]**.

Since the early 1990's, phylogenetic analyses supported the paradigm of brain as a reservoir for HIV-1 infection sequestered from blood. In studies of the V3 region, both polar and non-polar AA residues were prevalent for brain and blood with predominantly negative and neutral AA for the brain. Entropy calculations for HIV-1 derived from six patients indicated lower entropy in V3 sequences from brain vs. blood. Thus, the complex sequence and structure relationships for HIV-1 sequences in brain need to be dealt with as well, for vaccine design and production **[26-32]**.

Materials and Methodology:

Datasets:

Structures:

The structural data for gp120 **(Table 1 see Supplementary material)** and gp41 **(Table 2 see Supplementary material)** were obtained from the PDB [http://www.pdb.org/pdb/home]. This dataset was created using the PDB interface by keyword search followed by manual curation. It should be noted that gp120 structures are available in the ligand-bound state.

Sequences:

Gp120 and gp41 sequences (from blood and brain) were downloaded from the LANL database [http://www.hiv.lanl.gov/] **[33]**. These sequences represent multiple HIV-1 clades from different geographical locations including Africa, Asia, North America, South America, Europe, and Oceania. Several sequences were also derived from the PDB structural database [http://www.pdb.org/pdb/home].

Sequence Alignment:

We developed the multiple sequence alignment (MSA) for gp120 **(Figure 1)** and gp41 **(Figure 2)** sequences with known structures **(Tables 1 and 2)** in Protein databank (PDB). The alignment was performed using Clustal W with a gap-opening penalty of 10 and a gap extension penalty of 0.2 **[34]**.

Structure analysis:

Superimposition:

The gp120 **(Table 1)** and gp41 **(Table 2)** structures in its monomer and trimer states **(Figure 3)** were superimposed using SPDBV (Swiss PDB Viewer version 3.7).

Solvent Accessibility:

The solvent accessibility of residues in gp120 and gp41 structures was assessed using ASA calculations **(Figure 4)**. ASA was calculated using the Lee and Richards **[35]** algorithm implemented in the software, Surface racer **[36]**. The probe radius used was 1.4 Å for the calculation of ASA.

Compositional polarity:

We used percent compositional polarity to estimate sequence variations among known clades. This allows calculation of percent polarity range among clades, within which the sequences vary. The percent compositional polarity (S, T, N, Q, H, Y, D, E, R, W, C and K) and non – polarity (G, A, P, V, I, L, F, M residues) were calculated for gp120 and gp41, which included blood and brain sequences **(Figure 5)**. It should be noted that W and C were included in the polar group due to their partial polar property.

Shannon Entropy:

Shannon entropy for each AA residue was calculated as –SUM {Paa.log(Paa)}. Paa is the proportion of each AA in its respective site **[37]**. This equation is based on a calculation of informational entropy and the LANL methods have been described in detail **[38, 39]** (http://www.hiv.lanl.gov/content /sequence/ENTROPY/entropy_one.html).

Polarity range and Shannon Entropy relation:

Pearson correlation of Polarity range and Shannon Entropy was calculated using Microsoft Excel **(Figure 6)**. The statistical significance analysis between Shannon entropy and polarity of gp120 and gp41 sequences in blood and brain were calculated using Two Way ANOVA. In addition, the post test for linear trend has been applied to confirm the linear regression between the groups using GraphPad Prism (version 5) software.

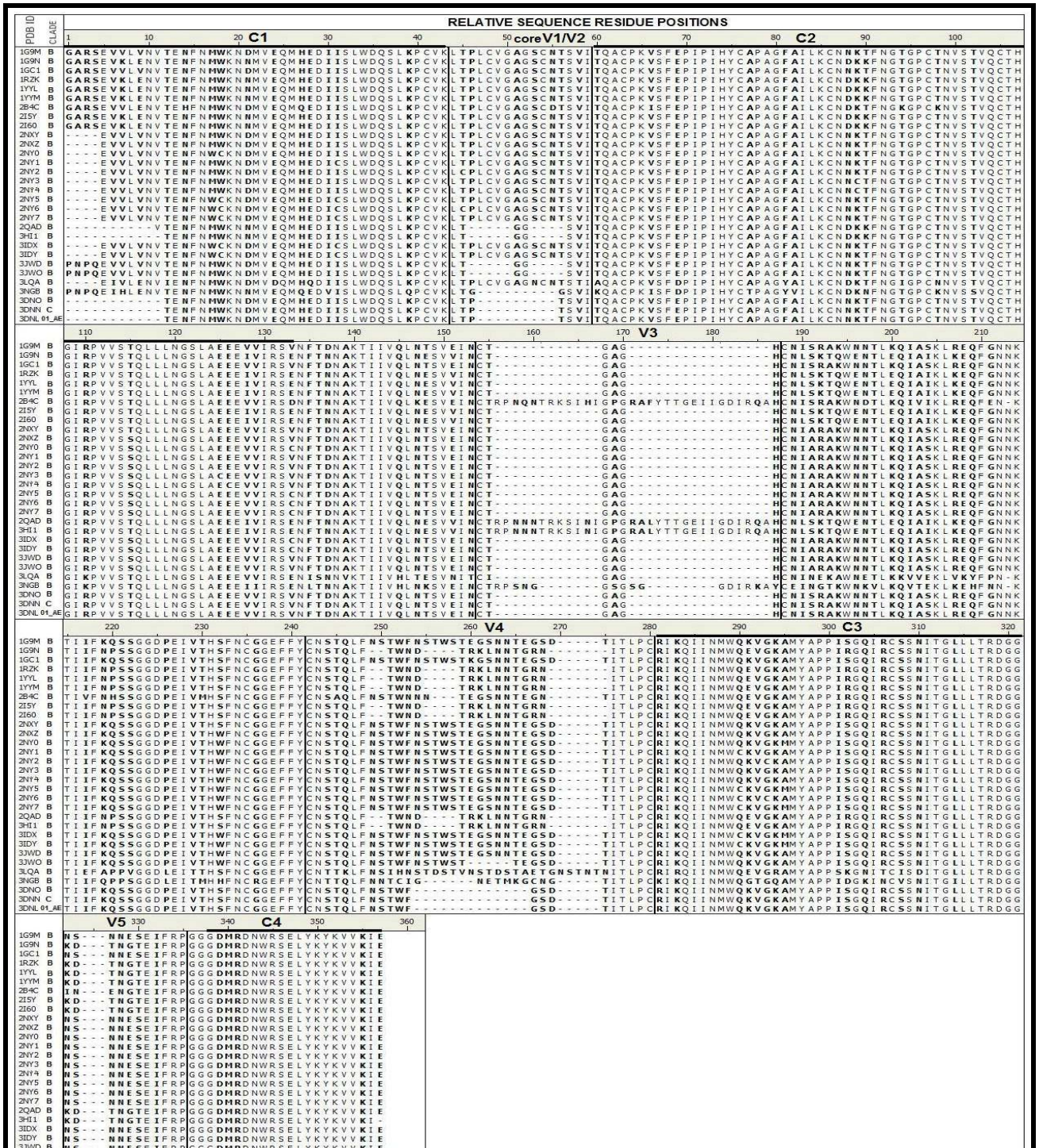

Figure 1: MSA of gp120 sequence dataset (**Table 1**) derived from known structures from the PDB database. [http://www.pdb.org/pdb/home] The position specific mutations are in bold letters. The incomplete sequences in the alignment are indicated with dashes. The V1/V2, V3, V4 and V5 regions are marked. However, all the variable loops are not yet solved by X-ray crystallography in gp120 structures and hence their corresponding atomic coordinates are not available **[40]**. These regions are indicated with incomplete sequences.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 6(2): 48-56 (2011)

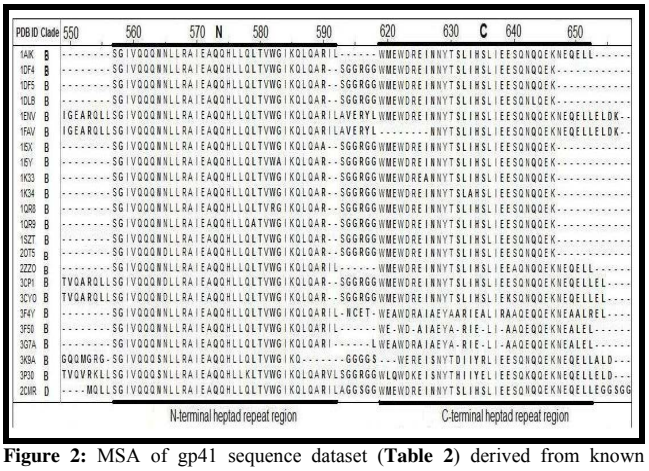

Figure 2: MSA of gp41 sequence dataset (**Table 2**) derived from known structures from the PDB database. [http://www.pdb.org/pdb/home]. The position specific mutations are in bold letters. The incomplete sequences in the alignment are indicated with dashes. The gp41 structures shown consist of the N and C helices representing the N heptad repeat and C heptad repeat, respectively.

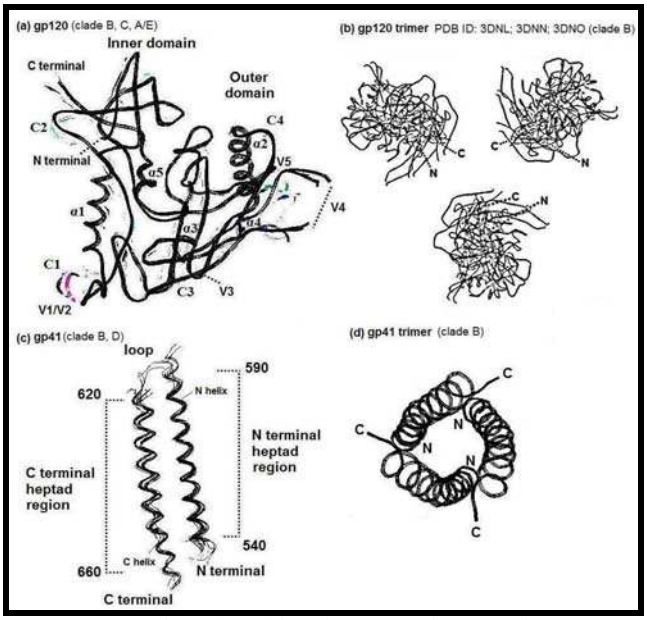

Figure 3: Structural superimposition of (a) gp120, (b) gp120 trimer, (c) gp41, and (d) gp41 trimer. The gp120 and gp41 structures obtained from PDB [http://www.pdb.org/pdb/home] (**Table 1 and Table 2**) were superimposed using the software SPDBV (Swiss PDB Viewer version 3.7). [http://www.expasy.org/spdbv/]. The variable loops (V1-V5) and the constant regions (C1-C4) of the GP120 structure, with inner and outer domains are shown (a). The superimposition of the gp120 trimer (PDB ID: 3DNL), solved through NMR at 20.0Å is shown in (b). The N heptad and the C heptad regions of the superimposed gp41 structure, solved by X-ray crystallography with a resolution of 2.10Å, are shown in (c). Only the N heptad region of gp41 is available in trimeric form, as illustrated in (d).

Results:

The Graphical Abstract outlines the flow of analysis from the databases to the tables and figures. The LANL and PDB databases were utilized to obtain sequence and structure information for the HIV-1 gp120 and gp41. We generated datasets from PDB of known structures produced by X-ray crystallography for gp120 and gp41 shown in **Tables 1 and 2**, respectively. These datasets include clades B, C, and A/E for gp120, and clades B and D for gp41. From **Tables 1 and 2**, we produced MSAs for gp120 and gp41 shown in **Figures 1 and 2**, respectively. In **Figure 3**, the corresponding structures for the sequence alignments were then used for constructing structural superimposition of gp120 (a), its trimer (b), gp41 (c), and its trimer (d). The superimposition of multiple structures shows that several clades share the same structural folds for gp120 (clades B, C, A/E) and gp41 (clades B, D). The ASA distribution, mean, and standard deviations are shown for each residue position for gp120 and gp41 in **Figure 4** based on **Tables 1 and 2**. These values show the degree of structural variation for the differences in sequence, which is represented by the ASA measure and help identify residue positions that are solvent exposed (ASA > 0 $Å^2$). The mean distribution show that most residues in gp120 and gp41 are solvent exposed.

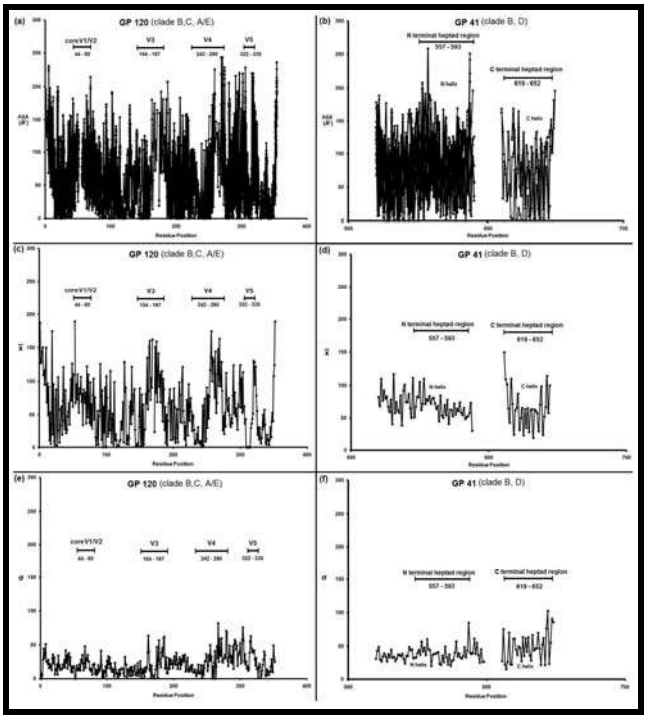

Figure 4: Residue position specific ASA distribution profiles for (a) gp120 and (b) gp41, (c) and (d) respective means, (e) and (f) respective standard deviations. The structural datasets were obtained from PDB [http://www.pdb.org/pdb/home] as shown in **Tables 1 and 2**. The ASA distribution was calculated using Surface racer **[41]**.

As of 12-31-2010, there were approximately 14,925 gp120 and 14,472 gp41 sequences in the LANL database for clades A-K. **Figure 5** shows the sequence percent polarity for several clades, blood, and brain by geographical location from this database. In addition, the sequence percent polarity range is shown for gp120 (clades B, C, A/E) and gp41 (clades B, D) for which structures are known from the PDB database. (The figures that utilize sequences from the full LANL database for clades A-K and the sequences from known structures from PDB do not specify geographical location.) The ranges in percent compositional polarity among clades in blood and brain sequences are shown in **Table 3 (see Supplementary material)**. The polarity range of gp120 clade B brain vs. blood sequences is 19.47% vs. 18.0%, respectively in Europe, clade B brain vs. blood, 20.13% vs. 21.64%, respectively, in North America, and clade A brain vs. blood 17.78% vs. 19.44%, respectively, in Africa. The range of polarity of gp41 clade B brain vs. blood sequences is 11.81% vs. 10.27%, respectively in Europe, and clade B brain vs. blood, is 12.76% vs. 10.72%, respectively, in North America. There is a correlation for gp120 Shannon entropy vs. gp120 polarity range across clades and for gp41 Shannon entropy vs. gp41 polarity range among clades. The Pearson correlation coefficients (r) are 0.734 and 0.588, respectively, for gp120 and gp41 (**Figure 6**). Two-Way ANOVA of Shannon entropy and polarity of gp120 as well as gp41 sequences shows significant variation in both blood ($F = 509.6$; $P<0.0001$) and brain ($F = 790.9$; $P < 0.0001$). Furthermore, the post-test for linear trend, between the Shannon entropy and polarity, is positively correlated by coefficient of determination in blood ($R^2 = 0.558$) and in brain ($R^2 = 0.483$).

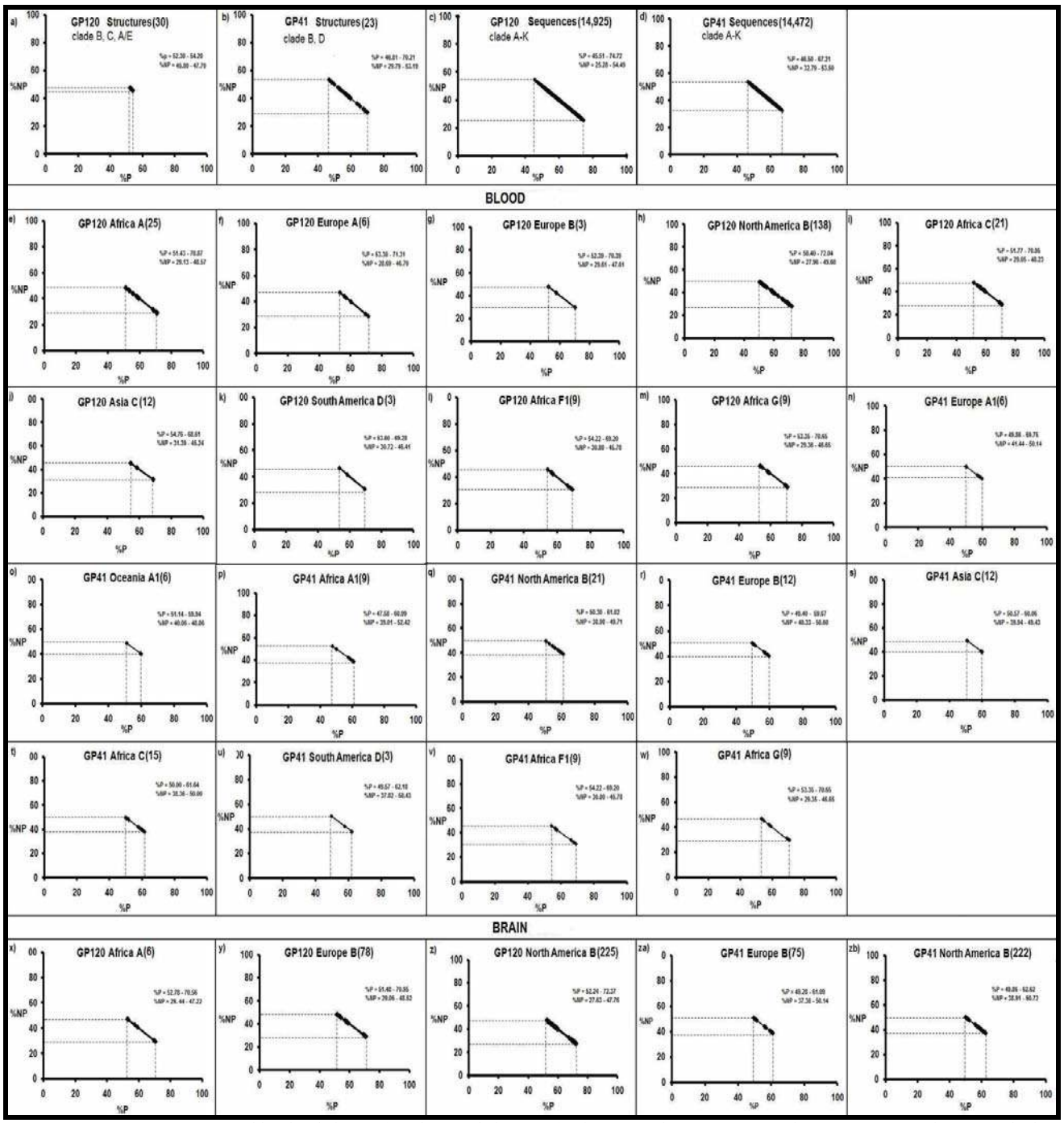

Figure 5: Percent sequence compositional polarity (%P) and non-polarity (%NP) of gp120 and gp41 sequences.These sequences are from HIV database sequences, the PDB structural database (a, b) [http://www.pdb.org/pdb/home] and the LANL sequence database (c, d) [http://www.hiv.lanl.gov/]. (Geographical information is not available from PDB in (a) and (b) and the A-K sequence geographical information in (c) and (d) from LANL are not shown). Additional sequence compositional polarity comparisons are from blood (e to w) and brain (x to zb). Clade and source geographical locations are stated (from LANL). The numbers in parenthesis represent the number of sequences. The ranges for %P and %NP are stated as well in each figure.

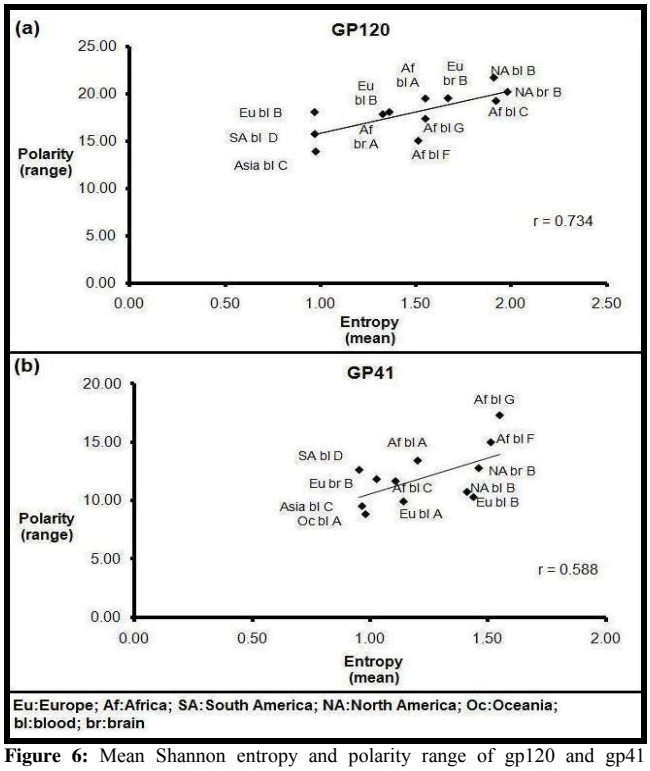

Figure 6: Mean Shannon entropy and polarity range of gp120 and gp41 sequences in brain and blood samples from different geographical locations. Shannon entropy of gp120 and gp41 sequences is calculated at the LANL website [http://www.hiv.lanl.gov/]. The mean Shannon entropy is correlated with polarity range by regression coefficient (r) of 0.734 for gp120 (a) and 0.588 for gp41 (b) sequences. The Pearson correlation co-efficient was calculated using Microsoft EXCEL software.

Discussion:

The gp120-gp41 complex trimer protein is a potential HIV-1 vaccine candidate **[42]**. The gp120 and gp41 protein subunits interact with each other at an interface, forming a gp120-gp41 complex involved in trimer assembly. The gp41 interactive region of the gp120 protein has a layered structure that has conformational mobility – flexibility - at the interface with gp41 **[43]**. The gp120 structures in the PDB database are available only in ligand-bound states, which may be partly due to the limited stability of the protein without support ligands. The trimer is unstable when produced in vitro and this may be caused by its sequence composition and conformation **[44-46]**. The polar composition of the gp120-gp41 complex trimer protein influences its surface, immunological, and stability properties. The bottleneck for the in vitro synthesis and production of this protein in stable form may be the prevalence of solvent exposed polar residues as described in our findings. Moreover, trimer instability is probably also due to the difficulty in exactly mimicking the in vivo environment, in vitro, for protein folding and assembly of the complex. The structure of the gp120-gp41 complex is similar among clades and homologous sequences share common structural folds and shapes although they have differences among side chain packing and residue orientation **[47, 48]**. Our analysis, using a solvent exposure measure, ASA, shows three-dimensional structural variation for each AA position and that the residues are solvent-exposed. In addition, in major segments of gp120 and gp41, we demonstrate a large percent polarity for solvent exposed residues across clades, blood, and brain. Thus, our findings support more open and dynamic ENV structures and conformations.

The variation in biochemical properties is relevant to manufacturing a stable and effective HIV-1 ENV protein. An additional concern in using the ENV as an HIV/AIDS vaccine candidate is its high sequence variation among clades from different geographical locations; the LANL database (as of 12-31-2010) contained 14,925 gp120 and 14,472 gp41 sequences. Potential gp120-gp41 global vaccine candidates should incorporate the issue of immunological specificity and AA mutant variation across clades from different geographical locations. Our analysis of the known sequences for gp120 and gp41 to estimate the polarity changes caused by AA variation show that the percent polarity range among clades, blood, and brain correlates with the mean Shannon entropy. This reflects sequence variation that changes surface properties within and across clades, blood, and brain and this is anticipated to affect their respective immunological responses. Despite the structural similarities determined so far across clades for blood, this is not yet known for blood vs. brain. The difference in sequence polarity range for brain is comparable to blood for gp120 and is greater for brain than blood for gp41. These sequence variations could be due to differences in immune selection between brain and blood as well as due to the structure of the ENV, e.g. the gp120 juts further into the solvent than does gp41, since gp41 is partially submerged in the membrane. These findings help quantify the percent compositional polarity range within which the gp120 and gp41 sequences vary among clades; we infer from this structural folding related to conformations pertinent to the immune response. This is further relevant in the design of suitable HIV-1 ENV vaccines specific for multiple HIV-1 clades across blood- and brain-derived sequences. Vaccines that induce neutralizing antibody are currently insufficient to the task; thus, new methodologies are needed to optimize this approach. A recent method is under development that originates from a neutralizing antibody, works its way back to reconstruct the epitopes (reverse engineering), and then uses a structure based design technology to optimize the epitopes **[49]**. The structural information presented in this article should enhance methods that augment the stability of the gp120-gp41 complex and trimer. This work points to the need for the development of supporting ligands that assist in protein conformation stabilization as well as producing AA mutants and other protein modifications that neutralize antigen charge where needed. In addition, brain-related HIV-1 should be dealt with as strains of HIV-1 that similarly require a specific antigenic enhancement approach. For this tactic, it is important to understand the charge characteristics of the antigen; thus, we characterize the HIV-1 ENV from the point of view of compositional percent charge polarity and its variation across clades, brain, and blood.

There are several additional concerns in developing HIV-1 ENV vaccines that indicate the complexity of the problem and the sophistication required. These include issues of antigen integrity, delivery, and cross-reactivity across clade, blood, and brain viral strains. Furthermore, potential vaccines may have advantageous or deleterious effects depending on other factors as well. Such vaccine-related factors include host immune responses that might cause inflammation resulting from vaccines that could be more deleterious in the presence of HIV-1 infection. Thus, application of the knowledge of antigen structure may be different for vaccines in uninfected vs. HIV-1 infected individuals. Virologic and immunologic factors (e.g., pre-existing viral strains and cognate host immune responses), once one introduces a vaccine into the CNS of an infected individual, might result in further inflammation as well. Moreover, it is unknown how selective pressure on virus evolution might lead to vaccine-resistant strains that vitiate vaccine effects. Would a vaccine be possible that stimulates the immune response sufficiently rapidly to halt virus replication prior to mutant virus spread? It is also unknown about the potential interactions with vaccines amongst neurovirulent macrophage-tropic HIV-1 strains (that predominantly infect the CNS). In addition, which potential vaccines will inhibit or prevent their neurovirulence as well as suppress or prevent brain infections due to CNS strains? In host genetics for example, the CCR5-32-delta polymorphism results in a less or non-functional HIV-1 co-receptor. Therefore, those individuals with that allele would be anticipated to have improved potential survival with a vaccine. However, NeuroAIDS susceptibility allele SNPs may play a deleterious role during vaccine exposure coupled with virus exposure, pre- or post-vaccinations. Because NeuroAIDS appears to result largely from host immune state and viral strain, host variability in the immune response to vaccine introduction is a crucial concomitant concern **[10, 27, 29, 30, 31, 50, 51]**.

The effects of HIV-1 infection can be directly measured in the brain and the neuropathology of HIVE has changed in the brain since the use of HAART commenced in 1995 **[52]**. Concomitant use of vaccines and anti-viral therapy may have an interactive impact on these therapies systemically and in the brain. The production of escape mutants could occur within brain that has been demonstrated systemically by automated deep sequencing techniques because of anti-CCR5 strain antiviral therapy **[53]**. During the last several years, the HIV-1 infected population has shown increased aging and increased incidence of Alzheimer's disease and these factors may further complicate vaccine use in at-risk and already infected individuals **[52]**. An additional complicating factor in the use of vaccines is the possibility of autoimmunity as a component of increased inflammation that could occur peripherally as well as within the brain **[54]**. Finally, in this article, we do not deal with the problem of protein

glycosylation; however, this is an important issue in vaccine production. On the one hand, the virus, as an additional means of immune escape may use glycosylation, though contributing to immunogenicity. On the other hand, deglycosylation of HIV-1 ENV proteins, in vitro, may enhance their use in vaccines **[55-58]**.

Acknowledgments:
We thank the members of Biomedical Informatics (Pondicherry, India) for extensive discussions.

References:

[1] Alter G *et al. Hum Vaccine*. 2008 **5**:119
[2] Bass E *et al. BETA*. 2009 **21**: 24 [PMID: 19517626]
[3] Rerks-Ngarm S *et al. N Engl J Med.* 2009 **361**: 2209 [PMID: 19843557]
[4] Wayne CK & Berkley SF. *N Engl J Med*. 2010 **363**:e7 [PMID: 20830827]
[5] Zhou T *et al. Science* 2010 **329**: 811[PMID: 20616231]
[6] Zolla-Pazner S & Cardozo T. *Nat Rev Immunol.* 2010 **10**:527 [PMID: 20577269]
[7] Wu X *et al. Science* 2010 **329**: 856 [PMID: 20616233]
[8] http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/Sodroski.html
[9] Kwong PD *et al. Nature* 1998 **393**: 648 [PMID: 9641677]
[10] Shapshak P *et al. AIDS* 2011 **25**: 123 [PMID: 21076277]
[11] Liang B *et al. Biochem Cell Biol.* 2007 **85**: 259 [PMID: 17534408]
[12] Kumar S *et al. J Cell Tissue Res.* 2010 **10**: 2359
[13] Gabuzda D *et al. J Acquir Immune Defic Syndr.* 1991 **4**: 34 [PMID: 1984054]
[14] Jiang X *et al. Nat Struct Mol Biol.* 2010 **17**: 955 [PMID: 20622876]
[15] Dong XN *et al. Immunol Lett.* 2001 **75**: 215 [PMID: 11166378]
[16] Lorizate M *et al. Biochemistry* 2006 **45**: 14337 [PMID: 17128972]
[17] Hollier MJ & Dimmock NJ. *Virology* 2005 **337**: 284 [PMID: 15913700]
[18] Fontaine Costa AL *et al. AIDS* 2010 **24**: 211[PMID: 19904197]
[19] Yang OO. *PLoS ONE*. 2009 **4**: e7388 [PMID: 19812689]
[20] Fischer W *et al. PLoS One.* 2010 **5**: e12303 [PMID: 20808830]
[21] Gnanakaran S *et al. J Virol.* 2007 **81**: 4886 [PMID: 17166900]
[22] Yusim K *et al. J Virol.* 2002 **76**: 8757 [PMID: 12163596]
[23] Bai X *et al. Biochemistry* 2008 **47**: 6662 [PMID: 18507398]
[24] Tan K *et al. Proc Natl Acad Sci U S A.* 1997 **94**: 12303 [PMID: 9356444]
[25] Woo J *et al. J Virol.* 2010 **84**: 12995 [PMID: 20881050]
[26] Korber BT *et al. J Virol.* 1994 **68**: 7467 [PMID: 7933130]
[27] Kangueane P *et al.* The Spectrum of NeuroAIDS Disorders: Pathophysiology, Diagnosis, and Treatment (Goodkin K, Verma A, & Shapshak P, eds), ASM Press, Washington, 2008, pg 105
[28] Chang J *et al. AIDS Res Hum Retroviruses*. 1998 **14**: 25 [PMID: 9453248]
[29] Shapshak P *et al. AIDS Res Hum Retroviruses*. 1999 **15**: 811 [PMID: 10381169]
[30] Shapshak P *et al.* The Spectrum of NeuroAIDS Disorders: Pathophysiology, Diagnosis, and Treatment (Goodkin K, Verma A, & Shapshak P, eds) ASM Press, Washington, 2008
[31] Shah M *et al. AIDS Res Hum Retroviruses.* 2006 **22**: 177 [PMID: 16478400]
[32] Zarate S *et al. J Virol*. 2007 **61**: 6643 [PMID: 17428864]
[33] Kuiken CL *et al.* HIV Sequence Compendium 2010. Los Alamos National Laboratory, 2010.
[34] Larkin MA *et al. Bioinformatics* 2007 **23**: 2947 [PMID: 17846036]
[35] Lee B & Richards FM. *J Mol Biol*. 1971 **55**: 379 [PMID: 5551392]
[36] Tsodikov OV *et al*. *J Comput Chem*. 2002 **23**: 600 [PMID: 11939594]
[37] Korber BT *et al. J Virol.* 1994 **68**: 6730 [PMID: 8084005]
[38] Reza FM. An introduction to Information Theory. Dover Publ NY. 1994: ISBN 0-486-68210-2.
[39] Calefa C *et al.* HIV Molecular Immunology. (Korber BTM, Brander C,Haynes BF, Koup R, Moore JP, Walker BD & Watkins DI, eds). Los Alamos, New Mexico, 2005: pg 33.
[40] Wyatt R *et al.* Human Retroviruses and AIDS 1998: (Korber B, Kuiken CL, Foley B, Hahn B, McCutchan F, Mellors JW & Sodroski J, eds), Los Alamos National Laboratory, 1998, pg 3.
[41] Liu J *et al. Nature* 2008 **455**: 109 [PMID: 18668044]
[42] Burton DR *et al. Nat Immunol*. 2004 **5**: 233 [PMID: 14985706]
[43] Pancera M *et al. Proc Natl Acad Sci U S A*. 2010 **107**: 1166 [PMID: 20080564]
[44] Moore JP *et al. J Virol*. 1992 **66**: 235 [PMID: 1727487]
[45] Moore JP *et al. AIDS Res Hum Retroviruses*. 1990 **6**: 1273 [PMID: 2078408]
[46] Chen B *et al. Nature* 2005 **433**: 834 [PMID: 15729334]
[47] Abagyan RA & Batalov S. *J Mol Biol*. 1997 **273**: 355 [PMID: 9367768]
[48] Kinjo AR *et al. Eur Biophys J*. 2001 **30**: 1 [PMID: 11372527]
[49] Sattentau QJ & McMichael AJ. *F1000 Biol Rep*. 2010 **2**: 60 [PMID: 21173880]
[50] Levine AJ *et al. AIDS Behav*. 2009 **13**: 118 [PMID: 18264751]
[51] Levine AJ *et al. Neurobehav HIV Med*. 2009 1: 1 [PMID: 20725607]
[52] Everall IP *et al. Neurotox Res*. 2005 **8**: 51[PMID: 16260385]
[53] Tsibris AM *et al. PLoS One*. 2009 **4**: e5683 [PMID: 19479085]
[54] Zandman-Goddard G & Shoenfeld Y. *Autoimmun Rev*. 2002 **1**: 329 [PMID: 12848988]
[55] Go EP *et al. J Proteome Res*. 2009 **8**: 4231 [PMID: 19610667]
[56] Huang Z *et al. Curr HIV Res*. 2008 **6**: 296 [PMID: 18691028]
[57] Kong L *et al. J Mol Biol*. 2010 **403**: 131 [PMID: 20800070]
[58] Zhang C *et al. AIDS Res Hum Retroviruses*. **26**: 569 [PMID: 20455767]
[59] Diskin R *et al. Nat Struct Mol Biol*. 2010 **17**: 608 [PMID: 20357769]
[60] Zhou T *et al. Nature* 2007 **445**: 732 [PMID: 17301785]
[61] Chen L *et al. Science* 2009 **326**: 1123 [PMID: 19965434]
[62] Stricher F *et al. J Mol Biol*. 2008 **382**: 510 [PMID: 18619974]
[63] Huang CC *et al. Science* 2007 **317**: 1930 [PMID: 17901336]
[64] Huang CC *et al. Structure* 2005 **13**: 755 [PMID: 15893666]
[65] Huang CC *et al. Proc Natl Acad Sci U S A*. 2004 **101**: 2706 [PMID: 14981267]
[66] Kwong PD *et al. Structure* 2000 **8**: 1329 [PMID: 11188697]
[67] Chan DC *et al. Cell* 1997 **89**: 263 [PMID: 9108481]
[68] Shu W *et al. J Biol Chem*. 2000 **275**: 1839 [PMID: 10636883]
[69] Shu W *et al. Biochemistry* 2000 **39**: 1634 [PMID: 10677212]
[70] Weissenhorn W *et al. Nature* 1997 **387**: 426 [PMID: 9163431]
[71] Zhou G *et al. Bioorg Med Chem*. 2000 **8**: 2219 [PMID: 11026535]
[72] Lu M *et al. J Virol*. 2001 **75**: 11146 [PMID: 11602754]
[73] Wang S *et al. Biochemistry* 2002 **41**: 7283 [PMID: 12044159]
[74] Ji H *et al. J Virol*. 1999 **73**: 8578 [PMID: 10482611]
[75] Luftig MA *et al. Nat Struct Mol Biol*. 2006 **13**: 740 [PMID: 16862157]
[76] Watabe T *et al. J Mol Biol*. 2009 **392**: 657 [PMID: 19616557]
[77] Horne WS *et al. Proc Natl Acad Sci U S A*. 2009 **106**: 14751 [PMID: 19706443]
[78] Shi W *et al. J Biol Chem*. 2010 **285**: 24290 [PMID: 20525690]
[79] Frey G *et al. Nat Struct Mol Biol*. 2010 **17**: 1486 [PMID: 21076402

Edited by F Chiappelli
Citation: Sowmya *et al.* Bioinformation 6(2): 48-56 (2011)

BIOINFORMATION open access

Supplementary material:

Table 1: GP120[a] structural dataset from PDB [http://www.pdb.org/pdb/home]

PDB ID	Clade	Length (aa)	Res[b] (Å)	Mutation	Protein Region	Reference
3JWD	B	379	2.61	-[c]	Core[d]	[43]
3JWO	B	357	3.51	T257S, S375W	Core	[43]
3LQA	C	332	3.4	T89I, N226D, T232I, N285T, S329N, T388I, N447D	Core	[59]
3NGB	01_AE	353	2.68	-	Core	[60]
3HI1	B	321	2.9	-	Core	[61]
3IDX	B	317	2.5	M95W, T257S, S375W, A443M, W96C, V275C, I109C, Q428C	Core	[61]
3IDY	B	317	3.2	M95W, T257S, S375W, A443M, W96C, V275C, I109C, Q428C	Core	[61]
3DNL	B	35			Core	[41]
		170	20	-		
		83				
3DNN	B	35			Core	[41]
		170	20	-		
		83				
3DNO	B	35			Core	[41]
		170	20	-		
		83				
2I5Y	B	313	2.2	-	Core	[62]
2I60	B	313	2.4	-	Core	[62]
2NXY	B	317	2	S334A	Core	[60]
2NXZ	B	317	2.04	T257S, S334A, S375W	Core	[60]
2NY0	B	317	2.2	M95W, W96C, T257S, V275C, S334A, S375W, A433M	Core	[60]
2NY1	B	317	1.99	I109C, T257S, S334A, S375W, Q428C	Core	[60]
2NY2	B	317	2	T123C, T257S, S334A, S375W, G431C	Core	[60]
2NY3	B	317	2	K231C, T257S, E267C, S334A, S375W	Core	[60]
2NY4	B	317	2	K231C, T257S, E268C, S334A, S375W	Core	[60]
2NY5	B	317	2.5	M95W, W96C, I109C, T257S, V275C, S334A, S375W, Q428C, A433M	Core	[60]
2NY6	B	317	2.8	M95W, W96C, I109C, T123C, T257S, V275C,S334A, S375W, Q428C, G431C	Core	[60]
2NY7	B	317	2.3	M95W, W96C, I109C, T257S, V275C, S334A, S375W, Q428C, A433M	Core	[60]
2QAD	B	322	3.3	-	Core with V3	[63]
1YYL	B	313	2.75	-	Core	[64]
1YYM	B	313	2.2	-	Core	[64]
2B4C	B	344	3.3	-	Core with V3	[64]
1RZK	B	313	2.9	Variable Loops substituted	Core	[65]
1G9M	B	321	2.2	-	Core	[66]
1G9N	B	313	2.9	-	Core	[66]
1GC1	B	321	2.5	(GARS)[e] Substitution at N Terminus, Gly/Ala/Gly substitution for V1/V2 and V3 Loops	Core	[9]

[a]GP120 = Larger subunit of ENV glycoprotein; [b]Res = Atomic resolution of PDB structure; [c]no information available; [d]Core region of gp120; [e]GARS = Glycine, Alanine, Arginine, Serine.

Table 2: GP41 structural dataset from PDB [http://www.pdb.org/pdb/home]

PDB ID	Clade	Length (aa)	Res [a](Å)	Mutation	Protein region	Reference
1AIK	B	[b]N=37;[c]C=35	2.0	-[d]	Core[e]	[67]
1DF4	B	68	1.45	-	Core	[68]
1DF5	B	68	2.7	-	Core	[68]
1DLB	B	68	2.0	Q65L	Core	[69]
1ENV	B	123	2.6	V61, L9I, N13I, L16I, V20I, L23I, V27I	Ectodomain[f]	[70]
1FAV	B	79	3	V6I, L9I, N13I, L16I, V20I, L23I, V27I	Trimeric core	[71]
1I5X	B	68	1.8	R579A	Core	[72]
1I5Y	B	68	2.1	G572A	Core	[72]
1K33	B	68	1.75	I48A	Core	[73]
1K34	B	68	1.88	I55A	Core	[73]
1QR8	B	68	2.1	W571R	Core	[74]
1QR9	B	68	1.6	-	Core	[74]
1SZT	B	68	2.4	-	Thermostable subdomain	[70]
2CMR	D	226	2	-	Inner core mimetic 5-helix	[75]
2OT5	B	68	1.8	N43D	Core	[23]
2ZZO	B	N=37;C=35	2.2	S138A in C	Fragment N36 and fusion inhibitor C34/S138A	[76]

3CP1	B	86	2	N43D	Core domain	**[23]**
3CYO	B	86	2.1	N554D,E648K	Core domain	**[23]**
3F4Y	B	N=38; C=40	2	M1T, M4E, E5A, E9A, N11A, N12E, T14A, S15A, L16R, H18E, S19A, N21E, Q33A, E34A, L36R in C	Six-helix bundle	**[77]**
3F50	B	38	2.8	-	Six-helix bundle	**[77]**
3G7A	B	36	2.8	-	Six-helix bundle	**[77]**
3K9A	B	108	2.1	HR1+4XGly+HR2+MPER[g]	Ectodomain	**[78]**
3P30	B	96	3.3	-	Ectodomain	**[79]**

[a]Res= Atomic resolution of PDB structure; [b]N= N terminal of gp41; [c]C= C terminal of gp41; [d]no information available; [e]Core = Core structure of gp41; [f]Ectodomain of gp41; [g]MPER= Membrane-proximal external region

Table 3: The range of compositional polarity for ENV, GP120 and GP41 structures from PDB [http://www.pdb.org/pdb/home] and sequences from LANL [http://www.hiv.lanl.gov/]

		Region[a]	**Clade**	**Number of sequences**	**Range[b] of % compositional polarity**
GP120	Sequences	–[c]	A-K	14,925	29.21
	Structures[d]	–	B,C,A/E	30	1.90
	Blood Sequences	Africa	A	25	19.44
		Africa	C	21	19.18
		Africa	F1	9	14.98
		Africa	G	9	17.30
		Asia	C	12	13.85
		South America	D	3	15.68
		North America	B	138	21.64
		Europe	A	6	18.01
		Europe	B	3	18.00
	Brain Sequences	Africa	A	6	17.78
		Europe	B	78	19.47
		North America	B	255	20.13
GP41	Sequences[e]	–	A-K	14,472	20.71
	Structures[f]	–	B,D	23	23.40
	Blood Sequences	Africa	A1	9	13.41
		Africa	C	15	11.64
		Africa	F1	9	14.98
		Africa	G	9	17.3
		Oceania	A1	6	8.8
		Asia	C	12	9.49
		Europe	A1	6	9.89
		Europe	B	12	10.27
		North America	B	21	10.72
		South America	D	3	12.61
	Brain Sequences	Europe	B	75	11.81
		North America	B	222	12.76

[a]Region = sample source geographical location; [b]Range = Maximum % compositional polarity - Minimum % compositional polarity; [c]No geographical locations specified; [d]gp120 sequences based on structures from PDB; [e]gp41 sequences from LANL; [f]gp41 sequences based on structures from PDB