ALIGNMENT OF HIV-1/SIVCPZ GENOMES

This year many new full-length viral sequences have become available, originating from diverse geographic origins and representing the spectrum of known HIV variation. We have decided to publish only full length HIV-1/CPZ sequences in our printed nucleotide alignment section, as this set is now becoming an adequate representation of the overall diversity of the virus.

As of December 1999 there were 161 complete or nearly complete (defined as greater than 8,000 consecutive basepairs of sequence) HIV-1 genomes in the database. Of these, some were not included in the printed alignment, as they are very closely related to a sequence already included in the alignment, and our intent is to print a hardcopy alignment representative of global diversity. The complete alignment including all sequences is available at our web and ftp sites.

`http://hiv-web.lanl.gov/ALIGN_CURRENT/ALIGN-INDEX.html`

Ninety-nine HIV-1 sequences plus viral strains isolated from chimpanzees, CPZANT, CPZGAB, and CPZUS comprise the printed alignment. In phylogenetic analyses, the CPZ sequences are the simian-derived viruses most similar to HIV-1; in fact HIV-1 M, N and O group sequences are roughly as distant from one another as they are from the CPZ sequences (see Figure 1 below, and caption next page).

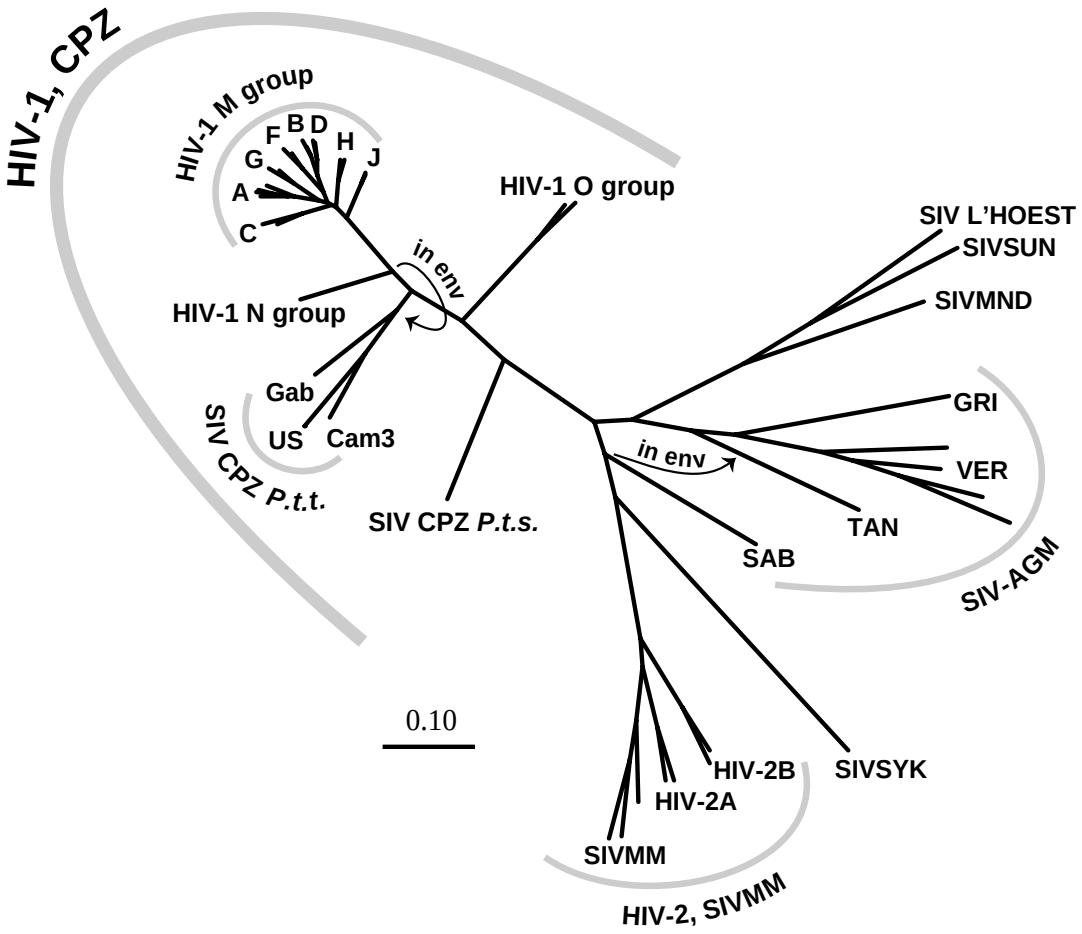

Figure 1: Primate lentivirus phylogenetic relationships.

The *pol* gene was used to illustrate primate lentivirus relationships as it is relatively conserved and so particularly useful for reconstructions of very distant sequence relationships. Two of the viruses depicted in this tree (HIV-1 N group, and SIVagm SAB, from the African green monkey *sabaeus* subspecies), have mosaic genomes. This means that the phylogenetic relationships and clustering depends upon the region of the genome under consideration, indicating recombination. The HIV-1 N group sequence is more closely related to the HIV-1 M group in the *pol* gene region, as shown; in *env* gene phylogenies, the N group clusters within the SIVcpz clade from the chimpanzee subspecies *Pan troglodytes troglodytes (P.t.t.)* (F. Gao et al., *Nature* **397**:436 (1999), S. Corbet et al., *J Virol* **74**:529 (2000)). The SIVagm SAB viruses are recombinant between a virus that infects other species of African green monkeys (Tantalus, Vervet and Grivet) and a virus which infects Sooty mangabeys, SIVmm. There is a recombination breakpoint within the *pol* gene of SIVagm SAB, which is why the branch occupies an intermediate position between the SIVmm and SIVagm; this position does not reflect a true evolutionary history. Because of the inclusion of the recombinant, this is not an accurate phylogeny, however it serves to give a reasonable portrait of the relationships between primate lentivirus; the small arrows indicate where the sequences would branch in an *env* gene reconstruction.

Viruses depicted in this tree:

Human:

HIV-1 M (Main) group, including reference strains from subtypes A-J. Group M is responsible for the pandemic

HIV-1 O (Outlier) group, most commonly found in West Africa

HIV-1 N (Not-M, Not-O) group, found in a very small number of individuals in West Africa

HIV-1 M group reference strains: A_UG.U455, A_KE.Q2317, B_US.JRFL, B_US.WEAU160, C_ET.ETH2220, C_IN.21068, D_ZR.NDK, D_ZR.ELI, F_FI.FIN6393, F_BE.VI850, G_SE.SE6165, G_BE.DRCBL, H_CF.90CF056, H_BE.VI997, J_SE.SE91733, J_SE.SE92809, and CRF01 AE_CF.90CF402 and AE_TH.CM240, which are subtype A in *pol*.

HIV-1 N group: N_CM.YBF30

HIV-1 O group: O_CM.ANT70, O_CM.MVP5180

HIV-2 subtypes A and B: H2A_DE.BEN, H2A_SN.ST, H2B_GH.D205, and H2B_CI.EHO

Simian:

SIVcpz from chimpanzee Pan troglodytes troglodytes (P.t.t.):
SIVcpz.GAB, SIVcpz.US, and SIVcpz.Cam3

SIVcpz from chimpanzee Pan troglodytes shweinfuthii (P.t.s.): SIVcpz.ANT

SIV African Green Monkey (SIVagm):
Tantalus (TAN): SIVagm.TAN1
Vervet (VER): SIVagm.VERTYO, SIVagm.VERAGM3, SIVagm.VER9063, SIVagm.VER155
Grivet (GRI): SIVagm.GRI677
Sabaeus (SAB): SIVagm.SAB1C

SIV Sooty Mangaby (SIVsm) (also found in macaques): SIVsm.mac251, SIVsm.smm9

SIV L'hoest: SIV.LHOEST

SIV Mandrill: SIV.MNDGB1

SIV Sun: SIV.SUN

The primary references for these viral sequences can be found through our website, the common names are noted after the period.

All 161 of the complete genomes for HIV-1 have been updated with annotation of the major gene start and end sites. All are available as fully annotated database entries, with subtype and country of origin included, from the HIV database WWW site

`http://hiv-web.lanl.gov/`

by using the sequence search interface

`http://hiv-web.lanl.gov/cgi-bin/hivDB3/public/wdb/ssampublic`

to search for HIV-1 sequences with length greater than 8,000 bases (this will select for only the full length or near complete genome sequences when using in the search tool).

The sequences in this section are identified by their common name preceded by the HIV subtype designations and country of origin appropriate for the sequence. The primary sequence reference, country of origin, database accession number, and brief notes describing the isolate and sequence, with some additional relevant references, can be found in Table 2 for the set of sequences included in the printed alignment. The sequences that have been found to be recombinants with portions of the genetic sequences associated with different subtypes are indicated by listing all of the subtypes in the prefix to the name. For example, the prefix AG simply indicates that some regions of the sequence are subtype A-like, others G-like. The subtypes are organized alphabetically and not meant to reflect the proportion of either subtype in the mosaic genome. In the HIV-2/SIV section two new whole genomes were published this year: SIV-sun (*J Virol* **73**(9):7734–7744 (1999)), a relative of SIV-L'Hoest (*J Virol* **73**(2):1036–1045 (1999)), and HIV-2 ALI (Unpublished (1998)), a new HIV-2 subtype A strain. Amino acid sequences of all genes for these strains have been incorporated in the gene alignments. In addition, a few dozen sequences were added to the amino acid sequences of various genes. In the amino acid alignments, we have not adhered strictly to the requirement that all sequence encompass the full-length gene, as for some HIV-1 and SIV subtypes this would imply that they drop out of the printed alignment altogether.

Alignment This alignment was generated by using the HMMER Hidden Markov Model sequence alignment software developed by Sean Eddy.

`http://genome.wustl.edu/eddy/hmmer.html`

An iterative process was used involving alignment of the genomes using HMMER, followed by hand-editing (using an in-house revised version of the MASE alignment editing program (Faulkner, D., and Jurka, J., *Trends in Biochem. Science*, **13**:321–322 (1988)), and BioEdit.

`http://www.mbio.ncsu.edu/RNaseP/info/programs/BIOEDIT/bio-edit.html`

The resulting final alignment is not suggested to be an "optimal alignment" with the absolute minimum number of gaps and mismatches. It is a compromise between optimal alignment, readability, and an attempt to keep insertions and deletions from altering the protein reading frame presentation. Most gaps have been introduced in multiples of 3 bases to maintain open reading frames when translated directly from the alignment. Frameshifting gaps were added at the *gag-pol* slip site, at the end of *pol*, and at the end of *vif*.

After the final alignment was generated, a HMMER model was built with the hmmb program, using this alignment as the input or training set. The final HMMER model based on the full length genomes has been tested here with partial genomes as well. Using the HMMER -R option for ragged ends (gaps inserted at the ends of sequences are given very low weight) the HMMER program did a reasonable job of aligning the complete and partial env genes to each other. The model was used again to align the complete genomes plus the env gene sequences, and in this case all sequences were reasonably aligned to each other. We are in the process of making these models available at our web site.

The annotation. The annotation for the precursor peptide cleavage sites in Gag and Gag-Pol is based on the information published in [Tozser et al.(1991), Le Grice et al.(1989)]. The annotation of the Gag-Pol ribosomal slip site is based on information published in [Reil et al.(1993), Kollmus et al.(1994), Le et al.(1989)]. The annotation for the cis-acting transcriptional activation domains in the LTR section is based on information published in [Zhang et al.(1997), Estable et al.(1996), Montano et al.(1997), Gao et al.(1996)]. There are a varying number of NF-κB binding sites in C subtype sequences, with some sequences carrying an additional site [Gao et al.(1996), Carr et al.(1996), Montano et al.(1997)]. The annotation for the Rev responsive element (the RRE) is based on [Charpentier et al.(1997)].

The B_FR.HXB2 reference nucleotide reference sequence is translated into all three reading frames at the top of the alignment using the single character amino acid designation. At the bottom of the alignment, protein sequences, based on the B_FR.HXB2 sequence are indicated; the HIV genome has many overlapping coding regions, and all are shown. For more complete annotation of functional domains see the protein sequence alignments in Part II.

REFERENCES

[Carr et al.(1996)] J. K. Carr, M. O. Salminen, C. Koch, D. Gotte, A. W. Artenstein, P. A. Hegerich, L. S. D., D. S. Burke, & F. E. McCutchan. Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. *J Virol* **70**:5935–43, 1996.

[Charpentier et al.(1997)] B. Charpentier, F. Schultz, & M. Rosbash. A dynamic *in vivo* view of the HIV-1 Rev-RRE interaction. *J Mol Biol* **266**:950–962, 1997.

[Estable et al.(1996)] M. C. Estable, B. Bell, A. Merzouki, J. S. Montaner, M. V. O'Shaughnessy, & I. J. Sadowski. Human immunodeficiency virus type 1 long terminal repeat variants from 42 patients representing all stages of infection display a wide range of sequence polymorphism and transcription activity. *J Virol* **70**:4053–62, 1996.

[Gao et al.(1996)] F. Gao, D. L. Robertson, S. G. Morrison, H. Hui, S. Craig, J. Decker, P. N. Fultz, M. Gerard, G. M. Shaw, B. H. Hahn, & P. M. Sharp. The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. *J Virol* **70**:7013–29, 1996.

[Kollmus et al.(1994)] H. Kollmus, A. Honigman, A. Panet, & H. Hauser. The sequences of and distance between two cis-acting signals determine the efficiency of ribosomal frameshifting in human immunodeficiency virus type 1 and human t-cell leukemia virus type ii *in vivo*. *J Virol* **68**:6087–91, 1994.

[Le et al.(1989)] S. Y. Le, J. H. Chen, & J. V. Maizel. Thermodynamic stability and statistical significance of potential stem-loop structures situated at the frameshift sites of retroviruses. *Nucleic Acids Res* **17**:6143–52, 1989.

[Le Grice et al.(1989)] S. F. Le Grice, R. Ette, J. Mills, & J. Mous. Comparison of the human immunodeficiency virus type 1 and 2 proteases by hybrid gene construction and trans-complementation. *J Biol Chem* **264**:14902–8, 1989.

[Montano et al.(1997)] M. A. Montano, V. A. Novitsky, J. T. Blackard, N. L. Cho, D. A. Katzenstein, & M. Essex. Divergent transcriptional regulation among expanding human immunodeficiency virus type 1 subtypes. *J Virol* **71**:8657–65, 1997.

[Reil et al.(1993)] H. Reil, H. Kollmus, U. H. Weidle, & H. Hauser. A heptanucleotide sequence mediates ribosomal frameshifting in mammalian cells. *J Virol* **67**:5579–84, 1993.

[Tozser et al.(1991)] J. Tozser, I. Blaha, T. D. Copeland, E. M. Wondrak, & S. Oroszlan. Comparison of the HIV-1 and HIV-2 proteinases using oligopeptide substrates representing cleavage sites in gag and gag-pol polyproteins. *FEBS Lett* **281**:77–80, 1991.

[Zhang et al.(1997)] L. Zhang, Y. Huang, H. Yuan, B. K. Chen, J. Ip, & D. D. Ho. Genotypic and phenotypic characterization of long terminal repeat sequences from long-term survivors of human immunodeficiency virus type 1 infection. *J Virol* **71**:5608–13, 1997.