HIV-1

Introduction This year many new full length viral sequences have become available, originating from diverse geographic origins and representing the spectrum of known HIV variation. Thus we have decided to publish only full length HIV-1/CPZ and HIV-2/SIV 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 November 1997 there were 63 complete (or nearly complete) HIV-1 genomes in the database (all are listed in Table 1). Of these, some were not included in the printed alignment, as they are very closely related, but the complete alignment including all sequences is available at our WEB (http://hiv-web.lanl.gov/) and ftp sites. Ten sequences were excluded because they were related to the IIIB/LAI lab strain of HIV-1, which is represented by the two sequences LAI and HXB2 in the alignment. Three more sequences have regions which contain IIIB/LAI, spliced onto regions of new isolate sequence (F12 Z11530, and WC001 AF003888, AF003887). A fourteenth sequence was omitted because it was a re-sequencing of the Z2Z6 sequence (Z2Z6 K03458) already represented by Z2Z6 M22639. This left 49 HIV-1 sequences in the alignment. Sequences from viral strains isolated from chimpanzees, CPZANT and CPZGAB, were also included. In phylogenetic analyses, the CPZ sequences are the closest simian-derived viruses to HIV-1; in fact HIV-1 M group and HIV-1 O group sequences are roughly as distant from one another as they are from the CPZ sequences.

All 63 of the complete genomes for HIV-1 have been updated with annotation of the major gene start and end sites. All are available from the HIV database WWW site (http://hiv-web.lanl.gov/) by using the sequence search interface (http://retro.lanl.gov/hivDB_search/index.html) to search for HIV-1 sequences with length ranging from 5,000 to 12,000 bases (this will select for only the full length or near complete genome sequences when using in the search tool).

The sequences are identified by their common name preceded by the HIV subtype designations 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 the 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. The “I” subtype is still in the process of being characterized. (See Robertson, D., et al., part III pages 25-30, of this compendium, (1997), for further details concerning the I subtype; the patterns of the inter-subtype recombination breakpoints for the all of the recombinant full length genomes are also can be found in Robertson et al.)

Alignment This alignment was generated by using the HMMER Hidden Markov Model sequence alignment software developed by Sean Eddy (available at http://genome.wustl.edu/eddy/hmmer.html and http://predict.sanger.ac.uk/mirrors/hmm/hmm.html), using an iterative process 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, A., Trends in Biochem. Science, 13:321-322 (1988)), and SE-AL V1.d1, 1995, a beta test version of an alignment program developed by Andrew Rambaut at Oxford), followed by rebuilding a new HMMER model and realigning the sequences, and then more hand editing. 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.

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 same set of complete genomes were then realigned using this model, to test the performance of HMMER with this model. The hmma -R option (for ragged-end alignment due to not all LTRs being complete) was used. The resulting alignment was very comparable to the alignment presented here, except that gaps were not preferentially placed to preserve codons. The average nucleotide identity between B_WEAU and the other 60 sequences was 0.88572 in the automated alignment and 0.88520 in the alignment presented here. Thus the automated alignment is closer to optimal than this alignment which has been adjusted by hand to be more presentable (*i.e.*, gaps moved from within codons to between codons).

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 cleaveage 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-kappaB 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 WEAU 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 WEAU 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.

HIV-2/SIV

Alignments of the 26 HIV-2/SIV full length genomes are included as a separate nucleotide alignment in Part I, following a similar annotation strategy as the HIV-1/CPZ alignment. Table 3 includes a list of all sequences included here, the subtype associated with each sequence, the GenBank accession number, and the primary reference. Sometimes a second reference is included if it contains basic information about the sequence or the isolate.

This alignment is also available at our WEB (http://hiv-web.lanl.gov/) and ftp sites. All corresponding sequence entries are available from the HIV database WWW site (http://hiv-web.lanl.gov/) by using the sequence search interface (http://retro.lanl.gov/hivDB_search/index.html) to search for HIV-2 and SIV sequences with length ranging from 9,000 to 12,000 bases.

The alignment was prepared by adding new sequences to an existing alignment using an in-house revised version of MASE, Multiple Alignment Sequence Editor (Faulkner, D., and Jurka, A., Trends in Biochem. Science, 13:321-322 (1988)). Phylogenetic analysis using neighbor-joining was used to check subtype designations. The STM (stump-tailed macaque) sequence is not included in the SIV-SD subtype since it is phylogenetically distinct from the other simian sequences.

The HIV-2 isolate BEN was selected as the reference sequence for this alignment since it is representative of the larger HIV-2 subtype, A, and it is one of the longer HIV-2 sequences available.

SIV sequences from other simian species such as African Green monkeys (AGM), sykes (SYK), and mandrill (MND) are not presented in this alignment. They are phylogenetically more distant from the HIV-2s than the macaques and mangabeys. The complete genome nucleotide sequence alignment of the sequences from African Green monkeys will be presented on the web and ftp sites when it is complete.

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. LA - Eng. *J Virol* **70**:5935–43, 1996. OTE: (Medline: 96323109), (Genbank: U54771).

[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.