HIV-1

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-1/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 December 1998 there were 129 complete or nearly complete, (defined as greater than 7,000 consecutive basepairs of sequence) 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 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_99/ALIGN-INDEX-99.html

Fifty-nine HIV-1 sequences plus viral strains isolated from chimpanzees, CPZANT and CPZGAB, comprise the printed alignment. The CPZUS sequence (AF103818) was not released in time to be included. In phylogenetic analyses, the CPZ sequences are the closest simian-derived viruses 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.

All 129 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 7,000 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 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. The "I" subtype has been characterized and AGI(CY032) circulating recombinant form. See Carr, J., *et al*., part III pages III-10 to III-19, of this compendium, for further details concerning this recombinant form; the patterns of the inter-subtype recombination breakpoints for most of the recombinant full length genomes also can be found in Robertson *et al*. *Human Retroviruses and AIDS 1997*, pages III-20 to III-25. The classification of the G subtype is still under discussion, and it has some regions which are associated with the A, and possibly even the E subtype. This year we are including a new classification system, circulating recombinant forms (CRFs), and the E subytpe has been renamed as a AE CRF; the AE CRF has a reference strain CM240, the first isolate to be sequenced that had the recombinant structure of E in env, A essentially everywhere else. See Carr, J., *et al*., part III pages III-10 to III-19, of this compendium, for further discussion of circulating recombinant forms.

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

http://predict.sanger.ac.uk/mirrors/hmm/hmm.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 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. 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 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-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.HXB2R 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.HXB2R 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 were included as a separate nucleotide alignment in Part I of *Human Retroviruses and AIDS 1997*. For the 1998 Compendium, we present an alignment of all classes of primate immunodeficiency viruses from HIV-1 to HIV-2 and representatives of each type of SIV sequenced to date. Complete description and tables for that alignment follow the HIV-1 alignment.

HIV Database Compendia from previous years, as well as electronic copies of these alignments in a variety of formats are available on our WWW site at

http://hiv-web.lanl.gov/HTML/compendium.html

and

http://hiv-web.lanl.gov/ALIGN_99/old_alignments.html

respectively. The electronic copies of this years alignments are at:

http://hiv-web.lanl.gov/ALIGN_99/ALIGN-INDEX-99.html

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.