Introduction

This compendium is the annual printed version of the HIV database, summarizing our effort to compile, organize, and rapidly publish as much relevant molecular data concerning the human immunodeficiency viruses (HIV) and related retroviruses as possible. The scope of the compendium and database is best summarized by the three parts that it comprises: (I) Nucleic Acid Alignments, (II) Amino Acid Alignments, and (III) Reviews and Analyses. Information concerning all the parts is updated throughout the year on the Web site:

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

This year we are not including a part IV regarding related human DNA sequences, although the 1997 compendium had such a summary that is available on our web site. We may include this section again in future years.

While this publication could take the form of a review or sequence monograph, it is not so conceived. Instead, the literature from which the database is derived has simply been summarized and some elementary computational analyses have been performed upon the data. Interpretation and commentary have been avoided insofar as possible so that the reader can form his or her own judgments concerning the complex genetic information. The exception to this are reviews submitted by experts in areas deemed of particular and basic importance to research involving AIDS viral sequence information. These are included in Part III, and are contributed by scientists with particular expertise in the area of interest. In addition to the general descriptions below of the parts of the compendium, the user should read the individual introductions for each part.

This year we have changed our nomenclature slightly. There have been several recombinant forms that have been identified in multiple unrelated individuals, and so appear to be important epidemic strains. We have decided to call these circulating recombinant forms, and append the name of the first isolate sequenced with a given mosaic structure to the subtype designation. A major consequence of this change is that sequences that were formerly designated subtype E will now be identified as circulating recombinant form AE (CM 240), as subtype E was only clearly defined in the env region, and the gag and pol regions of these AE genomes were not clearly distinguishable from the A subtype.

Part I. Nucleic Acid Alignments. Annotated nucleic acid sequence alignments of representative complete genomes of HIV-1/CPZ (Section A) and HIV-1/HIV-2/SIV (Section B) are presented. The hard-copy annotation includes coding regions, regulatory structures, splice sites, and other features of functional significance. The basis for this annotation is often conservation between strains, (the recurrence of patterns such as TATAA and AATAAA), and a brief listing of some of the references used for the annotation is included. Although our practice has been to conservatively annotate, we caution the user against docility: sequence information regarding transcripts, for example, is neither certain or complete at this time. Descriptions of the sequences, and the isolates from which they were derived, are provided, as well as additional references. Beginning in 1997, only full-length nucleotide sequences are presented in the hard copy edition, to save space. Other HIV and SIV sequences are catalogued, and long fragments of coding regions for particular genes are aligned as part of our effort; these coding region alignments are available at our Web or ftp sites. The full formatted GenBank style entries of these sequences are located on the Web site:

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

and the database FTP server. To a basic GenBank entry, we add extensive comment lines and additional fields such as date of isolation, phenotype, and health status of the patient, and updated references that refer to a given sequence, as is possible. The comment lines for the full length genomes included in the

printed alignment are also included in the printed version. As of this writing, there are approximately 30,000 HIV sequences in GenBank. Clearly we cannot add annotation to all of them, and so we emphasize enriching entries of full length genomes and genes. We also provide alignments of all gene regions on the web site, and a tool for users to tailor-make alignments including only sequences that meet user-selected criteria (sequences from a specific gene region, country of origin, or subtype).

Part II. Amino Acid Alignments. This section contains alignments of the amino acid sequences (mostly full-length) of all known coding regions, and open reading frames of HIV-1/CPZ (Section A), HIV-2/SIV (Section B), and SIV AGM/SYK (Section C). Sequences representing the range of global variation, including commonly used reference strains, were selected for inclusion in the printed copy. Other alignments with more complete sets of sequences are available on the Web site. Protein processing sites are annotated when known, as well as key functional domains and overlapping reading frames. The reader should consult the introduction to Part II for further explanation of the presentation and annotation of the amino acid sequences.

Part III. Analyses. This section is open-ended with the constraint that the sequence analyses and compilations be basic and of interest to the diverse community of database users. In 1998, analyses and curatorial contributions include: updating of tables of mutations relating to drug resistance, a summary of SHIV constructs, a review of the structure of gp120, an updated reference set of sequences for use in HIV-1 subtyping, background information on the second receptor usage of different isolates, and a numbering scheme to facilitate finding positions in the HIV genome.

Inquiries regarding the database should be addressed to:

Bette Korber and Carla Kuiken
Theoretical Division T-10, MS K710
LANL Los Alamos, NM 87545

(505)-665-4453; fax (505)-665-3493
e-mail:
btk@t10.lanl.gov
kuiken@t10.lanl.gov

A short glossary follows.