An HIV Database

Three repositories for nucleotide sequence data exist, all established in the eighties. One is headquartered at Los Alamos National Laboratory (see "GenBank" by Walter B. Goad in *Los Alamos Science,* Number 9, 1983), another at the European Molecular Biology Laboratory in Heidelberg, West Germany, and the third at the National Institute of Genetics in Mishima, Japan. Those databases provide easy access to nucleotide sequences determined by researchers worldwide.

The rationale for establishing a separate database devoted exclusively to nucleotide sequences of human immunodeficiency virus samples is well stated in the preface to *Human Retroviruses and AIDS 1987, the* first edition of that new database:

> From the initial publications of the nucleotide sequences of [HIV samples] in 1985, as well as from the early comparisons of restriction enzyme maps of many HIV isolates, it became evident that the HIV genome could exercise considerable heterogeneity. These early results encouraged further sequencing aimed at delineating the extent and meaning of the initially observed variation. These follow-up studies, reported in the summer of 1986, showed conspicuous sequence variation among isolates obtained from North America and Africa and, in one study, pointed to the phenomenon of "swarming''-distinct sequences could be cloned over time from individual patients. The earlier hypothesis of a small number of stable variants-a New York strain, a San Francisco strain, etc.-could not be straightforwardly upheld. And, for a while at least, sequencing would remain an ongoing endeavor. Thus it became likely that systematic compilation and analysis of HIV sequences would contribute to the identification of conserved and variable elements of the viral genome, and perhaps even to the geographical and temporal tracking of variants.
>
> At that time the U.S. and European nucleic acid sequence databases-Genbank and EMBL-Heidelberg-having the formidable task of annotating and maintaining more than ten million bases of sequence data, were hard pressed to preferentially annotate and enter the numerous sequences associated with the new human retroviruses, and the likelihood that individual laboratories could keep up with the growing information was unthinkable. We proposed, then, to the Acquired Immunodeficiency Syndrome Program of the National Institute of Allergy and Infectious Diseases that an HIV sequence database be established [at Los Alamos] and that a quarterly publication of the compiled information be made readily available, at no cost, to all interested investigators. The principal goal of the database would be to eliminate all acquisitional barriers to sequences so that comparison and analysis could keep pace.

The NIAID agreed to our proposal, and the third edition of the database will be appearing in early 1989. Updates to the yearly editions are published as acquisition of new data warrants.

The database now includes sequences for about thirty HIV samples. The sequences for about a quarter of the samples span the entire viral genome of approximately 10,000 nucleotides. In addition, the database includes nucleotide sequences for several other retroviruses of relevance to the study of HIV: three simian immunodeficiency viruses, three non-human lentiviruses, and two human oncoviruses. Also included are the amino-acid sequences (as predicted from the nucleotide sequences) for various protein-coding regions of the HIV genome. Floppy diskettes containing the nucleotide and amino-acid sequence data are provided.

The staff of the database center not only compiles and publishes sequence data but also, as time permits, carries out some sequence analysis, the results of which are included in the database. For example, we have aligned the nucleotide sequences for the 5' long terminal repeats of fifteen HIV1 samples and have deduced evolutionary relationships among various samples from differences among their nucleotide sequences. We have also undertaken, in collaboration with Chang-Shung Tung, helix-twist analyses of the regulatory sequences of the long terminal repeats of HIV.

It now seems likely that the HIV sequence project will be continued at least through 1995. Its future scope will include research on molecular aspects of HIV immunology. The results of such research should prove valuable in developing antiviral drugs and vaccines. ■