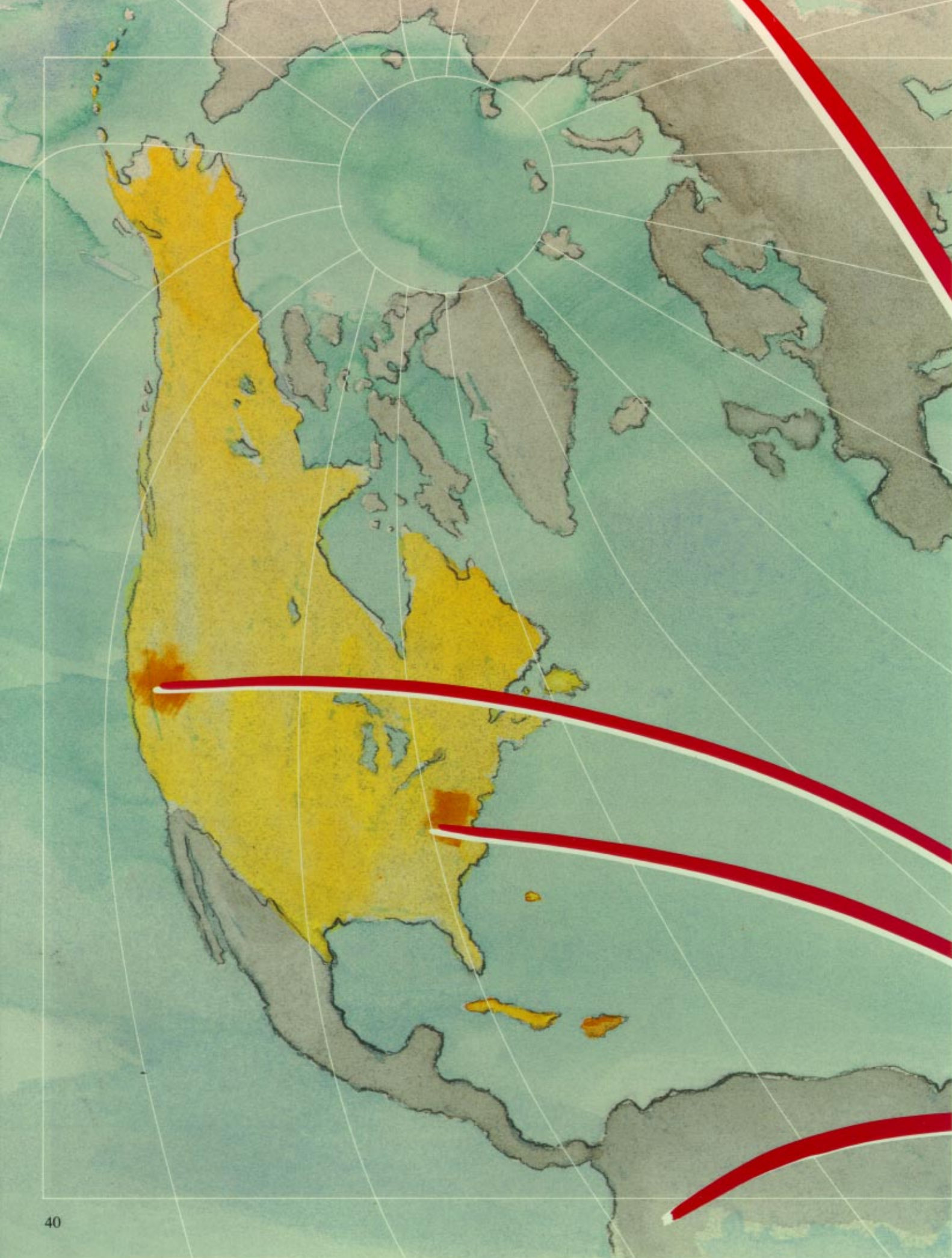
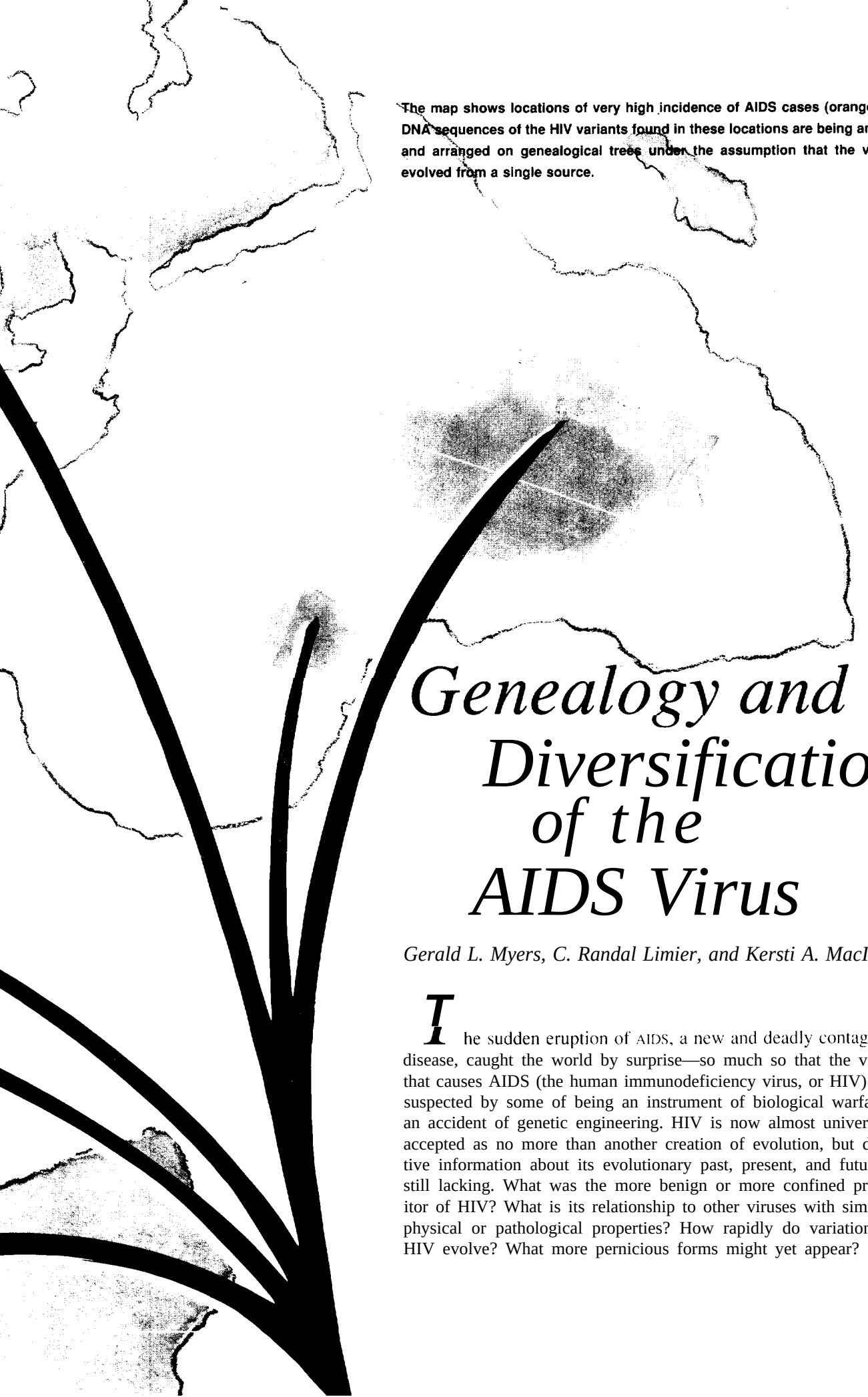




The map shows locations of very high incidence of AIDS cases (orange). The DNA sequences of the HIV variants found in these locations are being analyzed and arranged on genealogical trees under the assumption that the variants evolved from a single source.

Genealogy and Diversification of the AIDS Virus

Gerald L. Myers, C. Randal Limier, and Kersti A. MacInnes

The sudden eruption of AIDS, a new and deadly contagious disease, caught the world by surprise—so much so that the virus that causes AIDS (the human immunodeficiency virus, or HIV) was suspected by some of being an instrument of biological warfare or an accident of genetic engineering. HIV is now almost universally accepted as no more than another creation of evolution, but definitive information about its evolutionary past, present, and future is still lacking. What was the more benign or more confined progenitor of HIV? What is its relationship to other viruses with similar physical or pathological properties? How rapidly do variations of HIV evolve? What more pernicious forms might yet appear?

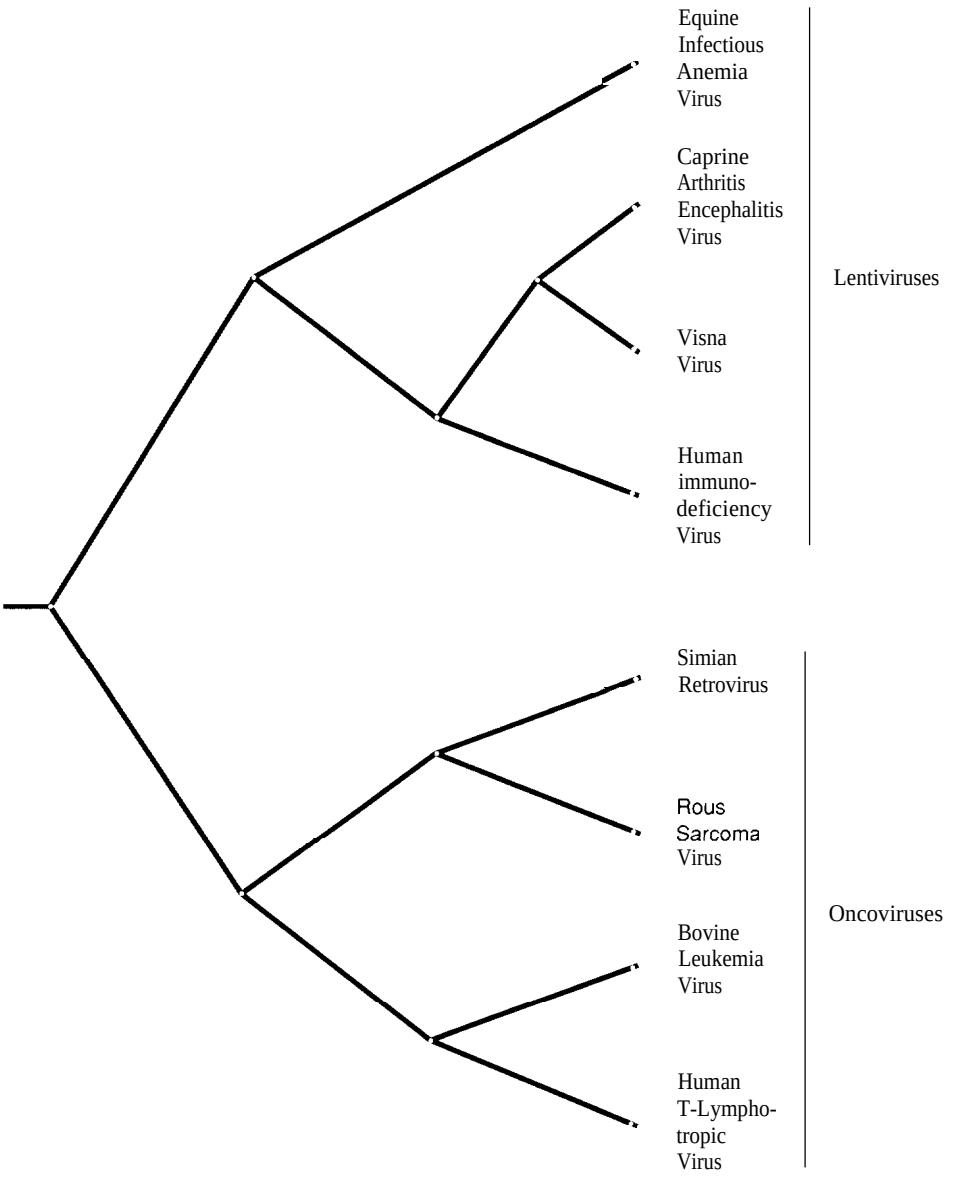


Fig. 1. A family tree for a set of retroviruses based on differences among the amino-acid sequences of their proteases and reverse transcriptases. The tree depicts the "consensus" pattern of evolution, that is, the pattern in agreement with analyses by a number of investigators. Unlabeled dots denote assumed intermediate ancestors. The closer relationship of HIV to known lentiviruses than to known oncoviruses agrees with the classification of HIV as a lentivirus.

Such questions are being addressed by analyzing the characteristics of HIV at the molecular level, as the following example illustrates. HIV is a retrovirus and as such has a genome composed of RNA rather than DNA (see "Viruses and Their Lifestyles"). The first step in the replication of a retrovirus is synthesis of DNA from the RNA template provided by the viral genome. That synthesis is catalyzed by enzymes—reverse transcriptases—that are virtually unique to retroviruses. Likely evolutionary relationships among HIV and other disease-causing retroviruses have been deduced from the differences among the sequences of amino acids that compose their reverse transcriptases. The same has also been done for retroviruses by focusing on their proteases, enzymes common to all organisms and essential to the breakdown of other proteins.

Figure 1 shows those evolutionary relationships depicted, as is customary, in the form of a phylogenetic "tree." Note that the analysis relates HIV more closely to the lentiviruses (which cause slowly developing, chronic diseases affecting the lungs, joints, and nervous, hematopoietic, and immune systems) than to the oncoviruses (which cause cancer, often of blood cells). That closer relation agrees with the classification of HIV as a lentivirus on the basis of other characteristics, including the pathology of AIDS. However, the similarity between the protease amino-acid sequences of HIV and of, for example, the visna lentivirus (a homology of about 40 percent) implies only a relatively close relation between the two viruses, comparable to that between humans and fungi.

The idea of deducing evolutionary relationships from molecular rather than, say, anatomical or morphological data was first proposed in the early sixties, roughly a decade after Sanger and his colleagues published the first amino-acid sequence of a protein (insulin). Figure 2 shows an early example of that application of amino-acid sequence data: a phylogenetic tree for aerobic organisms based on differences among the amino-acid sequences of the cytochrome *c's* of about thirty extant species. (Cytochrome *c*, a protein essential to all aerobic organisms, is among the more highly "conserved" proteins; in particular, a time lapse of 20 million years after the divergence of two evolutionary lines is required to produce a change of 1 percent in the amino-acid sequences of their cytochrome *c's.*) The tree has the same topology, or branching pattern, as do trees based on more conventional biological data.

Changes in the amino-acid sequences of proteins are among the raw materials for natural selection and evolution of new organisms. But those changes are themselves the direct results of changes in the sequences of nucleotides that compose DNA (or, in the case of retroviruses, RNA) and encode the proteins. Now, roughly a decade after Sanger and Maxam and Gilbert developed methods for sequencing DNA (and RNA), nu-

EVOLUTIONARY RELATIONSHIPS AMONG AEROBIC ORGANISMS

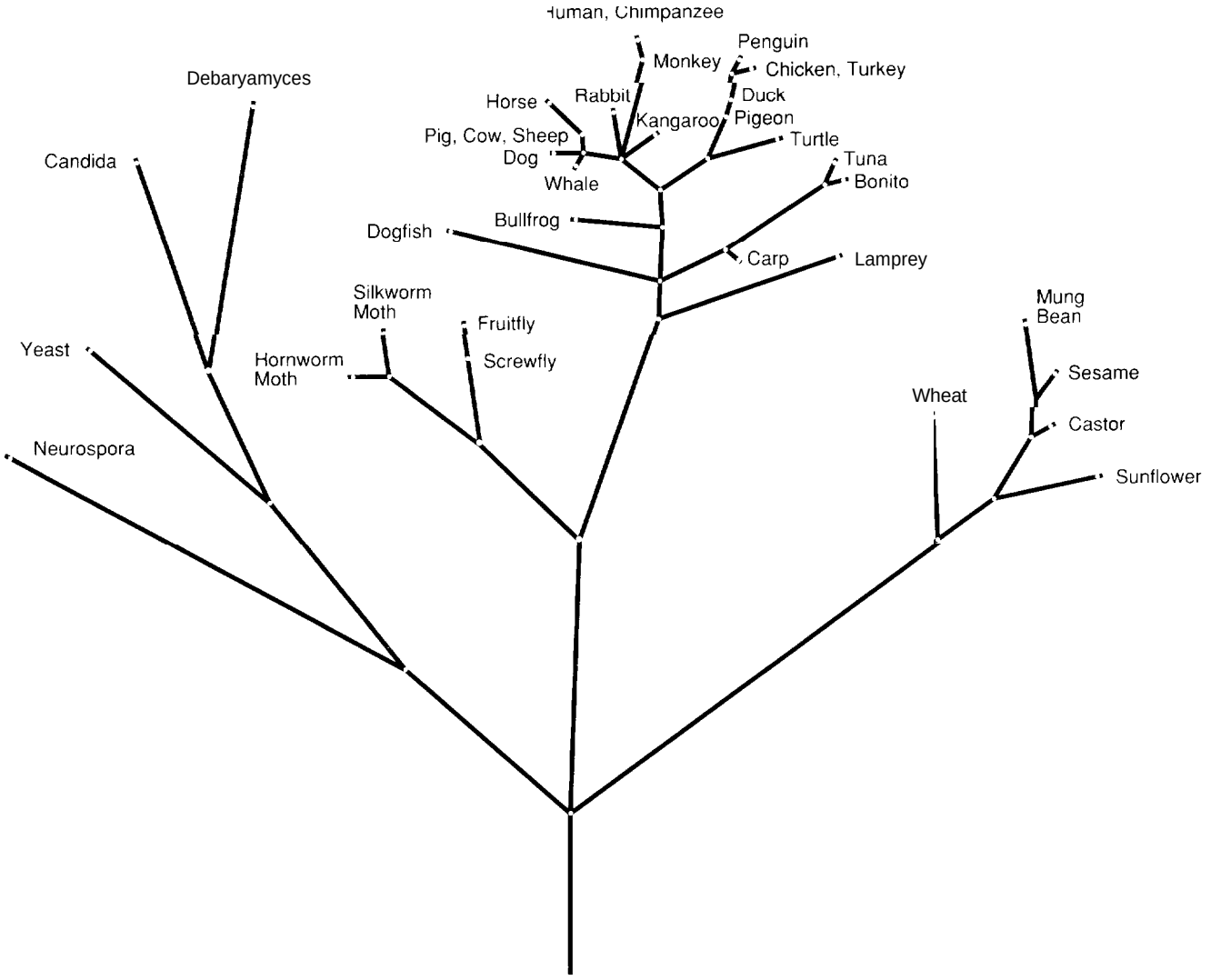
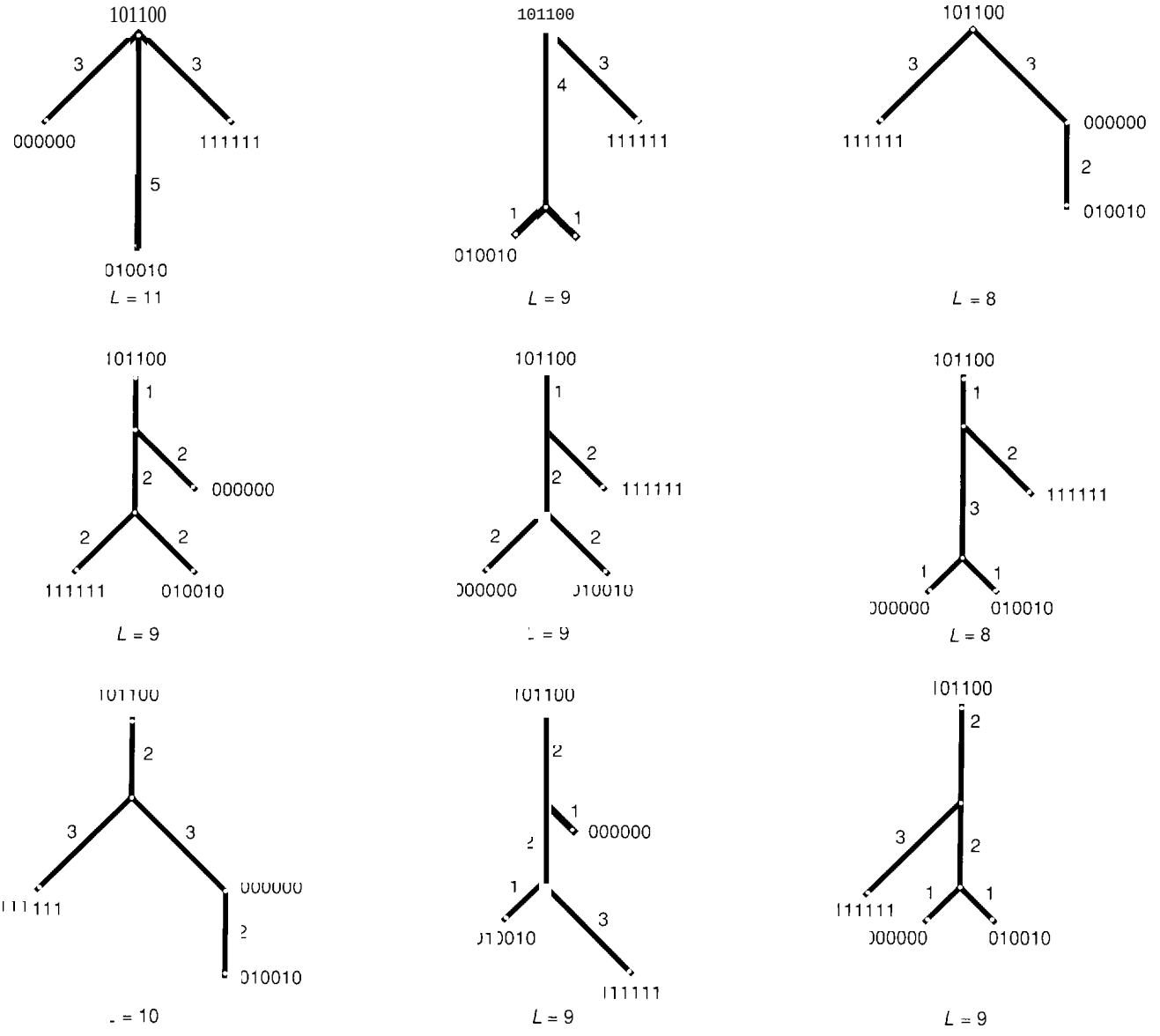


Fig. 2. A phylogenetic tree for aerobic organisms based on differences among the amino-acid sequences of the cytochrome *c*'s of about thirty extant fungi, insects, amphibians, mammals, birds, fish, and plants. The distance between any two dots is proportional to the dissimilarity between the cytochrome *c*'s of the organisms represented by the dots. Unlabeled dots represent assumed intermediate ancestors. (Adapted from a figure in the article "Building a phylogenetic tree: Cytochrome *c*" by M. O. Dayhoff, C. M. Park, and P. J. McLaughlin. In *Atlas of Protein Sequence and Structure 1972,* edited by Margaret O. Dayhoff, 7-16. National Biomedical Research Foundation, Washington, D. C., 1972.)

cleotide sequence data are supplying-information relevant to nearly all aspects of biology and medicine, including evolution and AIDS. Such data can provide more detail about recent evolutionary changes than amino-acid sequence data. (The genetic code is highly "degenerate"; that is, almost all amino acids are specified by more than one triplet of nucleotides. Thus many nucleotide changes do not result in protein changes.) Recognizing that, we have been using nucleotide sequence data to construct phylogenetic trees for HIV and its close relatives. The number of HIV samples for which such data are available is at present rather modest (the first appeared only in 1985) but is increasing rapidly—in fact, so rapidly that a center for compilation and analysis of the data has been established at the Laboratory (see "An HIV Sequence Database").

The procedure for constructing a phylogenetic tree from nucleotide sequence data is based on a concept introduced by Stanislaw Ulam: a "distance" between a pair of sequences. The simplest definition of such a genetic distance, and the one we employed, is the number of nucleotide substitutions required to transform one sequence into the other. More complicated definitions of a genetic distance include other biologically possible changes to a nucleotide sequence, such as insertions or deletions of nucleotides or relocations of fragments of sequences.

To determine a phylogenetic tree for a set of sequences, one must first assume a "root," that is, a sequence ancestral to all the given sequences. (A random se-



EVOLUTION OF BINARY SEQUENCES

Fig. 3. Consider the set of binary sequences 000000, 111111, end 010010. Assume that those sequences have been produced from an "ancestral" sequence 101100 by successive substitutions of 1's for 0's and O's for 1's. There are many paths by which the given sequences might have "evolved" from the ancestral sequence. Some of those paths are shown above as phylogenetic trees. Assumed intermediate ancestors are depicted in gray. Below each tree is listed its length *L*, which equals the total number of substitutions involved in the path represented by the tree. Note the two trees of minimum length *(L* = 8)The idea of the evolution of binary sequences can be extended to the evolution of new organisms by point mutations In the quaternary nucleotide sequences that encode a protein. In that context the tree (or trees) of minimum length is assumed to represent the most likely course of evolution, subject to the condition that all of the given (or extant) sequences appear at branch tips.

quence is often chosen as the root.) Then one considers all branching patterns that lead from the root to the given sequences. (Most of the branching patterns will include postulated branch points, or intermediate ancestors.) The branching pattern that minimizes the sum of the branch lengths (each of which is the genetic distance between a sequence and its immediate ancestor) is the phylogenetic tree that is most consistent with the assumption that evolution proceeds for the most part with maximum parsimony. (HIV exemplifies the conditions under which that assumption is most valid: a high mutation rate and recent divergence.)

The procedure outlined above is illustrated in Fig. 3 for a small set of short sequences of the digits O and 1. (Nucleotide sequences are of course not binary but quaternary, since DNA and RNA are polymers of four different nucleotides. Binary sequences were chosen as the example in Fig. 3 for simplicity.) The analysis is more complex when, as is true in practice, the sequences are longer and their number is greater. In addition, since insertions and deletions of nucleotides do

EVOLUTIONARY RELATIONSHIPS AMONG HIV1, HIV2, AND SIV

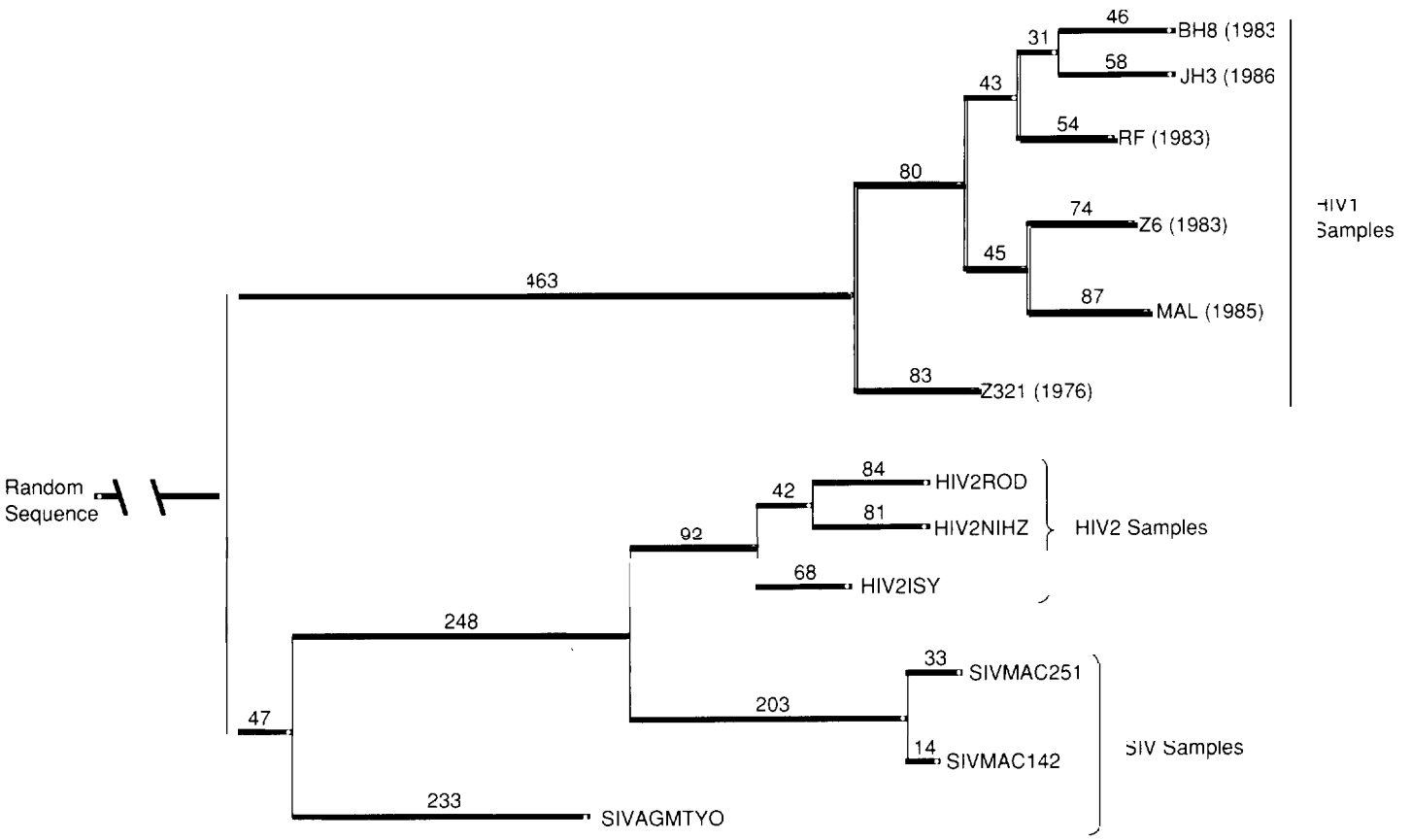


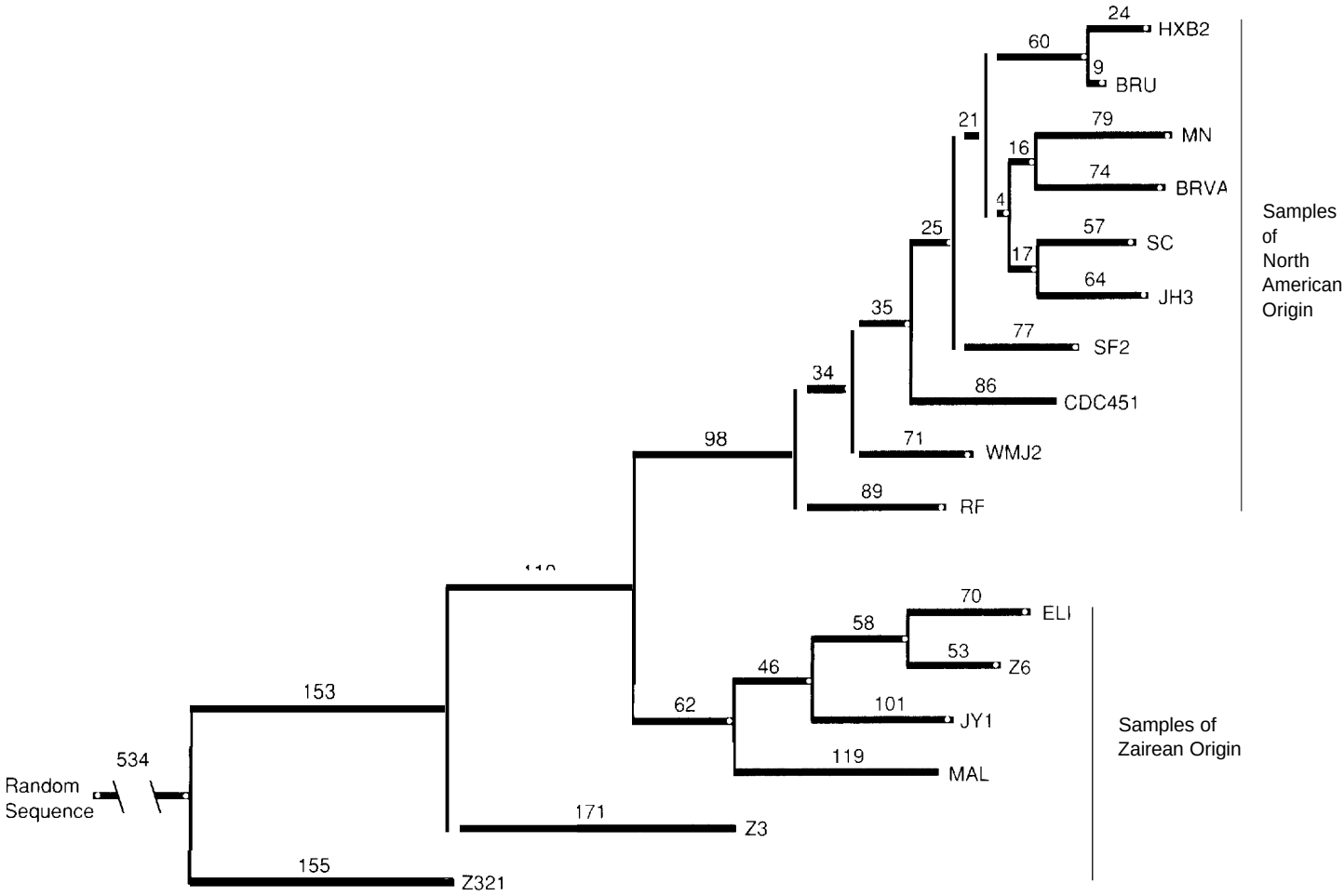
Fig. 4. A phylogenetic tree for the two currently recognized types of HIV (HIV1 and HIV2) and the simian Immunodeficiency virus (SIV) based on the differences among the nucleotide sequences of the *env* regions of the viral genomes. The total number of nucleotide sites included in the analysis is 1290. The horizontal distance between any two dots is proportional to the listed branch length, that is, to the genetic distance between the sequences represented by the dote. (Unlabeled dots denote assumed Intermediate ancestor.) Thus, for example, the *env* nucleotide sequences of the samples labeled BH8 and RF differ at 131 = 46+ 31 +54 sites out of 1290. Note the relatively close relationship between SIV isolated from captive macaques and HIV2 and the relatively distant relationship between SIV isolated from macaques and SIV isolated from wild African green monkeys. The ten-year span between the dates of isolation of the Z321 and JH3 samples permits an approximate temporal calibration of the tree. That calibration places the latest possible date for divergence of HIV1 and HIV2 at about 1950.

occur, the sequences are often not of the same length. Then gaps in the shorter sequences must be assumed as necessary to align regions of the sequences that are observed to be more highly conserved. Such regions are assumed to code for critical features of the protein. The nucleotides in the longer sequences corresponding to the gaps in the shorter sequences are ignored in the analysis. In other words, the phylogenetic trees are based on information derived only from nucleotide positions common to all the sequences and showing some variation. The alignment of sequences, therefore, must be performed with great care.

Another complication arises from the extremely large number of branching patterns that must be examined in the search for the one (sometimes more than one) of minimum length, Algorithms are available for directing a computer to perform the search; we used the PAUP (phylogenetic analysis using parsimony) algorithm developed by David L. Swofford of the University of Illinois.

Finally, as mentioned above, an abundance of nucleotide sequence data for HIV and its near relatives is not yet available, Despite advances in the technology of sequencing, the process is still time-consuming and expensive: the viral samples must be isolated and cultured and the genomes fragmented, cloned, and then sequenced. Furthermore, some researchers are reluctant or unable (because of agreements with private companies) to make public the sequence data they have obtained. And, understandably, many professional sequencers do not want such a dangerous human virus in their laboratories.

The trees we construct for HIV and its close relatives are based on nucleotide sequences for various regions of the viral genomes. The topologies of the trees are remarkably similar irrespective of the genomic region upon which each is based, and that similarity



EVOLUTIONARY RELATIONSHIPS AMONG HIV1 SAMPLES

Fig. 5. A phylogenetic tree for a set of HIV1 samples based, like the tree in Fig. 4, on differences among the nucleotide sequences of the *env* regions of the viral genomes. Again horizontal distances are proportional to the listed branch lengths, and unlabeled dots denote assumed intermediate ancestors. Note that the tree cleanly separates the samples by geographic origin. The story behind the close relation between the samples labeled JH3 and SC, which were isolated from a Japanese victim of AIDS and a Californian victim of AIDS, is told in the text.

gives us considerable confidence in our results. Figure 4 shows a version of a tree based on the *env* regions (see "Viruses and Their Lifestyles") of the HIV1, HIV2, and SIV (simian immunodeficiency virus) genomes. HIV1 and HIV2 are the currently known types of HIV. Both cause similar but not identical pathologies; that of HIV2, the less widespread type, is perhaps less lethal. SIV causes an AIDS-like disease in some captive non-human primates. The tree indicates that HIV2 is more closely related to SIV than is HIV1 and that HIV2 is more closely related to the type of SIV isolated from captive macaque monkeys than to the type isolated from wild African green monkeys. Resolving the question of whether the progenitor of HIV and SIV is a human or a simian virus will require sequence data for a greater number of samples, particularly of simian viruses. (The nucleotide sequences of immunodeficiency viruses isolated from wild chimpanzees and mandrills will be known in 1989.)

The samples labeled Z321 and JH3 in Fig. 4 date to 1976 and 1986, respectively. (Z321 was isolated from a stored blood sample.) The ten-year span between those two samples, together with the genetic distance between them, permits an approximate temporal calibration of the tree, The calibration, which assumes a linear relation between genetic distance and time, leads to an estimate of about 1950 as the latest possible date of divergence of HIV1 and HIV2. That time scale for viral evolution is broadly consistent with the known history of the AIDS epidemic. However, to date both the epidemiologic records and the nucleotide sequence data are extremely sparse in information derived from occurrences of AIDS earlier than the eighties. Such "fossil" data, which may lie hidden in stored blood samples, would be especially revealing.

Figure 5 shows a version of a phylogenetic tree focused entirely on HIV I isolates; it also is based on nucleotide sequence data for the *env* region. That tree is of particular interest for several reasons. First, its "bushiness," which reflects an abundance of distinguishable variants, suggests that the HIV1 variants are not competing among each other for a restricted ecologic niche. (The tree remains bushy even when

pruned of those variants not yet proved to be infectious.) In contrast. the phylogenetic tree of the influenza virus is much less bushy, despite the fact that the data now available indicate that the two viruses exhibit roughly comparable rates of change in genetic distance (on the order of 1 substitution per 100 nucleotides per year). The changes in the genome of the influenza virus are manifest primarily as relatively infrequent appearances of different types. Apparently the influenza virus is occupying a narrow ecologic niche in which competition among variants is intense.

Second, the two main branches of the tree correspond to a grouping of the viral samples by geographic origin. In other words, the North American samples are more closely related to each other than to the African samples and vice versa. That geographic intelligibility mitigates strongly against a hypothesis that the variants of HIV1 have existed independently confined for a long time. How then did all break loose more or less simultaneously? Much more consistent with the phylogenetic analysis (and with the demonstrated existence in the same, singly infected individual of more than one variant*) is the hypothesis that diversification occurs in step with the growth of the epidemic. Such a hypothesis is not unreasonable since replication of the viral genome, which is necessary for production of daughter viruses that can spread to other cells in the same individual or to other individuals, often introduces changes in the viral genome. (Replication of the genome of a retrovirus is a two-step process: first, the genomic RNA serves as the template for synthesis of DNA, catalyzed by reverse transcriptase; and second, the resulting DNA serves as the template for synthesis of new genomic RNA. The first step, called reverse transcription, is not subject to any proofreading or error-correction mechanisms, and the DNA produced is often not an exact "transliteration" of the viral genome. Thus the second step often leads to a new viral genome that is not an exact copy of the original. Furthermore, evidence has recently been found that reverse transcription of the HIV genome is more error-prone than that of other retroviruses.)

A final point to note about the tree in Fig. 5 concerns the sample labeled JH3, which was isolated from a Japanese hemophiliac who contracted AIDS from transfusion with HIV-infected blood. The tree relates JH3 most closely to SC, a sample isolated from a Californian. How can the genetic similarity of the two variants be reconciled with their widely separated geographic origins? The answer lies in the source of the transfused blood: Japan imports much of its blood supply, primarily from the United States. That tale illustrates the utility of sequence data and phylogenetic analysis to tracking the course of the AIDS epidemic, a utility so great that the National Institutes of Health has embarked on an ambitious program in "molecular epidemiology"-a worldwide viral sampling and sequencing project to track HIV variants through space and time. Phylogenetic trees will not only provide a measure of the velocity of the AIDS epidemic but also help guide research and policy about vaccines.

The course of HIV evolution will become better defined as more nucleotide sequence data accumulate. But the data now at hand clearly reveal great diversification of the virus. That diversification will beget a spectrum of disease states under the umbrella of AIDS: "HIV1 disease, " "HIV2 disease," and so on.

What are the implications of HIV diversification for the hopes to control, to cure, to eradicate AIDS? There are several, and all but one are negative. HIV may become capable of infecting an even wider range of cell types (it now targets primarily T4 lymphocytes, macrophages, and glial cells) and may become pathogenic more rapidly. Conceivably, different modes of transmission of HIV may arise: insect vectors, colostrum, or respiratory aerosols, for example. HIV may develop resistance to azidothymidine (the only antiviral drug now available in the United States) and to future antiviral drugs. Tests for the presence of HIV may not detect newly evolved variants. (Data reported in the most recent edition of the HIV sequence database, *Human Retroviruses and AIDS 1989,* include HIV variants that, although detectable by the current test, reside on branches of the phylogenetic tree different from the one containing the group of variants on which the test is based.) And the diversification of HIV will only exacerbate the difficulties of developing and testing a vaccine for AIDS.

To offset the litany of negative implications is a single positive one. Analyses of nucleotide sequence data have pinpointed areas of the HIV *env* region that are relatively conserved. Those areas may prove to be "soft spots" at which to aim in the search for a vaccine. The more we learn about the invariant regions of the HIV genome, the better equipped we shall be to design intervention strategies.

The AIDS epidemic is the first major medical crisis to occur since molecular biology came of age. In the brief time since HIV was identified as the cause of that new disease, much has been learned about the virus, and few doubt that some detail of its molecular biology will be the key to its ultimate conquest or, at least, containment. ■

* Much research is currently centered on the diversification of HIV within a singly infected individual and on the changes in that diversification with time. However, to date very few results have been published.

Viruses and Their Lifestyles

Few can have escaped learning that once again a deadly virus is loose among humans. But what are viruses, and how do they subsist and reproduce?

Viruses are freeloaders, parasites that carry out their only function—multiplication-only by making free use of the metabolic and biosynthetic machinery of "host" cells, particularly their machinery for protein synthesis. The parasitism of some viruses is fatal to host cells; that of others is benign. Many kinds of viruses have evolved, each adapted to some bacterial, plant, or animal host. Of course defenses aginst viruses have also evolved, ranging from the restriction enzymes of bacteria to the immune systems of vertebrates. And, in the case of some animal viruses, such as poliovirus and rabies virus, research has provided vaccines that greatly strengthen the natural immune response to viral attack.

The complete extracellular form of a virus is called a virion. Its components are few: a genome of nucleic acid, a proteinaceous housing for the genome, and, in certain instances, a few molecules of a virus-specific enzyme. Multiplication of a virus requires replication of its genome and synthesis of the proteins the genome encodes. The host cell provides all of the energy and many of the biochemical needed to carry out those processes.

The genome of a virus may consist of either one of the two nucleic acids, DNA and RNA. The nucleic acid polymer may be single- or double-stranded, linear or circular. Viruses with genomes of RNA **are** unique: in all other organisms RNA is involved only in synthesis of proteins and not also in storage of genetic information. The smaller viral genomes encode as few as four proteins; the larger, which approach the size of small bacterial genomes, encode several hundred. (The human genome is thought to encode about 100,000 proteins.)

The housing enclosing a viral genome consists of a coat (capsid) made up of many copies of a very few types of virus-specific proteins. The architecture of the capsid is geometric; in fact, all simple viruses exhibit helical or icosahedral symmetry (that of a twenty-sided regular polyhedron) or a combination of the two. The housing of many of the more complex viruses includes an "envelope" surrounding the capsid. The envelope is very similar in structure and composition to the plasma membrane of the host cell, containing lipids derived from the cell and virus-specific glycoproteins.

The processes involved in the life cycle of a virus (more properly, its multiplicative, or reproductive, cycle, since viruses are not "living" organisms) include delivery of the viral genome to the interior of a host cell, replication of the viral genome, synthesis of the proteins encoded by the viral genome, assembly of the newly produced viral components into new virions, and exit of the virions from the host cell. Since the details of those processes are complex and vary from one kind of virus to another, only their general features are sketched here.

Delivery of the viral genome to the interior of a host cell ("infection" of a cell) is accomplished through site-specific and often cell-type-specific interaction of the capsid or its envelope with the cellular membrane. The site- and cell-specificities are the result of selective interaction between viral housing and receptors on the surface of the cellular membrane. The mechanisms of infection are varied. For example, the T4 phage infects *Escherichia coli* by injection, the Semliki Forest virus infects mosquito cells by receptor-mediated endocytosis (a normal cellular process by which proteins enter cells), and the human immunodeficiency virus infects T4 lymphocytes by fusion of the viral envelope and the lymphocyte membrane.

Infection of a cell is followed by replication of the viral genome and synthesis of the proteins it encodes. Since the features of those processes depend foremost on whether DNA or RNA composes the viral genome, that property is the basis for dividing viruses into two major classes.

The genome of a DNA virus is processed (transcribed and replicated) by an infected cell in much the same way that the cell processes its own DNA. That is, the viral DNA is used as the template for synthesis of viral messenger RNAs (which in turn serve as templates for synthesis of viral proteins) and as the template for synthesis of new viral DNA. However, only the simplest of viruses (whether DNA or RNA) entrust the production of new viral components entirely to the normal workings of a host cell. Instead, the genomes of most viruses include genes for enzymes that "reprogram" the cellular machinery toward preferential (sometimes exclusive) processing of the viral genome. Such reprogramming is necessary, for example, to achieve rapid replication of the genome of a DNA virus, since a cell normally synthesizes DNA only in preparation for cell division. In addition, the genomes of most viruses include sequences that regulate the timing and extent of gene expression.

Processing of the genomes of some DNA viruses does not always immediately follow infection. Instead the viral genome can become incorporated into that of the host cell. There it lies latent, its gene expression repressed, being passed silently through (typically) many generations of daughter cells. Ultimately, some stimulus triggers the exit of the viral DNA from that of the host, and its processing then begins.

Three types of RNA viruses are recognized. Two are distinguished on the basis of whether the genomic RNA or its complement serves as messenger RNA, that is, as the template for synthesis of

REPRODUCTIVE PATHWAYS OF VIRUSES

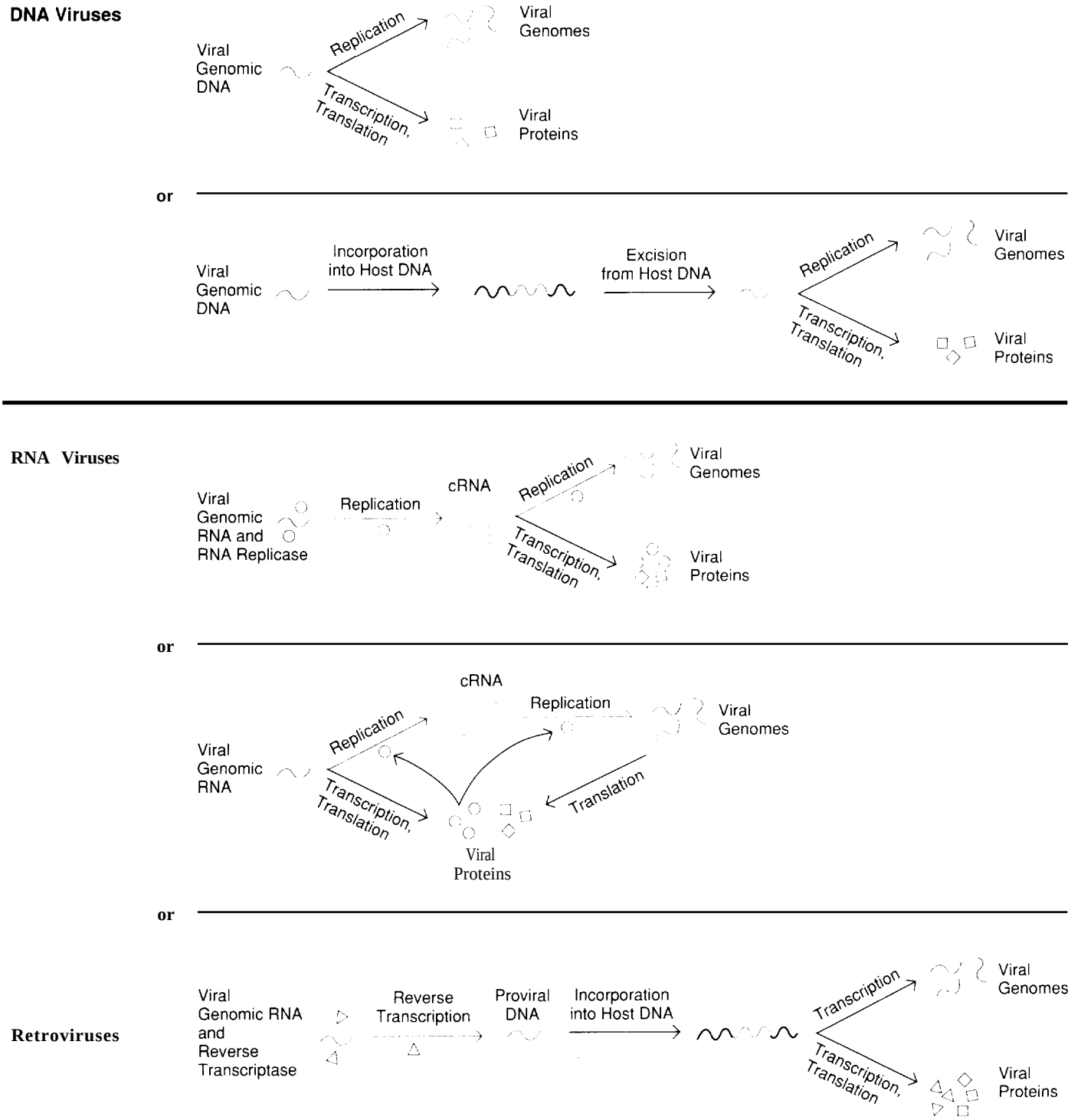


Fig. 1. Reproduction of a virus involves replication of its genome and synthesis of the proteins encoded therein. Shown here schematically are general features of the pathways by which those processes are carried out, Each biochemical reaction is catalyzed by an enzyme; however, only the virus-specific enzymes (those not supplied by the host cell) are listed. The squares represent all viral proteins other than RNA replicase and reverse transcriptase. For simplicity the viral and cellular genomes are assumed to be single-stranded.

viral proteins. (We assume here that the genomic RNA is single-stranded; double-stranded genomic RNA adds only minor complications.) Encoded in the genomes of both those types of RNA viruses is an enzyme, an RNA replicase, that catalyzes the synthesis of RNA from an RNA template. (Host cells cannot supply such an enzyme because they never replicate RNA.)

In the case of an RNA virus whose genomic RNA serves as messenger RNA, its RNA replicase is the first of the viral proteins to be synthesized from the template of the genomic RNA. The replicase catalyzes the synthesis first of RNA Complement to the genomic RNA and then of RNA complementary to the complement of the genomic RNA, that is, of RNA identical to the genomic RNA. Some of the replicas of the genomic RNA serve as genomes for daughter virions, and some serve as messenger RNA for further synthesis of viral proteins.

In the case of the second type of RNA virus, the complement of the genomic RNA, and not the genomic RNA itself, serves as messenger RNA. Therefore some RNA replicase is needed initially to synthesize the complement and allow synthesis of viral proteins, including the RNA replicase. The cycle is started by entry into the cell, along with the viral genome, of a few molecules of RNA replicase produced during the previous reproductive cycle. Those molecules catalyze the synthesis of the complement of the genomic RNA (which then serves as the template for synthesis of viral proteins) and as the template for synthesis of replicas of the viral genome.

The third type of RNA virus follows an entirely different reproductive pathway in which neither the genomic RNA nor its complement serves as the template for protein synthesis. Instead, the genomic RNA serves as the template for synthesis of DNA. Viruses of that type, known as retroviruses, are the only known exception to the "central dogma" of molecular genetics, which asserts that genetic information flows from DNA to RNA. The synthesized DNA, known as proviral DNA, is incorporated into that of the host cell and processed by the cellular machinery under control of viral regulatory mechanisms. Unlike the incorporated DNA of DNA viruses, the incorporated DNA of retroviruses is not excised from that of the host cell before processing.

HIV STRUCTURE

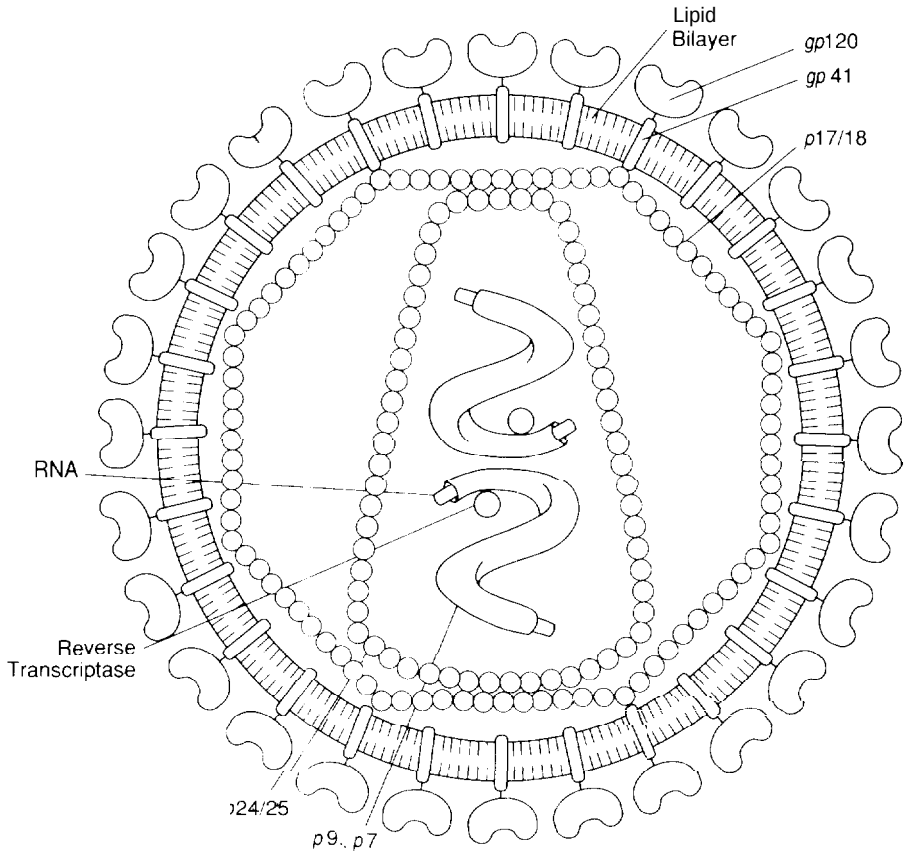


Fig. 2. The diplold genome of HIV, together with two molecules of reverse transcriptase, is housed within a capsid made up of many copies of the protein *p24/25.* The capsid itself is encased within an envelope made up of the glycoproteins *gp120* and *gp41* and a lipid bilayer derived from the membrane of the host cell. (Adapted, with permission, from a figure in the article "AIDS in 1988" by Robert C. Gailo and Luc Montagnier. *Scientific American,* October 1988.)

Since cells never synthesize DNA from an RNA template (the reverse of transcription), a retrovirus must have encoded in its genome an enzyme, a reverse transcriptase, for catalysis of that reaction. Furthermore, since the genomic RNA of a retrovirus is not translated into proteins, it must be accompanied into the cell by a few molecules of reverse transcriptase.

The various pathways for synthesis of viral proteins and replication of viral

GENETIC MAP OF HIV

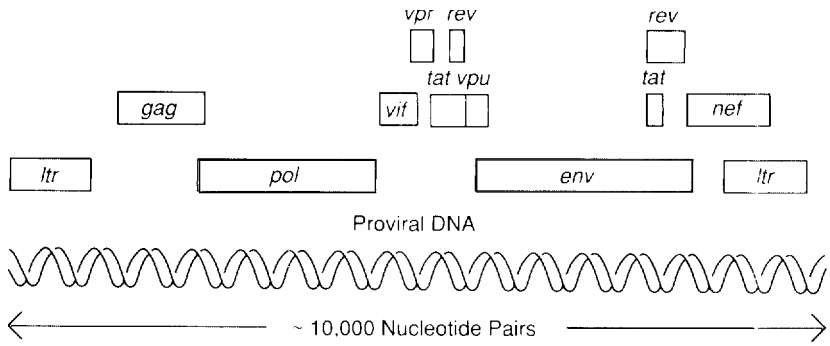


Fig. 3. The DNA of the HIV provirus includes two noncoding long terminal repeats *(ltrs)* that flank at least nine genes. Three are genes for viral components: *gag,* which encodes the proteins *p24/25* end *p9,p7; pol,* which encodes the enzyme reverse transcriptase; and *env,* which encodes the proteins *gp120* and *gp41.* The genes called *tat, rev, vif,* and *nef* encode biochemicals that regulate expression of the viral-component genes. Note that both the fat and rev genes consist of two seperate segments. The functions of *vpr* and *vpu* are not known. (Adapted, with permission, from a figure in "The Molecular Biology of the AIDS Virus" by William A. Haseltine and Flossie Wong-Staal. *Scientific American,* October 1988.

genomes are illustrated in Fig. 1.

The next step in the reproductive cycle of a virus is assembly of the newly synthesized components into daughter virions. That occurs by sequential stages of spontaneous aggregation involving formation of weak bonds, such as hydrogen bonds. Details of viral morphogenesis have provided insight into the development of more complex organisms.

The final step is escape of the new virions from the host cell. Some naked (non-enveloped) viruses escape by natural protein-secretion mechanisms of the cell, others by destroying the cell membrane with virus-specific proteins. Enveloped viruses escape—and become enveloped—by "budding," a process akin to the reverse of receptor-mediated endocytosis.

To conclude this primer on viruses, we present a few more details about retroviruses, particularly the human immunodeficiency virus (HIV), the cause of AIDS.

The unusual nature of retroviruses was not recognized until 1970, although some of the diseases they cause, such as the "swamp fever" that afflicts horses, had been known for many years. Some retroviruses cause cancers, others slowly degrade various physiological systems, and others apparently cause no disease. Only four human retroviruses have been identified, all within the past decade. Two cause rare and fatal cancers; the others are the two recognized types of HIV.

Like all retroviruses, HIV is enveloped and diploid (that is, its genome consists of two copies of its RNA "chromosome"). Figure 2 shows its structure and constituents. The reproductive cycle of HIV is basically that of any retrovirus, but its ability to regulate that cycle, through both positive and negative feedback mechanisms, is much greater than that of any other known retrovirus. The very rapid reproductive tempo that HIV can achieve is the basis for one mechanism by which HIV may kill infected T4 lymphocytes. (Reproduction of most retroviruses is not lethal to host cells.)

HIV has been, and continues to be, the object of intensive research. The nucleotide sequence of its proviral DNA (and hence of its genome) has been determined, and so have the locations of its genes along that sequence (Fig. 3). Numerous details about the biosynthetic pathways involved in HIV replication have been ascertained, and many more will be. Not only are such details necessary to develop drugs and vaccines to combat AIDS; they also exemplify the awesome complexity of even those not quite living organisms we call viruses. ■

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:

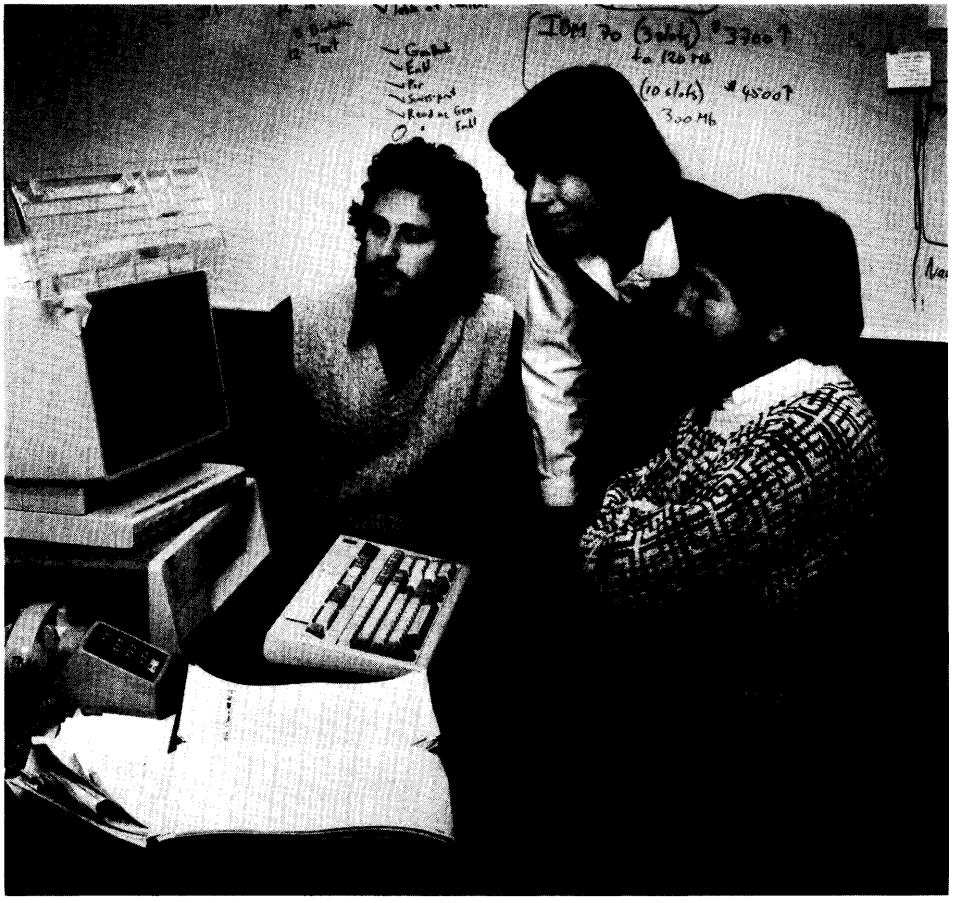
> From the initial publications of the nucleotide sequences of [HIV samples] in 1985, as well as from the early comparisons of restriction enzyme maps of many HIV isolates, it became evident that the HIV genome could exercise considerable heterogeneity. These early results encouraged further sequencing 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. ■



Gerald L. Myers (right) received a B.A. in zoology from the University of Colorado and, in 1969, a Ph.D. in biophysics from the University of Colorado Medical School. He was an American Cancer Society Postdoctoral Research Fellow at Yale University from 1969 to 1972 and then began teaching liberal arts (including classical Greek) at St. John College in Santa Fe. Since 1982 he has been a collaborator in the Theoretical Biology and Biophysics Group at the Laboratory. He has been the principal investigator for the HIV Sequence Database and Analysis Project from its inception in 1986. This year he is working on developing new user-interface strategles for the HIV project as a Guest Researcher at the National Institutes of Health's Lister Hill National Center for Biomedical Communications. **C. Randal Linder** (left), a Graduate Research Associate in the Theoretical Biology and Biophysics Group. received a B.A. from St. John's College and, in 1987, an M.S. in ecology and evolutionary biology from Cornell Univcrsity. **Kersti A. Maclnnes** (center), a Data Analyst in the Theoretical Biology and Biophysics Group, received a B.S. from St. Mary's College and, in 1977, an M.S. in biology from the University of Delaware.

Further Reading

Jean L. Marx. 1988. The AIDS virus can take on many guises. *Science* 241, 1039--1040.

W.-H. Li, M. Tanimura, and P. M. Sharp. 1988. Rates and dates of divergence between AIDS vlrus nucleotide sequences. *Molecular Biology and Evolution* 5, 313–330.

Temple F. Smith, Mira Marcus, and Gerald Myers. 1988. Phylogcnetic analysis of HIV-1 and HIV-2. In *Vaccines 88,* edited by Harold Ginsberg, Fred Brown, Richard A. Lerner, and Robert M. Chanock, 317–321. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory.

T. F. Smith, A. Srinivasan, G. Schochetman. M. Marcus, and G. Myers. 1988. The phylogenctic history of immunodeficiency viruses. *Nature* 333, 573–575.

S. Yokoyama, L. Chung, and T. Gojobori. 1988. Molecular evolution of the human immunodeficiency and related viruses. *Molecular Biology and Evolution* 5, 237–251.

R. F. Doolittle, D.-F. Feng, M. S. Johnson, and M. A. McClure. Origins and evolutionary relationships of retroviruses. To be published in the March 1989 issue of *The Quarterly Review of Biology.*

Scientific American October 1988. The entirety of the issue is devoted to AIDS and HIV.