Computational Tools to Battle HIV

Bette T. M. Korber and Alan S. Perelson

In 1981, AIDS, the acquired immune deficiency syndrome, was initially detected in a handful of gay men and reported in the Mortality and Morbidity Report of the Centers for Disease Control. Once the disease had been defined and described, its presence was rapidly recognized not only in the homosexual community in the United States but in other populations and countries. By 1983, the etiological, or causative, agent was isolated: a retrovirus[1] called human immunodeficiency virus, type I, or HIV-1. With that discovery, diagnostic tests soon became available, and by the mid-1980s, it was clear that there was a new pathogen on the loose in the world. It was soon recognized that a global epidemic of terrible magnitude was coming, although the extent of the devastation now experienced in sub-Saharan Africa was unimagined in those early days. The World Health Organization established an international network (www.unaids.org) to track this disease as it moved through populations, and through the years, it has maintained a grim record of HIV-1's death toll and spread. Current estimates indicate that more than 65 million people have contracted HIV-1, and 25 million of those are already dead. Africa has been the most brutally affected. Life expectancies in some of the worst afflicted nations are projected to drop as much as 20 years by 2010. Eurasia sits at the brink of the rapidly advancing storm. Russia, China, and India are facing burgeoning epidemics with no cure and no vaccine yet in hand.

HIV has three peculiarities that make it a particularly challenging foe: its latent period, its devastation of the immune system, and its variability. It has, on average, a decade-long latent period during which infected individuals carry the virus and can transmit the virus but are not overtly ill; in fact, unless they are tested, they are likely to be unaware of their infection. During this period, the immune system responds to the virus but is unable to clear it, and the ability to make immune responses to any infection slowly deteriorates. Combinations of antiretroviral therapies have been developed that can control viral replication and prolong life and good health, but so far, these drugs have not been able to clear the infection. Therapy is very expensive and, because of side effects, can be difficult to take for years on end.

During the infection, HIV infects and decimates $CD4^+$ T lymphocytes, the very cells that are central to the immune response needed to counter a viral infection. But the most serious impediment to defeating the virus is its extraordinary variability. This virus evolves during the course of every infection. The immune system responds and inhibits one form of the virus, but a new form inevitably escapes from that response. This cyclical response and evasion continues throughout the infection, and the human host ultimately loses the race. The virus can quickly acquire drug-resistant mutations, particularly when therapy is only partially successful, and the drug-resistant forms of the virus can be transmitted. This rapid within-host evolution results in extraordinary variation at the epidemic level, and viruses from any two individuals are quite distinct.

The Theoretical Biology and Biophysics Group at Los Alamos has played an integral part in understanding and defining the nature of both the host-viral dynamics and the evolution of HIV. This article covers the history of some of the ideas that were developed by Los Alamos scientists in collaboration with experimental and theoretical colleagues worldwide.

The HIV-1 Databases: An International Resource

Los Alamos was the original home of GenBank, a database of all publicly available genetic sequences from virtually all organisms. GenBank is now housed at the National Library of Medicine at the National Institutes of Health (NIH). During the mid-1980s, HIV sequences began to be archived in GenBank, arriving not only from the United States and Europe, but from Africa as well. Gerald Myers, who was working at Los Alamos at the time, was struck by the extraordinary diversity of the incoming HIV sequences. He realized, with great prescience, that given this level of diversity, creating a vaccine

[1] A retrovirus is a virus that stores its genetic code in RNA rather than DNA, and once in the host, copies its RNA into DNA. The viral DNA is then incorporated into the host DNA and read to make new virus particles.

was going to be a tremendous challenge and that the HIV research community would benefit from additional curatorial effort. He proposed to generate viral sequence alignments to enable comparisons of similarities between the incoming and archived sequences, to correlate the sequences with any information about their gene products, to link sequences to additional information (for example, the health status of patients and geographic and phenotypic information), and to publish an annual compendium. Researchers with new HIV sequences could then quickly put them into the context of a global framework without having to extract and organize the information each time. The NIH, persuaded by Myers, subsequently funded a program to create the HIV sequence database.

In the mid-1990s, Myers initiated databases for other pathogens, leaving the HIV project to Los Alamos researchers Bette Korber and Carla Kuiken. Sequencing technologies were rapidly improving, and HIV sequences were being generated at an accelerating pace. With a dedicated staff contributing both computer and biological expertise, the HIV sequence database developed into an easily used, Web-accessible, relational database, which now houses roughly 80,000 HIV sequences. Web search tools were added to retrieve data in new ways. For example, a researcher working on a vaccine for South Africa can pull up an automated alignment of all regional sequences, organized by year of isolation, by subtype, or by phenotype (see Figure 1.) Pioneering computational tools were developed for handling such basic problems as detecting viral recombination, hypermutation, and polymerase-chain-reaction contamination, tools that served to improve vastly the overall quality of the HIV scientific literature.

In 1995, an immunology database that contains information concerning HIV immune responses was created and added to the sequence database. It allowed a researcher to integrate immune-response data with sequence variability data. In 1997, an HIV drug-resistance mutation database was added, and in 2002, a vaccine-trial database became part of the collection. The latter was developed by Los Alamos postdoctoral fellow John Mokili. It summarizes and allows direct comparisons of data sets from hundreds of vaccine studies conducted in primates. The HIV database Web pages, with access to all four databases, received more than a million hits per month during the peak use period in 2002, averaging 35,137 pages accessed per day. The Los Alamos databases became an integral part of global HIV research efforts, providing a foundation for continuing scientific work.

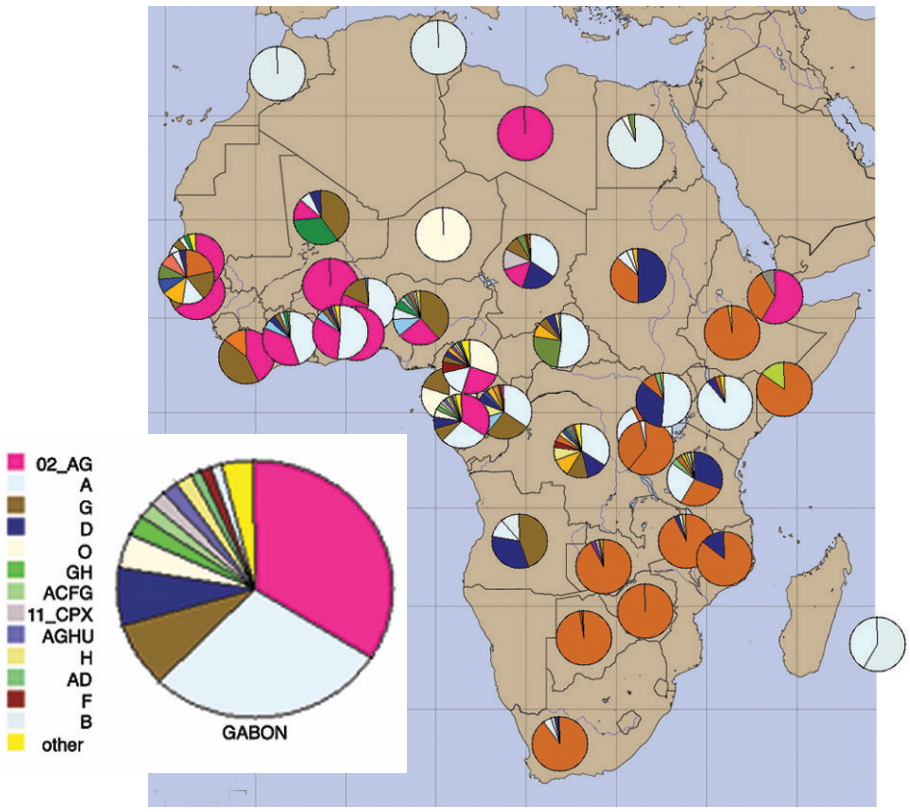

Figure 1. Tools to Combat HIV

Los Alamos is home to four databases that archive most of what is known about HIV and its associated immune responses, drug resistance, and vaccine trials. The database website at www.hiv.lanl.gov also contains numerous analysis tools. This figure is a screenshot from a geography tool that provides a bird's-eye view of viral diversity throughout a region, in this case the African continent. The pie chart indicates the distribution of viral strains for different subregions. Clicking on an individual pie chart on the Web brings up more information about the local viral strains. Clicking on a strain accesses the sequence information contained in the database.

HIV, the Shape Changer

The database was developed to be an international resource, but the integrated data it provides also serves as a basis for our own research efforts, and we have used it to study the evolution of the virus from many different perspectives. The exposed envelope protein of HIV (shown in Figure 2) is the most antibody-vulnerable part of the

virus. The envelope has two different protein components: gp41, which is locked into the outer membrane of the virus, and gp120, which is bound to gp41. The component gp120 projects from the surface of the virus, allowing HIV to recognize and bind to receptors on the human cell it infects. Gp120 is remarkably mutable, and sequences from different lineages, or subtypes, of HIV can differ in more than 30 percent of their amino acids.

But even the high fraction of amino-acid changes does not reflect HIV's true capacity to generate structural diversity. The virus has at least five mechanisms to alter its antigenic conformation—the molecular shape that is relevant to an immune response—and thus exhibits an extraordinary (and unique) capacity to avoid triggering the immune system. Following are the five mechanisms:

Mutational change through base substitution. This process typically comes to mind when one thinks about mutations—one nucleotide is replaced by another because the polymerase molecule made a mistake in copying genetic information. This process is the common mechanism—typically the only mechanism—by which viruses diversify and change. Substitutions occur at a high rate in HIV. The process is readily modeled and provides the basis for phylogenetic analysis of the evolutionary history of HIV.

Insertions and deletions. Hypervariable domains in HIV's genome lead to frequent alterations in the number of amino acids in the HIV-1 envelope proteins so that, within a few generations, the virus presents a new face to the immune system. The high frequency of insertion and deletion mutational events makes HIV-1 envelope proteins difficult to align and analyze, and the evolution of such regions cannot be modeled with currently available tools.

Shifts in glycosylation site pattern. The outer face of the HIV envelope protein is essentially covered by carbohydrates—that is, it is heavily glycosylated. (It is one of the most heavily glycosylated proteins in nature.) The carbohydrates are thought to partially shield HIV from antibody responses. The average gp120 molecule has 25 glycosylation sites, but the number varies between 18 and 33. The gain or loss of a carbohydrate moiety can alter the conformation of a protein and abrogate the ability of certain antibodies to bind.

Recombination. HIV carries two strands of RNA into newly infected cells. Because of frequent substitutions, insertions, and deletions, the two strands may be different. When they are copied, they can recombine, parts of one strand intermixing with parts of another. Researchers have found many examples of recombinant viruses that resulted from splicing together dramatically different strains of HIV. Undetected recombination can cause phylogenetic reconstruction to be inaccurate and evolutionary inferences to be incorrect.

Change at a distance. Mutational changes in gp41 can result in conformational changes in gp120 that can alter gp120 antibody binding sites.

Each of these mechanisms for change has been studied and tracked by the HIV database group, but it was base substitutions, the first kind of change, that allowed us to model the phylogenetic (or evolutionary) history of the virus. By ascertaining when HIV-1 entered the human population, we

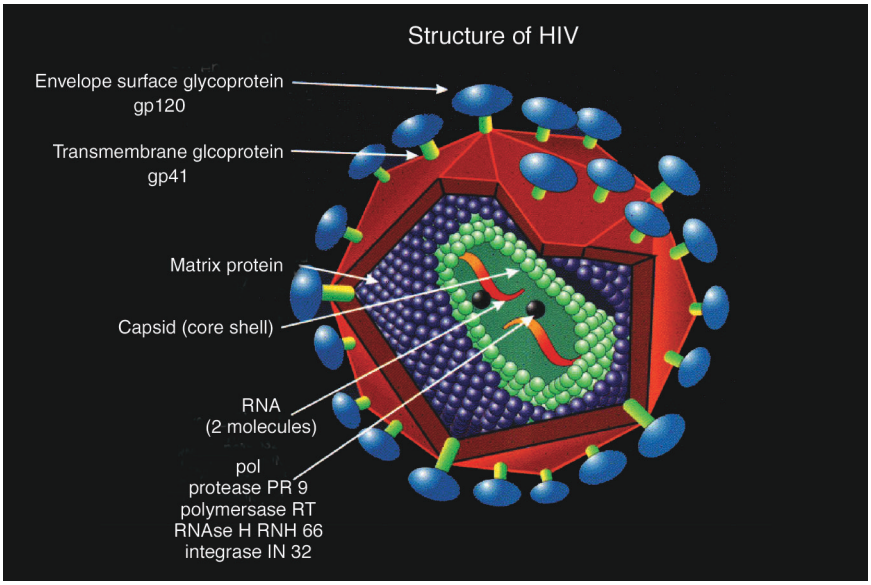

Figure 2. Human Immunodeficiency Virus, Type 1
This illustration depicts the essential features of HIV-1. The virus stores its genetic information in two strands of RNA. These lie within a protein layer known as the capsid, which is surrounded by the matrix protein. Also lying within the capsid is a metaprotein complex called pol, which contains four protein subunits: a reverse transcriptase, which will copy the RNA into DNA once the RNA has entered a host cell; an integrase, which inserts the viral DNA into the host cell's genome, an RNAse, and a protease that cleaves pol into the four subunits. Thus, upon entering a cell, pol literally cuts itself apart. The entire virus is encased in a lipid membrane that is actually a piece of the host's cell membrane. The two proteins gp41 and gp120 are the envelope protein of the virus. The highly exposed gp120 protein binds to CD4 receptors on the host cell as the first step toward infecting the cell.

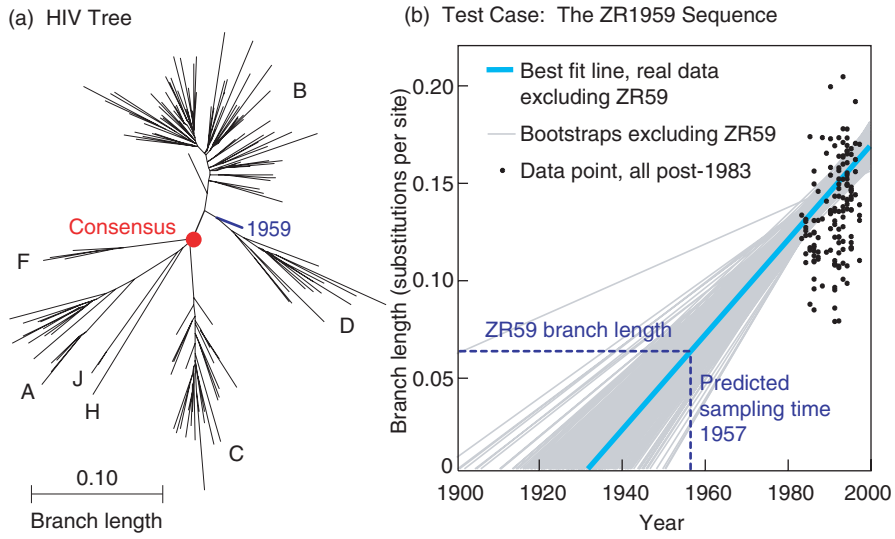

Figure 3. The Origin of HIV

(a) By aligning more than 150 HIV sequences, we constructed a phylogenetic tree that showed how HIV evolved. Each fanlike cluster of branches corresponds to an HIV subtype, and each terminating branch represents an HIV strain. The "root" of the tree (red dot) corresponds to the unobserved viral matriarch, the ancestral virus responsible for the strains now driving the AIDS pandemic. (b) The distance between the root and each branch (the branch length) is a measure of how many mutations occurred between the matriarch and each strain. All our HIV samples had a known sampling date, so given the tree and assuming a uniform rate of evolution, we could plot the branch length against the year of sampling and fit a line (turquoise) through the data points. Extrapolating the line backward indicates that the virus began to spread through the human population in 1930, give or take a decade. The gray lines are fits from a "bootstrap" method that was used to estimate the uncertainty. Our 1930 date was given support when we considered an "ancient" HIV sequence obtained in 1959. According to our tree, the sequence had a branch length of about 0.6, and according to our best fit, would have been spawned in 1957.

could estimate how rapidly the virus was diversifying and thus better understand the progression of the disease.

The oldest human blood sample documented to contain HIV-1 was taken in 1959 and came from an individual living in central Africa. It was sequenced in the late 1990s, and the sequence was analyzed at Los Alamos. This sample provided an "ancient" HIV sequence (relatively speaking), which allowed us to calibrate our evolutionary models. Drawing on a diverse set of Los Alamos scientists with interdisciplinary skills in computation, modeling, and statistics, and making use of the Los Alamos supercomputing facility to create optimized phylogenetic trees, we were able to estimate that the spread of HIV through the human population began in 1930, plus or minus a decade (see Figure 3).

Further modeling suggested a very slow initial spread of the virus, possibly indicative of a time when it was confined to rural areas with limited transmission possibilities. These estimates enabled better understanding of the rate of diversification of HIV and of how long HIV took to get to the present level of diversity. They helped rule out some controversial theories about the origins of HIV, and they moved us closer to understanding the history of this virus in its human host, a vital topic given how this epidemic has changed the landscape of human history in the 20 years since its discovery.

Another area of study has been an ongoing effort to understand how the human immune response influences viral variation in populations. Our cells routinely chop up internal proteins into short amino-acid segments and ferry those segments to the surface, where they are "examined" by a prowling immune-system cell. The immune system is triggered when the segment is deemed foreign to the body, but it is well known that certain amino acids are less likely to trigger an immune-system response. Mining the immunology literature, we were able to identify the parts of the virus that get chopped up—so-called antigenic regions—that are the focus of the immune responses in many individuals. An algorithm then allowed us to predict successfully where antigenic regions would be concentrated in less well studied HIV proteins. In the most variable regions of these viral proteins, we found a significant enrichment of certain amino acids, indicating that HIV had evolved to make itself less vulnerable to attack. Thus, human immune responses have left a clear imprint on the evolution of the virus.

The virus also may be leaving its evolutionary imprint on us. Certain human genes can influence our susceptibility to infection and our ability to live with HIV. In collaboration with the Santa Fe Institute and members of our theoretical group, we have been working to understand and define these genes. In populations where the virus is highly prevalent, such advantages may shift the human population in favor of those carrying such genes.

Finally, we have designed artificial consensus sequences (or reconstructed ancestor sequences) that are more similar to circulating strains of the virus than the various strains are to each other. Proteins from these artificial sequences are now widely used by experimentalists to probe and study the T cell immune responses of HIV-infected individuals. Our hope for these artificial proteins is that they will be more likely than

natural strains to elicit cross-reactive immune responses if used as a vaccine. They effectively reduce by half the number of amino-acid differences between a vaccine candidate strain and circulating viruses. Our colleagues at the University of Alabama and Duke University currently have experiments under way to test the immunological cross-reactivity of these proteins, and the initial results are encouraging. This kind of vaccine design strategy could be used in conjunction with other strategies—for example, those that deliberately expose immunologically vulnerable parts of the envelope that are usually hidden. Eventually, the result could be production of vaccines with better potential to protect individuals against infection from the extraordinarily diverse pool of circulating viruses.

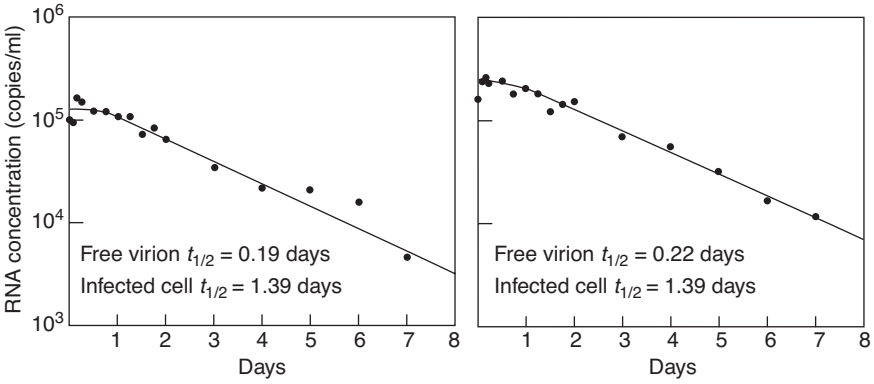

Figure 4. HIV Dynamics in the Presence of an Inhibitory Drug
This figure shows a fit of Equation (3) in the text to data from two representative patients. The circles show the concentration of HIV-1 RNA after Ritonavir treatment has begun on day 0. The theoretical curve was obtained by a nonlinear least-squares fit. From the fit, we could estimate the rate of clearance of the virus, the rate of loss of infected cells, and the initial viral load, that is, parameters c, δ, and V_0, respectively, in Equation (3).

A Model of HIV Dynamics

Developing a successful HIV vaccine is our ultimate goal, but in the meantime, we have made enormous progress in learning how to fight the virus with drugs. In pioneering studies conducted by Alan Perelson, in collaboration with David Ho's group at the Aaron Diamond AIDS Research Foundation, Rockefeller University, we addressed the question of whether the average 10-year time from HIV infection to AIDS reflects the fact that HIV grew slowly, and therefore did not need aggressive treatment.

By studying chronically HIV-infected individuals whose HIV concentrations in blood were relatively constant, we showed that giving a drug called Ritonavir[2], a potent inhibitor of HIV-1 protease, caused the HIV concentration in blood to drop about 100-fold in two weeks. We then showed that this rapid decline implied that the body was rapidly clearing HIV. With this and other experiments, we were able to estimate that the time to clear half of the HIV in blood (the half-life) was about one hour or less. But we also knew that infected T cells were constantly producing HIV. The virus' short half-life therefore implied that the T cells also had to die rapidly. By more detailed experiments and analysis, we were able to estimate that the $CD4^+$ T cells that were the major producers of HIV lived only about one day while producing HIV. The question of whether HIV kills these cells directly or the immune response plays some significant role is still being debated. Nevertheless, this work showed that HIV was being maintained in the body by a vicious cycle: the virus was being rapidly produced and cleared, and while it was present, it was infecting cells and killing them.

The analysis involved developing models of viral infection and the effect of treatment. To illustrate the approach, consider the following simple model of viral infection:

$$\begin{aligned} \frac{dT}{dt} &= s - \alpha T - kVT \\ \frac{dI}{dt} &= kVT - \delta I \\ \frac{dV}{dt} &= N\delta I - cV \end{aligned} \qquad (1)$$

where T is the concentration of uninfected cells, I the concentration of infected cells, and V is the concentration of virus particles or virions. Here, we assume uninfected cells are created from a source at rate s and die at a per capita rate α. In addition, we assume they are infected by the virus with a rate constant k. Infected cells are assumed to die at the *per capita* rate δ and to release a total of N viral particles during their lives. We also assume that the body clears the virus by a first-order process, with a constant rate of clearance per virion given by c. For simplicity, the loss of virus upon infecting cells is neglected, although it can be included. When drug therapy with a protease inhibitor is given, a fraction ε of newly formed virus particles is noninfectious. Thus, in the presence of the drug, the model equations become:

[2]Ritonavir goes by the tradename Norvir and is manufactured by Abbot Laboratories.

$$\begin{aligned}\frac{dT}{dt} &= s - \alpha T - kV_I T\\ \frac{dI}{dt} &= kV_I T - \delta I\\ \frac{dV_I}{dt} &= (1-\varepsilon)N\delta I - cV_I\\ \frac{dV_{NI}}{dt} &= \varepsilon N\delta I - cV_{NI}\end{aligned} \qquad (2)$$

where V_I and V_{NI} denote infectious and noninfectious virus, respectively. If one analyzes patient data obtained over the first week of therapy, T does not change greatly and can be assumed to be constant. Under this approximation, the system of equations becomes linear and can be solved exactly.

Two other approximations were made to allow us to gain insight into the solutions and to make comparisons with patient data. First, measurements of the total amount of virus in blood showed that, in most chronically infected patients, V was approximately constant over periods of about weeks or months. Thus, before drug therapy, a steady state was assumed, which implied that $c = NkT_0$, where T_0 was the measured level of T cells before therapy. Second, the drug efficacy was assumed to be 100 percent, that is, $\varepsilon = 1$. Under these considerations, one could show that

$$V(t) = V_0 e^{-ct} + \frac{cV_0}{c-\delta} \times \left[\frac{c}{c-\delta}\left\{e^{-\delta t} - e^{-ct}\right\} - \delta t e^{-ct}\right] \qquad (3)$$

where T is time on therapy and $V(0) = V_0$. Using nonlinear regression techniques, we tried to fit this model to measured values of V and obtained estimates of the parameters c and δ (see Figure 4). Further, we could show that if $\varepsilon < 1$, then these estimates of the clearance rate of HIV and the *per capita* death rate of T cells were minimal estimates (and thus infected cells and virus might be cleared even faster than estimated by this method). We could also show numerically that, if T changed by the amounts observed in patients, then the estimates would vary by only a few percent. Hence, by this method, the first estimates of the *in vivo* half-lives of HIV and infected cells were obtained, and they showed in a stunning manner that HIV infection was highly dynamic. From estimates of the rate of growth of the virus needed to maintain the observed constant levels of virus in the face of the estimated clearance, we showed that HIV would mutate sufficiently to become resistant to any single drug—an even more important finding. This determination, along with the observation that drug therapy could rapidly decrease the viral load, helped usher in the age of combination drug therapy for the treatment of AIDS. ■

Further Reading

Eberstadt, N. 2002. The Future of AIDS. *Foreign Aff.* **81** (6): 22.

Gaschen, B., J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang et al. 2002. Diversity Considerations in HIV-1 Vaccine Selection. *Science* **296** (5577): 2354.

Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes et al. 2000. Timing the Ancestor of the HIV-1 Pandemic Strains. *Science* **288** (5476): 1789.

The Los Alamos HIV sequence, immunology, vaccine, and drug-resistance databases. http://www.hiv.lanl.gov.

Perelson, A. S. 2002. Modelling Viral and Immune System Dynamics. *Nat. Rev. Immunol.* **2**: 28.

Perelson, A. S., A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho. 1996. HIV-1 Dynamics in Vivo: Virion Clearance Rate, Infected Cell Life-Span, and Viral Generation Time. *Science* **271** (5255): 1582.

Perelson, A. S., P. Essunger, Y. Cao, M. Vesanen, A. Hurley, K. Saksela et al. 1997. Decay Characteristics of HIV-1-Infected Compartments During Combination Therapy. *Nature* **387** (6629): 188.

The UNAIDS WHO site concerning the global status of HIV/AIDS. http://www.unaids.org.

Walker, B. D., and B. T. Korber. 2001. Immune Control of HIV: The Obstacles of HLA and Viral Diversity. *Nat. Immunol.* **2**: 473.

Bette Korber received her Ph.D. in immunology from the California Institute of Technology and went on to do postdoctoral work at Harvard University, studying the molecular epidemiology of retroviruses. She came to work at Los Alamos in 1990 as a Director's-funded postdoctoral fellow and for the past few years has managed the HIV immunology and sequence databases. Her research effort has focused on analysis of HIV sequences. Bette is a Pediatric AIDS Foundation Elizabeth Glaser Scientist, has been an advisor to the World Health Organization and UNAIDS on issues concerning HIV variability and vaccine design, and has collaborated with HIV scientists throughout the world.

Alan Perelson graduated from the Massachusetts Institute of Technology in 1967 with bachelor's degrees in life sciences, electrical engineering, and computer science. He received a Ph.D. in biophysics from the University of California at Berkeley in 1972. In 1974, Alan became a staff member in the Theoretical Division at Los Alamos. He left the Laboratory briefly to be an assistant professor of medicine at Brown University. Alan returned to Los Alamos, where he became a Laboratory fellow in 1991, group leader of Theoretical Biology and Biophysics from 1995 to 2001, and senior fellow in 2002. He is a member of the American Academy of Arts and Sciences and the recipient of a Merit Award from the National Institutes of Health. Alan's research has focused on developing models of the immune system and infectious disease. He has helped establish the field of theoretical immunology and, more recently, a field called viral dynamics.