Recent Evolutionary History of Human Immunodeficiency Virus Type 1 Subtype B: Reconstruction of Epidemic Onset Based on Sequence Distances to the Common Ancestor

Vladimir V. Lukashov, Jaap Goudsmit

Department of Human Retrovirology, Academic Medical Center, University of Amsterdam, Amsterdam, Meibergdreef 15, 1105 AZ The Netherlands

Received: 26 October 2001 / Accepted: 8 November 2001

Abstract. We obtained and studied HIV-1 sequences with a known sampling year from three outbreaks of the HIV-1 epidemic: 141 env V3 (270 nt) sampled between 1984 and 1992 and 117 pol prot/RT (804 nt) sequences sampled between 1986 and 1999 from Dutch homosexual men and injecting drug users (IDUs), as well as 77 env V3 sequences sampled between 1983 and 1994 in the United States. Since retrospective serological and/or epidemiological data on these populations are available, providing estimates of the dates of the onset of the HIV-1 epidemics, we had the opportunity to test different phylogenetic models for their accuracy in deriving the recent evolutionary history of HIV-1 subtype B and the onset date of the HIV-1 epidemic. We observed that, in any given year, individual sequences vary widely in their distances to the common ancestor, and sequences close to the ancestors were found decades after the onset of the epidemic. Nevertheless, the mean evolutionary distances of virus strains to ancestors were increasing significantly during the course of the studied epidemics, which indicates that the molecular clock is operational in the recent evolution of HIV-1. When the relationship between the sampling years of sequences and their nucleotide distances to the common ancestor was extrapolated to the past, analysis of pol sequences provided accurate estimates of the onset years of the epidemics, whereas analysis of V3 sequences by the maximum-likelihood or neighbor-joining methods led to an overestimation of the

age of the epidemics. Separate analysis of nonsynonymous and synonymous distances revealed that this overestimation results from nonsynonymous substitutions, whose numbers were not increasing significantly in all three virus populations over the observation period. In contrast, analysis of synonymous env V3 distances provided accurate estimates of the onset years for the outbreaks we studied.

Key words: HIV-1 — AIDS epidemic — Molecular history

Introduction

The molecular clock hypothesis is widely used to reconstruct phylogenetic relationships among organisms and to estimate dates of species divergence (Dickerson 1971). It assumes that the rate of molecular evolution of a given gene or gene region is approximately constant over time and equal in all descendent lineages (Kimura 1983). When the rate of accumulation of nucleotide substitutions—the evolution rate—for a certain nucleotide sequence has been experimentally established, the hypothesis offers the opportunity to date the divergence of two related sequences from their common root.

The concept of a molecular clock has been employed to estimate the date of divergence for DNA viruses such as the members of the hepadnavirus family (Orto et al. 1989) and mammalian herpesviruses (McGeoch et al. 1995) as well as RNA viruses such as influenza A viruses

(Buonagurio et al. 1986; Fitch et al. 1991), coronaviruses (Sanchez et al. 1992), flaviviruses (de A. Zannotto et al. 1996), picornaviruses (Villaverde et al. 1991), and retroviruses (Querat et al. 1990), including the human and simian immunodeficiency viruses (HIV and SIV, respectively) (Gojobori and Moriyama 1990; Gojobori et al. 1990; Eigen et al. 1990; Li et al. 1988; Sharp et al. 1988; Yokoyama et al. 1988; Salemi et al. 2000). Although the concept of a molecular clock is widely used, it is not universally accepted and applicable. The clock's existence in general is the subject of debate, and the concept is often applied to a specific virus group without proof of its clock-like evolution. So far, clock-like evolution has been demonstrated based on longitudinally collected sequence information only for influenza A viruses (Buonagurio et al. 1986; Fitch et al. 1991). On the other hand, the evolution of the vesicular stomatitis viruses has been shown to be inconsistent with the molecular clock hypothesis (Nichol et al. 1993; Gillespie 1993).

Several groups have applied the hypothesis to trace back the origin of HIV. The separation between the two types of HIV (HIV-1 and HIV-2) is estimated to have occurred between 140 and 280 years ago (Gojobori and Moriyama 1990; Gojobori et al. 1990; Eigen et al. 1990; Li et al. 1988; Sharp et al. 1988; Yokoyama et al. 1988; Salemi et al. 2001). The date of the subsequent event in the HIV-1 history—the onset of virus diversification within HIV-1 group M and the separation of HIV-1 genetic subtype lineages—has recently been estimated to be in the 1910s or 1930s, with 95% confidence intervals of several decades (Goudsmit and Lukashov 1999; Korber et al. 1998, 2000; Salemi et al. 2001). The above calculations, which were based on the analysis of evolutionary distances of contemporary HIV-1 sequences to the group M common node, do not support an earlier statement that HIV-1 group M viruses probably shared a common ancestor “in the 1940s or the early 1950s” (Zhu et al. 1998), which was based on an HIV-1 sequence sampled in 1959. Taken together, these estimates suggest that the separation within HIV-1 group M occurred decades before the onset of the global HIV-1 epidemic.

Another pitfall of the molecular clock approach is related to clock calibration. Based on the premise of evolution rate constancy, the approach requires exact estimates of the evolution rates to be used in calculations. For many viruses, including HIV, the evolution rate is highly variable over time within a lineage as well as among distinct lineages and different genome regions (Li et al. 1988; Lukashov et al. 1995a).

Phylogenetic methods are subject to various biases. In particular, the reconstruction of the nodal (ancestral) sequence is error-prone and depends on the nucleotide substitution model invoked, the distances of individual sequences to the ancestor may be not independent, etc. It has been shown that phylogenetic methods differ in their reliability in reconstructing a known HIV-1 transmission

chain (Leitner et al. 1999). In the present study, we attempted to test various existing phylogenetic methods for their accuracy in deriving the recent evolutionary history of HIV-1 subtype B and the onset date of the global epidemic. For this purpose, we collected and analyzed longitudinal sequence data representing three outbreaks of the global HIV-1 epidemic: (i) in the United States and among (ii) homosexual men and (iii) injecting drug users (IDUs) in The Netherlands. Since the onset years of these outbreaks had been established based on retrospective epidemiological and/or serological data, we could determine which phylogenetic model and which genomic region of HIV would yield the most reliable estimates. Our analysis was based on two genomic regions of HIV-1—a fast-evolving env gp120 V3 region [270 nucleotides (nt)] and a much more conserved region of the pol gene (protease/partial reverse transcriptase; 804 nt)—for which we collected extensive longitudinal sequence data with known date of sampling.

The concept of a molecular clock predicts that nucleotide substitutions are accumulating within individual HIV-1 sequences over time and that the evolutionary distances of individual HIV-1 genomes to their common ancestor are increasing in a linear fashion over the course of the AIDS epidemic (Kuiken et al. 1996a). The mean distance to the group's common ancestor increases over time according to the formula $D_T = r(T - T_0)$ [formula (1)], where D_T is the mean evolutionary distance to the common ancestor of the individual lineages in the population, r is the mean evolution rate of the individual lineages in this population, T_0 is the year when the group of virus lineages started to diverge from the common ancestor, and T is the year of individual sequence sampling.

The analysis of evolutionary distances to the common ancestor of individual HIV-1 sequences, in relation to their sampling years, opens the way to check whether the mean evolutionary distances to the common ancestor are indeed increasing in a linear fashion according to the proposed formula (1) and, if so, to identify the onset of virus diversification within a given population. A positive linear relationship between the sampling year of the individual sequences in an HIV-1 population and their evolutionary distances to the common ancestor of that population according to formula (1) allows us to put a time scale on this relationship, which can be extrapolated back to determine the diversification of the HIV-1 population ($D_T = 0$ when $T = T_0$). If the HIV-1 epidemic in a given human population started after the expansion of a single HIV-1 strain (the heterogeneity of the founder virus population was effectively equal to 0), the date of the onset of that epidemic can then be determined. If this condition is met, which is likely the case for the U.S. and Dutch drug users (Grmek 1990; Kuiken and Goudsmit 1994; Kuiken et al. 1996b), the most recent common ancestor of an epidemic coincides with the onset of this

Fig. 1. Phylogenetic tree demonstrating relations among HIV-1 subtypes in the env gp120 V3 region. For HIV-1 subtype B, 141 sequences are included. Indexes: homosexual men from The Netherlands (H); IDUs from The Netherlands (I), Thailand (TH), and Brazil (BR). The rest of the sequences are from the United States. For all subtype B sequences, sampling years are shown. For HIV-1 subtypes other than B, two reference sequences provided by the Los Alamos database

(<http://hiv-web.lanl.gov>) were included. The branching position of two group O sequences (MVP5180 and ANT70) is shown. The tree was built using the NJ method as implemented in the PHYLIP package (NEIGHBOR) based on gamma distances for the Jukes–Cantor method. The most recent common node of HIV-1 subtype B is shown (filled circle).

epidemic. If this condition is not met, and there have been several virus introductions into a particular risk group, which is likely the case for Dutch homosexual men (Kuiken and Goudsmit 1994; Kuiken et al. 1996b),

the onset of the epidemic would be later than the date of the most recent common ancestor. Therefore, the date calculated for the onset of the epidemic must generally be considered the earliest possible.

Materials and Methods

HIV-1 Sequences

For this study, we collected HIV-1 subtype B env and pol sequences, for which sampling years were known.

Sequences coding for the gp120 env V3 region were 270 nt long (de Wolf et al. 1994). The majority of env sequences was generated by us from homosexual men and injecting drug users (IDUs) participating in Amsterdam cohort studies (Lukashov et al. 1995a; Lukashov and Goudsmit 1997), the Baltimore cohort of IDUs (Lukashov et al. 1996a), individuals LAI and MN (NM) (Lukashov and Goudsmit 1995), the WHO UNAIDS network study (de Wolf et al. 1994), Russia (Lukashov et al. 1995b), and HIV-1 infected women in The Netherlands (Lukashov et al. 1996b). The sequences are available through GenBank (for accession numbers, see the original publications).

The pol sequences were 804 nt long and contained the full-length protease gene (294 nt) as well as the first 510 nt of the reverse transcriptase (RT) gene. Virtually all pol sequences were generated at our department from the participants of the Amsterdam cohorts and from HIV-infected individuals attending the Academic Medical Center in Amsterdam. To avoid the influence of drug-resistant mutations on the results of calculations, we used only wild-type pol sequences (without drug-resistant mutations).

In addition, we used both env and pol sequences available from the Los Alamos HIV-1 database (Kuiken et al. 1999). The description of the sequence sets, their origin, and references are provided in the Results section.

Phylogenetic Methods and Evolutionary Distance Calculations

For the U.S. epidemic and the two epidemics in The Netherlands, nucleotide sequences were aligned manually. Positions containing an alignment gap were excluded from pairwise sequence comparisons. To reconstruct the most recent common ancestor, or the common node, for sequences within each sequence set, several phylogenetic methods were used: the maximum-likelihood (ML) method for (i) nucleotide, (ii) nonsynonymous, and (iii) synonymous substitutions and neighbor-joining (NJ) method for (iv) nucleotide, (v) nonsynonymous, and (vi) synonymous substitutions. The ML and NJ methods for nucleotide distances were used as implemented in the PHYLIP package (DNAML and NEIGHBOR, respectively). The DNAML method was based on empirically found base frequencies and transition/transversion ratios, considering different rates of evolution at different positions. The NJ method was based on gamma distances for the Jukes-Cantor method. Our analysis of synonymous (D_s) and nonsynonymous (D_n) substitutions by the ML method was performed using the PAML codeml program available at <http://abacus.gene.ucl.ac.uk/software/paml.html> (models M1 and M2, respectively) (Yang et al. 2000). The analysis of D_s and D_n by the NJ method was performed using the MEGA package (Kumar et al. 1993) based on the Nei-Gojobori method with Jukes-Cantor correction. For each of the methods, reference sequences of HIV-1 subtypes other than B, provided by the Los Alamos database (<http://hiv-web.lanl.gov>), were used to root the trees. For each method, bootstrap resampling was performed (100 replications), resulting in a bootstrap support for the subtype B cluster of at least 80. For each distance estimation method, the evolutionary distances of individual sequences within each set to the most recent common ancestor of the set (Fig. 1) were calculated and analyzed using several statistical approaches.

Statistical Methods

All statistical calculations were done using the SPSS/PC+ software (version 5.0; SPSS Inc., Chicago, IL). For each sequence set, the relationship (correlation) between the sampling years of individual HIV-1 sequences and their distances to the most recent common ancestor was examined using linear regression analysis. Each sequence was considered to be statistically independent. To prove that the linear model was correct, a test for linearity was performed. No relationship was observed between the predicted and the residual values ($r = 0.000$), indicating that the assumptions of linearity and homogeneity of variance were met. In addition to the linear model, we used several models which included nonlinear members (the logarithmic and exponential) to check whether they could better describe the relations between the sequence sampling years and the evolutionary distances to the ancestor. Our attempts to fit the data by nonlinear models resulted in nonlinear members of the equation approaching 0.

For each sequence set, the data were analyzed using linear regression analysis in three ways. First, the distances of all individual sequences to the common node were analyzed in relation to their sampling year, with all sequences equally contributing to the analysis. Second, the weighted least-squares method was used. Third, the mean distances per year were analyzed, with all years equally contributing to the analysis. For each sequence set, 95% confidence intervals for the time of virus introduction into the respective human population were calculated (CI for the X intercept).

Results

The Accuracy of Various Phylogenetic Methods in Reconstructing the Molecular History of the U.S. HIV-1 Epidemic

It is believed that the worldwide HIV-1 subtype B epidemic started after the introduction of a single HIV-1 subtype B strain into the United States by patient 0. To analyze the US epidemic, we collected 77 HIV-1 V3 sequences sampled in the United States in 1983–1994, including sequences ASP1, ALA1, CDC42, JRCSF, NY5CG, RF, RJS, SBC, SC, and SF33 (Kuiken et al. 1999), sequences from the Florida dentist case (1990–1991) (Ou et al. 1992), a set of the early U.S. HIV-1 sequences (1983) (Lukashov and Goudsmit 1995), and the sequences from the Baltimore IDU cohort (1989–1994) (Lukashov et al. 1996a). The most recent common ancestor of the U.S. HIV-1 epidemic was reconstructed by the ML and NJ phylogenetic methods.

A linear regression analysis of the nucleotide distances of the U.S. sequences to their reconstructed common ancestor by either the NJ or the ML method revealed a significant positive correlation between the sampling years of the U.S. V3 sequences and their nucleotide distances to the common ancestor (Figs. 2a and b, respectively). There was a strong correlation between the nucleotide distances of individual sequences to the common root estimated by both methods (correlation, 0.81; $p < 0.001$). The years of the onset of the HIV-1 epidemic in the United States calculated by the NJ and ML methods based on nucleotide distances, were 1953 and 1967, respectively.

Fig. 2. Evolutionary distances of the U.S. V3 sequences to the reconstructed common ancestor of the U.S. HIV-1 epidemic in relation to sequence sampling years, according to several phylogenetic methods: (a) the NJ method based on nucleotide distances; (b) the ML method based on nucleotide distances; (c) the NJ method based on nonsynonymous distances; (d, e) the NJ method based on synonymous distances of all individual sequences (d) and mean synonymous distances per sampling year (e); (f, g) the ML method based on nonsynonymous (f) and synonymous (g) distances. Confidence intervals for the estimated date of the onset of the U.S. HIV-1 epidemic are calculated based on

synonymous distances of all individual sequences and shown in d and e, according to the NJ method and in g (next page), according to the ML method. Regression lines are shown. (a) $r = 0.27$; $r^2 = 0.07$; slope, 0.0021; X intercept, 1953; $p = 0.02$. (b) $r = 0.41$; $r^2 = 0.17$; slope, 0.0037; X intercept, 1967; $p < 0.001$. (c) $r = 0.14$; $r^2 = 0.02$; slope, 0.0012; X intercept, 1925; $p > 0.1$. (d) $r = 0.48$; $r^2 = 0.23$; slope, 0.0043; X intercept, 1976; $p < 0.00001$. (e) $r = 0.98$; $r^2 = 0.97$; slope, 0.0047; X intercept, 1976; $p < 0.000001$. (f) $r = 0.11$; $r^2 = 0.01$; slope, 0.0015; X intercept, 1950; $p = 0.32$. (g) $r = 0.32$; $r^2 = 0.10$; slope, 0.0058; X intercept, 1974; $p < 0.01$.

Fig. 2. Continued.

These results are identical to those reported by Korber et al. (2000), who estimated the year of HIV-1 introduction in the United States to be 1954 or 1967, based on the analysis of the full-length *env* gene and depending on the phylogenetic method used. However, these estimates are much earlier than those established based on epidemiological data (Jaffe et al. 1983; Selik et al. 1984). A clinical syndrome of AIDS was originally recognized in 1981, and earlier cases of AIDS in the United States have not been retrospectively identified before 1978 (Jaffe et al. 1983; Selik et al. 1984). Since HIV-1 infection has a minimal asymptomatic period of only a few years before AIDS development, the HIV-1 epidemic in the United States is unlikely to have started in the 1950s or 1960s.

To analyze whether characteristics of the evolution of the *env* region can result in overestimation of the age of the epidemic, we subsequently studied the synonymous and nonsynonymous distances of the U.S. V3 sequences to the reconstructed ancestor. Our analysis of the distances of all individual sequences to the common node (all sequences equally contributing to the analysis) using either the NJ or the ML methods revealed a highly significant positive correlation between the sequence sampling years and their synonymous distances to the common ancestor ($p < 0.00001$), whereas their nonsynonymous distances to the common ancestor were not significantly increasing ($p > 0.1$) during the observation period (Figs. 2d and g versus 2c and f, respectively). Extrapolation of the regression line of synonymous distances back to the date when no synonymous heterogeneity was present in the HIV-1 population in the United States allowed us to estimate 1974 (ML)–1976 (NJ) (95% CI, 1972–1977 for the NJ method; 1969–1978 for the ML method) as the onset date of the epidemic in the United States, which accords well with epidemiological data (Jaffe et al. 1983; Selik et al. 1984). Similarly, when the mean synonymous distances of all U.S. sequences sampled in each year were calculated and regression

analysis was based on the mean distances per year (Fig. 2e), a high correlation ($r = 0.98$, $r^2 = 0.96$, $p < 0.0001$) was revealed between the mean synonymous distances and the sequence sampling years.

The Origin of HIV-1 Epidemics Among Homosexual Men and Injecting Drug Users in The Netherlands

Homosexual men and IDUs in The Netherlands harbor distinct HIV-1 subtype B virus populations (Kuiken and Goudsmit 1994; Kuiken et al. 1996b), which form separate phylogenetic clusters. The HIV-1 epidemics among Dutch homosexual men and IDUs differ in origin, with the epidemic among Dutch IDUs being linked to that among U.S. IDUs (Lukashov et al. 1996a). Based on intensive retrospective serological studies, the onsets of the HIV-1 epidemics among Dutch homosexual men and IDUs have been dated 1977 and 1980, respectively (van Haastrecht et al. 1992).

We analyzed V3 sequences obtained at seroconversion or from the first HIV-1-positive sample from participants of the Amsterdam prospective cohorts of homosexual men ($n = 96$) and IDUs ($n = 45$) between 1984 and 1992; one sequence per individual was obtained. Additionally, we analyzed *pol* sequences obtained from the participants of the Amsterdam prospective cohorts of homosexual men ($n = 87$) and IDUs ($n = 30$) between 1986 and 1999, one sequence per individual. The most recent common ancestors for the two risk groups were calculated for the *env* and *pol* sequence sets separately, using the NJ and ML methods based on nucleotide distances as well as the NJ method based on synonymous and nonsynonymous distances. The distances of individual HIV-1 sequences as well as the mean distances per year to the common ancestor of the same risk group were analyzed in relation to sampling year (Figs. 3–5).

For the *env* V3 region, similar to our findings for the U.S. epidemic, we observed in both risk groups a significant correlation between the sampling year of the sequences and their nucleotide distances to the common root of their respective virus population, according to both the ML and the NJ phylogenetic methods (data not shown). Again, when based on nucleotide distances, the age of the HIV-1 epidemics was strongly overestimated for the Dutch homosexual population (1955 and 1967) and IDUs (1965 and 1975), according to the NJ and ML methods, respectively. Separate analysis of nonsynonymous and synonymous distances revealed that such overestimation results from nonsynonymous distances, which were not increasing significantly in either virus population over the observation period irrespective of whether the NJ or the ML method was used (data not shown). Nevertheless, for both risk groups and both the NJ and the ML methods, a significant correlation was observed between sequence sampling year and synonymous dis-

Fig. 3. Estimation of the onset years of the HIV-1 epidemics among homosexual men and IDUs in The Netherlands based on the NJ method for synonymous distances of all individual V3 sequences to the common ancestor of the respective risk group: (a) for homosexual men; (b) for IDUs. Confidence intervals for the estimated onset dates are shown. Regression lines are shown. (a) $r = 0.39$; $r^2 = 0.15$; slope, 0.0050; X intercept, 1977; $p < 0.00001$. (b) $r = 0.36$; $r^2 = 0.13$; slope, 0.0048; X intercept, 1981; $p < 0.01$.

tance to the ancestor (Fig. 3 and data not shown). The HIV-1 subtype B synonymous divergence rates were similar in Dutch homosexual men, Dutch IDUs, and U.S. subjects, indicating that the rate of the molecular clock is relatively stable across various HIV-1 populations. Our analysis revealed that the earliest possible years of the onset of HIV-1 epidemics among Dutch homosexual men and IDUs are 1977 (95% CI: 1976–1978) and 1981 (1980–1983), respectively, which is in agreement with the epidemiological data (van Haastrecht et al. 1992). Similar results for the years of virus introductions were obtained when regression analysis was based on the mean synonymous distances per sampling year (Fig. 4). As for the U.S. epidemic, this approach resulted in a great increase of the correlation coefficients. Our results suggest that the HIV-1 subtype B epidemic started in Europe 2 or 3 years after it started in the United States, which is in accord with the results of others (Holmes et al. 1995).

Fig. 4. Estimation of the onset years of the HIV-1 epidemics among homosexual men and IDUs in The Netherlands based on the mean synonymous distances per sampling year of the individual V3 sequences to the common ancestor of the respective risk group: (a) for homosexual men; (b) for IDUs. Confidence intervals for the onset dates are calculated based on data from Fig. 3. Regression lines are shown. (a) $r = 0.89$; $r^2 = 0.79$; slope, 0.0059; X intercept, 1978; $p < 0.001$. (b) $r = 0.80$; $r^2 = 0.64$; slope, 0.0048; X intercept, 1982; $p < 0.01$.

In contrast to the env V3 region, analysis by either the ML or the NJ method of nucleotide distances of pol sequences from Dutch homosexual men and IDUs to their common ancestors provided an accurate estimate for the years of the onset of HIV-1 epidemics in the two risk groups (Fig. 5, for the ML method). Analysis of synonymous and nonsynonymous distances of HIV-1 pol sequences to the common ancestors revealed that both were increasing significantly over the observation period (data not shown).

The Onset of the HIV-1 Subtype B Global Epidemic

To estimate the onset year of virus diversification within HIV-1 subtype B, we combined all env and pol sequences used to estimate the onset years of the outbreaks in the United States and The Netherlands with sequences with known sampling years obtained from other countries and/or risk groups. For the env region, we added sequences from Brazil (de Wolf et al. 1994; Morgado et

Fig. 5. Estimation of the onset years of the HIV-1 epidemics among homosexual men and IDUs in The Netherlands based on evolutionary distances of all individual pol sequences to the common ancestor of the respective risk group, calculated by the ML method: (a) for homosexual men; (b) for IDUs. Confidence intervals for the estimated onset dates are shown. Regression lines are shown. (a) $r = 0.52$; $r^2 = 0.27$; slope, 0.0017; X intercept, 1979; $p < 0.000001$. (b) $r = 0.63$; $r^2 = 0.40$; slope, 0.0018; X intercept, 1980; $p < 0.0001$.

al. 1994), Thailand (Ou et al. 1993; Kalish et al. 1994), and Russia (Lukashov et al. 1995b). In total, 283 env V3 sequences, sampled between 1983 and 1994, were used in the analysis. For the pol region, we added sequences obtained from HIV-infected individuals attending the Academic Medical Center in Amsterdam and sequences from untreated individuals provided by the Los Alamos (<http://hiv-web.lanl.gov/>) and Stanford HIV RT and Protease Sequence (<http://hivdb.stanford.edu/hiv/>) databases. In total, we analyzed 270 HIV-1 pol sequences obtained between 1983 and 2000.

Based on the synonymous distances of the env V3 sequences to the common node of HIV-1 subtype B, we estimated 1976 (95% CI: 1974–1977) as the onset year of the HIV-1 subtype B diversification and the subtype B global epidemic (Figs. 6a and b), a result in accord with the notion that the global HIV-1 subtype B epidemic started in the United States and spread worldwide from there. This result was obtained using either the synony-

mous distances of all subtype B sequences (Fig. 6a) or the mean distances per sampling year (Fig. 6b). The same result was obtained when the U.S. sequences were excluded from our analysis (data not shown). Additional analysis of HIV-1 subtype B sequences obtained from Africa (Kuiken et al. 1999) revealed that their synonymous distances to the subtype B common ancestor are similar to those of U.S. or Dutch sequences sampled in the same year (data not shown), indicating that HIV-1 subtype B strains found in Africa diversified from the common node at the same time as in the United States and Europe. Analysis of our international set of pol sequences gave 1977 (95% CI: 1976–1980) as the onset year of virus diversification within HIV-1 subtype B.

Discussion

In the present study, we used a population-based method to test the reliability of various phylogenetic methods in dating the onset of the AIDS epidemics caused by HIV-1 subtype B and the diversification within subtype B. We studied the evolution of HIV-1 env and pol regions in the course of three HIV-1 subtype B local epidemics: among mixed risk groups in the United States and among homosexual men and IDUs in The Netherlands. For each epidemic, the most probable common ancestors were reconstructed using several phylogenetic methods, and evolutionary distances of sequences to the common ancestors were analyzed using several statistical approaches. Since the onset of these epidemics is known from rigorous retrospective serological and/or epidemiological data, we could determine whether analysis of evolutionary distances of HIV-1 strains within an epidemic could be used to reveal their onset dates and, moreover, which phylogenetic model and HIV-1 genetic region would provide the most reliable estimate.

In each epidemic, irrespective of the phylogenetic method and HIV-1 genetic region used, distances to the common ancestor of individual sequences in any calendar year varied within broad ranges, and sequences close to the ancestors were found decades after the onset of the epidemic (Figs. 2–6). These findings confirm that the HIV-1 evolution rate is highly variable among individual virus lineages (Lukashov et al. 1995a; Lukashov and Goudsmit 1997; Wolinsky et al. 1996; Korber et al. 1994, 1998; Goudsmit and Lukashov 1999; Leitner et al. 1996). They strongly suggest that dating of an evolutionary event in the HIV-1 history or validation of a phylogenetic model cannot easily be based on a single sequence (Zhu et al. 1998; Korber et al. 2000).

Nevertheless, in the human/virus populations we studied, evolutionary distances of virus sequences to their ancestors increased significantly over the course of the epidemics, demonstrating that a molecular clock is operational in recent HIV-1 evolution. When we extrapo-

Fig. 6. Evolutionary distances of the globally collected HIV-1 subtype B V3 and pol sequences to their reconstructed common ancestors in relation to sequence sampling years. Regression analysis of synonymous distances of the V3 sequences [(a) based on all individual sequences and (b) based on the mean distances per sampling year], calculated using the NJ method, and evolutionary distances of the pol sequences, calculated using the ML method [(c) based on all individual sequences and (d) based on the mean distances per sampling year], is

shown. Confidence intervals for the estimated date of virus introduction are calculated based on synonymous distances of all individual V3 sequences (shown in a and b as well as ML distances of all pol sequences (shown in c and d). (a) $r = 0.33$; $r^2 = 0.11$; slope, 0.0044; X intercept, 1976; $p < 0.0000001$. (b) $r = 0.96$; $r^2 = 0.92$; slope, 0.0043; X intercept, 1976; $p < 0.00001$. (c) $r = 0.44$; $r^2 = 0.20$; slope, 0.0015; X intercept, 1977; $p < 0.000001$. (d) $r = 0.97$; $r^2 = 0.94$; slope, 0.0017; X intercept, 1979; $p < 0.000001$.

lated regression lines for the three epidemics back to the past, analysis of nucleotide distances among V3 sequences by either the ML or the NJ method predicted that the onset of the epidemics occurred in the 1950s–1960s. However, these estimates are much earlier than the onset dates established based on epidemiological data (Jaffe et al. 1983; Selik et al. 1984). The first U.S. cases of AIDS were retrospectively identified not earlier than 1978 (Jaffe et al. 1983; Selik et al. 1984). This suggests that the U.S. epidemic probably began in the early or mid 1970s, since the minimal asymptomatic period of an HIV-1 infection is several years. Retrospective epidemiological data suggest 1976 as the onset year of the HIV-1 epidemic in the United States (Grmek 1990). For the Dutch homosexual men and IDUs, intensive retrospective serological and epidemiological studies established that the HIV-1 epidemics in these risk groups started in 1977 and 1980, respectively (van Haastrecht et al. 1992; Coutinho et al. 1986, 1987; Lukashov et al.

1996a). The first AIDS cases were diagnosed in The Netherlands in 1982 in homosexual men and in 1985 in IDUs (van Haastrecht et al. 1992).

Our analysis of the nucleotide distances among the U.S. env sequences (Figs. 2a and b) resulted in exactly the same onset dates as those obtained in an earlier study using different phylogenetic approaches—before 1967 and even in 1954—based on the analysis of env sequences and depending upon the method used (Korber et al. 2000). The authors suggested, that there was a 5- to 15-year “hidden” period in the subtype B evolution before the “visible” U.S. epidemic started in 1976–1978 (Korber et al. 2000). However, the existence of such a long period in HIV-1 evolution in the United States, when the virus would be present in only a few lineages for up to 15 years, is not obvious from the phylogenetic tree of the U.S. HIV-1 strains, in which all individual viruses branch out from virtually a single node and no individual virus lineages—long internal branches—can

be distinguished. This indicates that the U.S. epidemic was the result of a rapid propagation of a single virus and suggests either that this hypothetical period has been very short (months, rather than years/decades) or that only one virus strain survived it and gave rise to the U.S. epidemic. This surviving strain is the most recent common ancestor of the current U.S. HIV-1 population. Since this strain did exist at the end of the hypothetical "hidden" period, our and earlier (Korber et al. 2000) calculations, which are aimed at the most recent common ancestor, should come to that later date.

Moreover, our results based on pol sequences from the two Dutch populations contrasted with estimates based on nucleotide distances among env sequences by us (this study) and by others (Korber et al. 2000). Furthermore, an earlier analysis of gag, another slow-evolving region, also resulted in a much more recent onset date for the U.S. epidemic—1972 (Korber et al. 2000). To learn whether the characteristics of the evolution of the fast-evolving env V3 region, compared to the slow-evolving pol or gag regions, could lead to overestimation of the age of the HIV-1 epidemic, we studied synonymous and nonsynonymous distances of the V3 sequences from the United States as well as from the Dutch homosexual men and IDUs to the reconstructed common ancestors of the viruses circulating in these human populations.

Our analysis demonstrated that this overestimation of the age of the HIV-1 epidemics based on the env sequences results from nonsynonymous substitutions, whose numbers were not increasing significantly in any of the three virus populations over the observation period. In contrast, analysis of synonymous env V3 distances provided estimates of the onset years for the outbreaks that were in agreement with the epidemiological data as well as with the analysis of slow-evolving regions, including pol (this study) and gag (Korber et al. 2000).

We previously reported that nonsynonymous heterogeneity within the V3 region is not currently increasing (Lukashov and Goudsmit 1997). The most likely explanation for this phenomenon is that the number of nonsynonymous substitutions that can accumulate within this region is limited due to its functional significance. The lack of increase may be considered an example of functional saturation. Soon after HIV-1 introduction into a human population, the nonsynonymous heterogeneity of the V3 region begins to increase, but this process appears to reach its plateau rapidly, due to the high evolution rate of the V3 region (Lukashov and Goudsmit 1997). Although the env V3 regions of each and every individual HIV-1 subtype B virus strain are evolving rapidly, and most of the nucleotide substitutions within this region are nonsynonymous, the mean intrasubtype nonsynonymous heterogeneity does not increase beyond a certain limit. This phenomenon can be compared to

Brownian movements, in which the continual traffic of a body's atoms does not lead to disintegration of the body. In contrast, synonymous heterogeneity within the V3 region is not subject to this limitation and continues to increase, as still does the nonsynonymous heterogeneity of the pol and gag genes, due to the much lower evolution rate of these genetic regions.

The onset of HIV-1 subtype B epidemics in our three study populations was dated without employing the epidemiological data but is in remarkable agreement with the history of AIDS epidemics based on retrospective seroepidemiological data. In our earlier analysis of the two outbreaks of the HIV-1 subtype C epidemic in Ethiopia, we also observed an agreement between the molecular and the seroepidemiological data (Abebe et al. 2001a, b). Our results suggest that the onset of virus diversification coincided with the onset of the HIV-1 epidemics in each population. They support the theory that the western AIDS epidemic arose in the United States and spread worldwide from there. Our results suggest that the subtype B lineage existed prior to 1975 and that its current worldwide diversity is a result of virus diversification within HIV-1 subtype B that coincided with the onset of the AIDS epidemic in the current risk groups.

For each of the three epidemics studied, the dates of virus introductions were determined within narrow 95% confidence intervals, whereas older evolutionary events, such as separation between HIV-1 subtypes, could be dated only within a period of decades (Goudsmit and Lukashov 1999; Korber et al. 1998, 2000). The very fact that the separation between HIV-1 subtype lineages is an older event than diversification within a single HIV-1 subtype is relevant to the breadth of the confidence intervals, but several other factors are also important. It has been noted that some HIV-1 subtypes are significantly more distant from the group M common node than others, which is often interpreted as evidence for different evolution rates of the HIV-1 subtype lineages. However, as our study demonstrates, HIV strains both close to and far from the common ancestor are present in an infected population at any given time point. In the early history of HIV-1, before the global epidemic started, massive population bottlenecks no doubt allowed the survival of only a limited number of virus strains, even a single virus, whose distance to the ancestor could be, by chance, either large or small. This subsequently resulted in distances of the offspring of this virus—current HIV-1 subtypes—being, by chance, either large or small. These random bottleneck events, which are quite difficult to model, would strongly bias reconstruction of the ancient evolutionary history of HIV-1. Thus, by observing a growing plant for a long time, we are able to estimate its growth rate and to trace back in time the moment when the seed started to grow. However, what we cannot easily estimate is the age of this seed and how much time it spent in the soil after being put there by the farmer.

Acknowledgment. The authors thank Lucy Phillips for editing the manuscript.

References

- Abebe A, Lukashov VV, Pollakis G, et al. (2001a) Timing of the HIV-1 subtype C epidemic in Ethiopia based on early virus strains and subsequent virus diversification. *AIDS* 15:1555–1561
- Abebe A, Lukashov VV, Rinke De Wit TF, et al. (2001b) Timing of the introduction into Ethiopia of subcluster C' of HIV type 1 subtype C. *AIDS Res Hum Retrovirus* 17:657–661
- Buonagurio DA, Nakada S, Parvin JD, Krystal M, Palese P, Fitch WM (1986) Evolution of human influenza A viruses over 50 years: Rapid, uniform rate of change in NS gene. *Science* 232:980–982
- Coutinho RA, Krone WJA, Smit L, et al. (1986) Introduction of lymphadenopathy associated virus or human T lymphotropic virus (LAV/HTLV-III) into the male homosexual community in Amsterdam. *Genitourin Med* 62:38–43
- Coutinho RA, Goudsmit J, Paul DA, et al. (1987) The natural history of HIV infection in homosexual men. *Ann Inst Pasteur/Virol* 138:67–74
- de A Zannotto, PM, Gould EA, Gao GF, Harvey PH, Holmes EC (1996) Population dynamics of flaviviruses revealed by molecular phylogenies. *Proc Natl Acad Sci USA* 93:548–553
- de Wolf F, Hogervorst E, Goudsmit J, et al. (1994) Syncytium-inducing and non-syncytium-inducing capacity of human immunodeficiency virus type 1 subtypes other than B: Phenotypic and genotypic characteristics. *AIDS Res Hum Retrovirus* 10:1387–1400
- Dickerson RE (1971) The structure of cytochrome c and the rates of molecular evolution. *J Mol Evol* 1:26–45
- Eigen M, Nieselt-Struwe K (1990) How old is the immunodeficiency virus? *AIDS* 4:S85–S93
- Fitch WM, Leiter JME, Li X, Palese P (1991) Positive Darwinian evolution in human influenza A viruses. *Proc Natl Acad Sci USA* 88:4270–4274
- Gillespie JH (1993) Episodic evolution of RNA viruses. *Proc Natl Acad Sci USA* 90:10411–10412
- Gojobori T, Moriyama EN (1990) Molecular clock of viral evolution, and the neutral theory. *Proc Natl Acad Sci USA* 87:10015–10018
- Gojobori T, Moriyama EN, Ina Y, et al. (1990) Evolutionary origin of human and simian immunodeficiency viruses. *Proc Natl Acad Sci USA* 87:4108–4111
- Goudsmit J, Lukashov VV (1999) Dating the origin of HIV-1 subtypes. *Nature* 400:325–326
- Grmek MD (1990) History of AIDS: Emergence and origin of a modern pandemic. Princeton University Press, Princeton, NJ
- Haastrecht van HJA, van den Hoek JAR, Mientjes GH, Coutinho RA (1992) Did the introduction of HIV among homosexual men precede the introduction of HIV among injecting drug users in the Netherlands? *AIDS* 6:131–132
- Holmes EC, Zhang LQ, Robertson P, et al. (1995) The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh. *J Infect Dis* 171:45–53
- Jaffe HW, Bergman DJ, Selik RM (1983) Acquired immune deficiency syndrome in the United States: the first 1,000 cases. *J Infect Dis* 148:339–345
- Kalish ML, Luo C-C, Weniger BG, et al. (1994) Early HIV type 1 strains in Thailand were not responsible for the current epidemic. *AIDS Res Hum Retrovirus* 10:1573–1575
- Kimura M (1983) The neutral theory of evolution. Cambridge University Press, Cambridge
- Korber BTM, MacInnes K, Smith RF, Myers G (1994) Mutational trends in V3 loop protein sequences observed in different genetic lineages of human immunodeficiency virus type 1. *J Virol* 68:6730–6744
- Korber B, Theiler J, Wolinsky S (1998) Limitations of a molecular clock applied to considerations of the origin of HIV-1. *Science* 280:1868–1871
- Korber B, Muldoon M, Theiler J, et al. (2000) Timing the ancestor of the HIV-1 pandemic strains. *Science* 288:1789–1796
- Kuiken CL, Goudsmit J (1994) Silent mutation pattern in V3 sequences distinguishes virus according to risk group in Europe. *AIDS Res Hum Retrovirus* 10:319–320
- Kuiken CL, Lukashov VV, Baan E, et al. (1996a) Evidence for limited within-person evolution of the V3 domain of the HIV-1 envelope in the Amsterdam population. *AIDS* 10:31–37
- Kuiken CL, Cornelissen MTE, Zorgdrager F, Hartman S, Gibbs AJ, Goudsmit J (1996b) Consistent risk group-associated differences in human immunodeficiency virus type 1 vpr, vpu and V3 sequences despite independent evolution. *J Gen Virol* 77:783–792
- Kuiken C, Foley B, Hahn B, et al. (eds) (1999) Human retroviruses and AIDS: A compilation and analysis of nucleic acid and amino acid sequences. Theoretical Biology and Biophysics, Group T-10, Los Alamos National Laboratory, Los Alamos, NM
- Kumar S, Tamura K, Nei M (1993) Molecular evolutionary genetics analysis (MEGA). Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park
- Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J (1996) Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc Natl Acad Sci USA* 93:10864–10869
- Leitner T, Albert J (1999) The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc Natl Acad Sci USA* 96:10752–10757
- Li W-H, Tanimura M, Sharp PM (1988). Rates and dates of divergence between AIDS virus nucleotide sequences. *Mol Biol Evol* 5:313–330
- Lukashov VV, Goudsmit J (1995) Increasing genotypic and phenotypic selection from the original genomic RNA populations of HIV-1 strains LAI and MN (NM) by peripheral blood mononuclear cell culture, B-cell-line propagation and T-cell-line adaptation. *AIDS* 9:1307–1311
- Lukashov VV, Goudsmit J (1997) Founder virus population related to route of virus transmission: A determinant of intrahost human immunodeficiency virus type 1 evolution? *J Virol* 71:2023–2030
- Lukashov VV, Kuiken CL, Goudsmit J (1995a) Intrahost human immunodeficiency virus type 1 evolution is related to length of the immunocompetent period. *J Virol* 69:6911–6916
- Lukashov VV, Cornelissen MTE, Goudsmit J, et al. (1995b) Simultaneous introduction of distinct HIV-1 subtypes into different risk groups in Russia, Byelorussia and Lithuania. *AIDS* 9:435–439
- Lukashov VV, Kuiken CL, Vlahov D, Coutinho RA, Goudsmit J (1996a) Evidence for HIV type 1 strains of U.S. intravenous drug users as founders of AIDS epidemic among intravenous drug users in Northern Europe. *AIDS Res Hum Retrovirus* 12:1179–1183
- Lukashov VV, Kuiken CL, Boer K, Goudsmit J (1996b) HIV type 1 subtypes in The Netherlands circulating among women originating from AIDS endemic regions. *AIDS Res Hum Retrovirus* 12:951–953
- McGeoch DJ, Cook S, Dolan A, Jamieson FE, Telford EAR (1995) Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. *J Mol Biol* 247:443–458
- Morgado MG, Sabino EC, Shpaer EG, et al. (1994) V3 region polymorphism in HIV-1 from Brazil: Prevalence of subtype B strains divergent from North American/European prototype and detection of subtype F. *AIDS Res Hum Retrovirus* 10:569–576
- Nichol ST, Rowe JE, Fitch WM (1993) Punctuated equilibrium and positive Darwinian evolution in vesicular stomatitis virus. *Proc Natl Acad Sci USA* 90:10424–10428
- Orito E, Mizokami M, Ina Y, et al. (1989) Host-independent evolution and a genetic classification of the hepadnavirus family based on nucleotide sequences. *Proc Natl Acad Sci USA* 86:7059–7062.
- Ou C-Y, Takebe Y, Weniger BG, et al. (1993) Independent introduction of two major HIV-1 genotypes into distinct high-risk populations in Thailand. *Lancet* 341:1171–1174

- Ou C-Y, Ciesielski CA, Myers G, et al. (1992) Molecular epidemiology of HIV transmission in a dental practice. *Science* 256:1165–1171
- Querat G, Audoly G, Sonigo P, Vigne R (1990) Nucleotide sequence analysis of SA-OMVV, a visna-related ovine lentivirus: Phylogenetic history of lentiviruses. *Virology* 175:434–447
- Salemi M, Strimmer K, Hall WW, et al. (2000) Dating the radiation of HIV-1 group M in 1930s using a new method to uncover clock-like molecular evolution. *FASEB J* on-line: 10.1096/fj.00-0449fje
- Salemi M, Strimmer K, Hall WW, et al. (2001) Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. *FASEB J* 15:276–278
- Sanchez CM, Gebauer F, Sune C, Mendez A, Dopazo J, Enjuanes L (1992) Genetic evolution and tropism of transmissible gastroenteritis coronaviruses. *Virology* 190:92–105
- Selik RM, Haverkos HW, Curran JW (1984) Acquired immune deficiency syndrome (AIDS), trends in the United States, 1978–1982. *Am J Med* 76:493–500
- Sharp PM, Li W-H (1988) Understanding the origins of AIDS viruses. *Nature* 336:315
- Villaverde A, Martinez MA, Sobrino F, Dopazo J, Moya A, Domingo E (1991) Fixation of mutations at the VP1 gene of foot-and-mouth disease virus. Can quasispecies define a transient molecular clock? *Gene* 103:147–153
- Wolinsky SM, Kunstman KJ, Safritz JT, Koup RA, Neumann AU, Korber BTM (1996) HIV-1 evolution and disease progression. *Science* 274:1010–1011
- Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449
- Yokoyama S, Chung L, Gojobori T (1988) Molecular evolution of the human immunodeficiency and related viruses. *Mol Biol Evol* 5: 237–251
- Zhu T, Korber BT, Nahmias AJ, Hooper E, Sharp PM, Ho DD (1998) An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* 391:594–597