JOURNAL OF VIROLOGY, June 1997, p. 4761–4770 Vol. 71, No. 6
0022-538X/97/$04.00+0

Tempo and Mode of Nucleotide Substitutions in *gag* and *env* Gene Fragments in Human Immunodeficiency Virus Type 1 Populations with a Known Transmission History

THOMAS LEITNER,[1]* SUDHIR KUMAR,[2] AND JAN ALBERT[3]

Theoretical Biology and Biophysics, Group T-10, Los Alamos National Laboratory, Los Alamos, New Mexico 87545[1]; Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802[2]; and Department of Clinical Virology, Swedish Institute for Infectious Disease Control, Karolinska Institute, S-105 21 Stockholm, Sweden[3]

Received 20 June 1996/Accepted 13 March 1997

The complex evolutionary process of human immunodeficiency virus type 1 (HIV-1) is marked by a high level of genetic variation. It has been shown that the HIV-1 genome is characterized by variable and more constant regions, unequal nucleotide frequencies, and preference for G-to-A substitutions. However, this knowledge has largely been neglected in phylogenetic analyses of HIV-1 nucleotide sequences, even though these analyses are applied to a number of important biological questions. The purpose of this study was to identify a realistic model of HIV-1 evolution and to statistically test if the application of such a model significantly improves the accuracy of phylogenetic analyses. A unique and recently reported HIV-1 transmission cluster consisting of nine infected individuals, for whom the direction and time for each transmission were exactly known, formed the basis for the analyses which were performed under a general model of nucleotide substitution using population sequences from the *env* V3 and p17^{gag} regions of the HIV-1 genome. Examination of seven different substitution models by maximum-likelihood methods revealed that the fit of the general reversible (REV) model was significantly better than that of simpler models, indicating that it is important to account for the asymmetric substitution pattern of HIV-1 and that the nucleotide substitution rate varied significantly across sites. The shape parameter α, which describes the variation across sites by a gamma distribution, was estimated to be 0.38 and 0.25 for *env* V3 and p17^{gag}, respectively. In *env* V3, the estimated average transition/transversion rate ratio was 1.42. Thus, the REV model with variable rates across sites (described by a gamma distribution) provides the best description of HIV-1 evolution, whereas simple models are unrealistic and inaccurate. It is likely that the accuracy of phylogenetic studies of HIV-1 and many other viruses would improve substantially by the use of more realistic nucleotide substitution models. This is especially true when attempts are made to estimate the age of distant viral ancestors from contemporary viral sequences.

Human immunodeficiency virus type 1 (HIV-1), which is a member of the *Lentivirus* genus (family *Retroviridae*), is characterized by a high level of genetic variation (30). Members of *Retroviridae* are RNA viruses that replicate through a DNA intermediate (2, 46). The viral RNA is copied into DNA by the viral enzyme reverse transcriptase. This process is quite error prone and forms the basis for the high genetic variability of these viruses (5, 12, 17, 45). However, the observed genetic variation is a product of a complicated process influenced by many factors, most notably mutation and selection. The effects of these factors are still not well understood or are difficult to study in practice. However, the mutation rate of reverse transcriptase has been suggested to be 3.4×10^{-5} mutations per base pair per replication cycle in vivo (30), and several investigations have documented differences in the selective pressures on different viral genes (19, 22, 37). In addition, the rate of genetic recombination in retroviruses is high, and this greatly contributes to the genetic variation (13, 35). The apparent substitution rates across the genome are not the same, as illustrated by the presence of the five variable regions (V1 to V5) in the HIV-1 *env* gene (26, 38).

Unlike bacterial genomes (40), the HIV-1 genome is A rich (≈36%) and C poor (≈18%). In vertebrates, the base composition varies between genes, and it has been suggested that the mosaic structure of the chromosomes reflects the varying G+C content (4, 14). Because of varying G+C content, the codon usage of a gene may vary with its genomic G+C context. It has been reported that HIV codon usage is dramatically different from that of cellular genes due to a high preference for A-rich codons and that this also results in a biased amino acid composition of viral proteins (3). Furthermore, HIV-1 has been shown to produce extensive and monotonous G-to-A nucleotide substitutions, especially in the GpA dinucleotide (48). It has been suggested that a skewed dCTP pool during reverse transcription is the cause of the G-to-A hypermutation (47).

Phylogenetic methods have been applied to various epidemiological questions, including studies of the spread of HIV-1 to different continents, the origin of the virus, routes of infection, and forensic investigations (1, 24, 27, 28, 32–34, 36). While considerable attention has been given to which genes and tree-building methods should be used, few attempts have been made to describe the actual evolutionary process in HIV-1 genes and to test different evolutionary models. It is clear that in order to make reliable phylogenetic inferences from HIV-1 sequences, the biased nucleotide composition, high mutation rate, differences of mutation rate in different regions, and preferred G-to-A substitutions should be accounted for.

In this study, we have used a unique data set of population

* Corresponding author. Mailing address: Theoretical Biology and Biophysics, Group T-10, MS K710, Los Alamos National Laboratory, Los Alamos, NM 87542. Phone: (505) 665-2594. Fax: (505) 665-3493. E-mail: tkl@t10.lanl.gov.

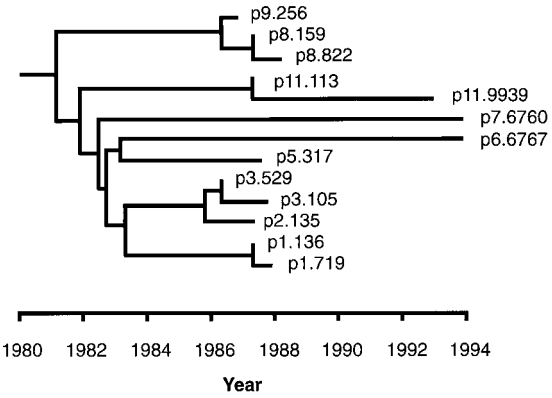

FIG. 1. The true phylogeny of the virus populations in a known HIV-1 transmission history from Sweden. This tree was constructed with the exact knowledge of who infected whom as well as when the transmissions occurred and when the blood samples were drawn from the infected individuals (23). Each lineage split indicates a transmission event, and each tip of a branch represents a sequence sample; patient and sample numbers are indicated. Note that for several individuals, serial samples were included.

DNA sequences of *env* V3 and p17^{gag} gene fragments derived from an HIV-1 transmission history with a known phylogeny (23). Different nucleotide substitution models ranging from the simple Jukes-Cantor (JC) model to the general reversible Markov process (REV) model were evaluated by maximum-likelihood calculations. The variation of substitution rates across sites was tested by comparison of different models with and without gamma-distributed rates across sites. We found that the inference of branch lengths was greatly improved by accounting for the special substitution pattern of HIV-1.

MATERIALS AND METHODS

Study population and the true tree. Viral populations from a group of HIV-1-infected individuals with well-characterized epidemiological relationships (23) were studied. Briefly, the index case, a Swedish male (p1) who became HIV-1 infected in Haiti 1980, infected several females (p2, p5, p7, p8, and p11) between 1981 and 1983. In addition, samples from a later male sexual partner (p6) and two children (p3 and p9) of the females were included in the phylogeny. Blood samples were obtained at different time points between 1986 and 1993. From some individuals, more than one sample was available. The information about when the transmissions had occurred and when the samples were obtained was compiled into a tree which shows the evolutionary history of the transmitted virus populations (Fig. 1). The branch lengths in the tree describe the actual time over which the viral populations have evolved. The tree topology (considering only the branching order) has been presented and examined before (23).

Sequence data. DNA sequences from the HIV-1 p17^{gag} and *env* V3 regions of the viral genome were determined by direct population sequencing as previously described (25). In the earlier study of reconstructed tree topologies (23), polymorphic nucleotide positions within a sample were described by using the IUPAC-IUB codes (15) (GenBank accession no. U68496 to U68521). Here, the population sequences were reanalyzed to determine a majority-rule consensus sequence for each sample. These sequences are available from the author upon request. This was necessary because available computer programs for estimating the nucleotide substitution patterns by maximum-likelihood methods cannot handle more than one character state at each site in a sequence. Since quantifications of mixed nucleotide positions are possible with the sequencing technique used (25), the most prevalent nucleotide at each such polymorphic site was assigned to that position. When two nucleotides existed in equal amounts at one position, the nucleotide that was more rare in total frequency over the fragment was chosen (the relative nucleotide frequency in *env* V3 was A > T > G > C, and in p17^{gag} it was A > G > T > C [Table 1]). For instance, in a 50% A and 50% G position (IUPAC-IUB code R), G would be chosen. Note that these polymorphisms describe variation between individual sequences within a population (intrasample variation) and not between populations (intersample variation). For clarity, we will refer to sites with polymorphisms within a population as multistate sites and between populations as polymorphic sites. The V3 and p17 fragments were divided into first, second, and third codon positions to allow estimations from each position; data sets consisting of first and second codon positions and the whole fragment (all three codon positions) were also evaluated.

Fit of nucleotide substitution models. Several substitution models were investigated by maximum-likelihood calculations by using the program PAML (51). All calculations were done by using the true tree topology of the investigated data set, ((256,(822,159)),((113,9939),(6760,((317,6767),((135,(529,105)),(719,136))))))), described above and shown in Fig. 1. The tree was treated as unrooted.

The rate matrix for a reversible homogeneous Markov process has the general form

$$Q = \begin{bmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \cdot \end{bmatrix} \quad (1)$$

where the diagonals are given as $Q_{ii} = -\Sigma_{i \neq j} Q_{ij}$ and the order of nucleotides is T, C, A, G. $Q_{ij}\Delta t$ is the probability that nucleotide i will change into nucleotide j in an infinitesimal time interval Δt. a, b, c, d, e, and f are the rate parameters, and the π's are the frequency parameters. One of the four rate parameters is redundant. The REV model assumes that the reversibility restriction, $\pi_i Q_{ij} = \pi_j Q_{ji}$ (44, 50), and eight free parameters need to be estimated. In other publications, the REV model is sometimes also referred to as the general-time-reversible model.

The other tested substitution models are special cases of the REV model. The simplest case is described by the JC model (16), where all π's are equal (0.25) and all rate parameters are equal ($a = b = c = d = e = f$). The Kimura (K2) model (18) has equal nucleotide frequencies but allows for different rates of transitions and transversions ($a = f$; $b = c = d = e$). Felsenstein described a model in 1981 (F81 model) that assumed empirical nucleotide frequencies and equal rate parameters (6). Hasegawa et al. (10) described a model (HKY model) where both empirical base frequencies and different rates of transitions and transversions were allowed ($a = f = \kappa\mu$ and $b = c = d = e = \mu$). Felsenstein implemented a similar model (F84 model) in later (current) versions of the program DNAML in the PHYLIP package (7) $\{a = [1 + \kappa/(\pi_T + \pi_C)]\mu, f = [1 + \kappa/(\pi_A + \pi_G)]\mu$ and $b = c = d = e = \mu\}$. The Tamura-Nei (TN) model (43) makes only the assumption that the transversion rates are equal ($b = c = d = e$). In all models, the common assumption of rate constancy over nucleotide sites was tested by incorporating gamma-distributed rates across sites (size and shape parameter α). The effects of different α's in the gamma distribution are shown in Fig. 2. For computational reasons, the continuous gamma distribution was approximated by a discrete gamma model with eight categories. The shape parameter was also estimated by three parsimony procedures: (i) the method of moments; (ii) a

TABLE 1. Observed average nucleotide frequencies in the investigated *env* V3 and p17^{gag} fragments

Codon position	Observed nucleotide frequency (range)			
	π_A	π_C	π_G	π_T
env V3				
1	0.4937 (0.4615–0.5385)	0.1192 (0.1099–0.1319)	0.2409 (0.2198–0.2637)	0.1462 (0.1319–0.1758)
2	0.3584 (0.3297–0.3956)	0.1986 (0.1758–0.2088)	0.1817 (0.1648–0.2198)	0.2612 (0.2527–0.2747)
3	0.5359 (0.4835–0.5604)	0.1243 (0.0989–0.1648)	0.0566 (0.0330–0.0879)	0.2832 (0.2527–0.3077)
1+2+3	0.4627 (0.4359–0.4725)	0.1474 (0.1355–0.1575)	0.1598 (0.1502–0.1722)	0.2302 (0.2125–0.2527)
p17^{gag}				
1	0.3179 (0.3147–0.3217)	0.2001 (0.1888–0.2168)	0.3330 (0.3147–0.3497)	0.1490 (0.1329–0.1678)
2	0.3948 (0.3846–0.4056)	0.1700 (0.1678–0.1748)	0.1985 (0.1888–0.2098)	0.2367 (0.2308–0.2448)
3	0.4857 (0.4615–0.4965)	0.1329 (0.1189–0.1468)	0.2146 (0.2028–0.2308)	0.1668 (0.1538–0.1818)
1+2+3	0.3995 (0.3916–0.4033)	0.1677 (0.1632–0.1748)	0.2487 (0.2471–0.2564)	0.1841 (0.1748–0.1935)

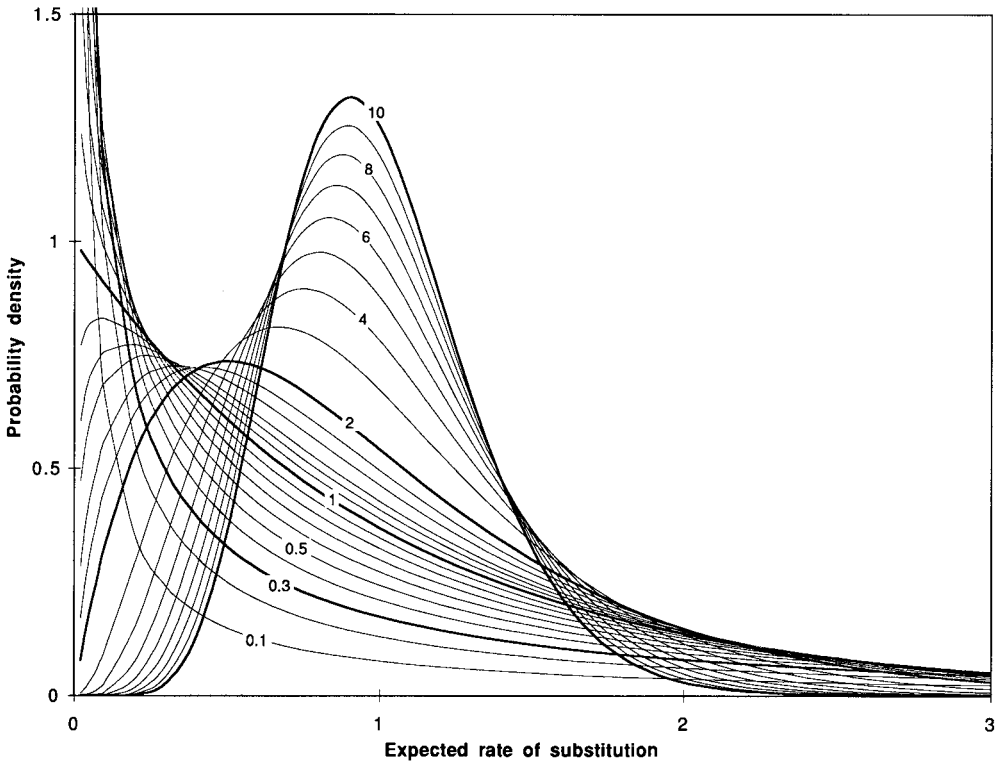

FIG. 2. The gamma distribution described by different shape parameters (α = 0.1 to 1.3, 1.5, 1.7, and 2 to 10). The mean and variance of the distribution are 1 and $1/\alpha$, respectively.

method according to Sullivan et al. (41); and (iii) a recent method by Yang and Kumar (YK model) (52), using the PAMP program (51).

The performances of the different substitution models and the discrete gamma model were evaluated by comparisons according to the log likelihood ratio test. If the log likelihood under a model with p parameters is l_1 and that for a submodel with q parameters fixed out of p is l_2, then $2\Delta l = 2(l_1 - l_2)$ can be compared with the χ^2 distribution with q degrees of freedom.

Branch lengths of the true tree were also calculated by a weighted maximum-parsimony procedure that accounts for differences among all nucleotide substitution types. The specific substitution events of each type were counted by using the true tree. These counts were then inverted, so that the most frequent type of substitutions was given the lowest score and the most rare events were to be considered the most informative (11). The inverted matrix was corrected for triangle inequality (42). Finally, this matrix was used to calculate branch lengths in the true tree. These calculations were performed by using the programs PAUP and MacClade (29, 42).

RESULTS

HIV-1 population sequences from a known phylogeny. In this study, we analyzed HIV-1 population sequences from a known HIV-1 transmission cluster consisting of nine infected individuals, for whom the direction and time for each transmission were exactly known. A total of 13 samples were included in the study, because more than one sample was available from some of the individuals. It was possible to construct a true phylogenetic tree based on the knowledge about when the transmission had occurred and when the samples had been obtained (Fig. 1). We have recently performed an analysis of the accuracy of the topologies (branching order) of trees obtained with different phylogenetic methods (23). Here we have investigated the accuracy of the branch lengths of trees obtained with different nucleotide substitution models under maximum-likelihood and maximum-parsimony calculations.

Variation of substitution rates across sites. Figures 3 and 4 show the alignments of the *env* V3 and p17^{gag} sequences, respectively. It is clear from a quick inspection that the V3 fragment is more variable than the p17 fragment and that there are regions within each fragment that are more variable than others. In total, 68% of the sites in the V3 fragment were observed as unvaried (60% in first, 71% in second, and 73% in third codon positions), compared to 85% of the sites in p17 (84% in first, 90% in second, and 82% in third codon positions) (Fig. 3 and 4). The differences in the substitution rates between fragments as well as between the three codon positions were expected, since both purifying and positive selection differ between genes and between codons, as illustrated by different synonymous/nonsynonymous ratios in different HIV-1 genes (8, 22, 53). However, it has never previously been tested if there are statistically significant differences in evolutionary rates across sites or if the accuracy of branch length estimates improves by accounting for rate variation across sites. To address these questions, we estimated the shape parameter (α; called gamma parameter) of the gamma distribution of substitution rates among sites under a REV model of nucleotide substitution for each codon position individually as well as for all codon positions together. The estimate of the gamma parameter is inversely related to the extent of differences in substitution rates across sites. If $\alpha > 2$, then the differences in substitution rates across sites are small and a uniform rate model (Poisson distribution) can fit the data (Fig. 2). The continuous gamma distribution was approximated by a discrete function with eight rate categories. When the number of rate categories was increased above eight, no significant increase of fit was observed (data not shown).

In our analyses, we attempted to use maximum-likelihood methods (discrete gamma model) with a REV model because

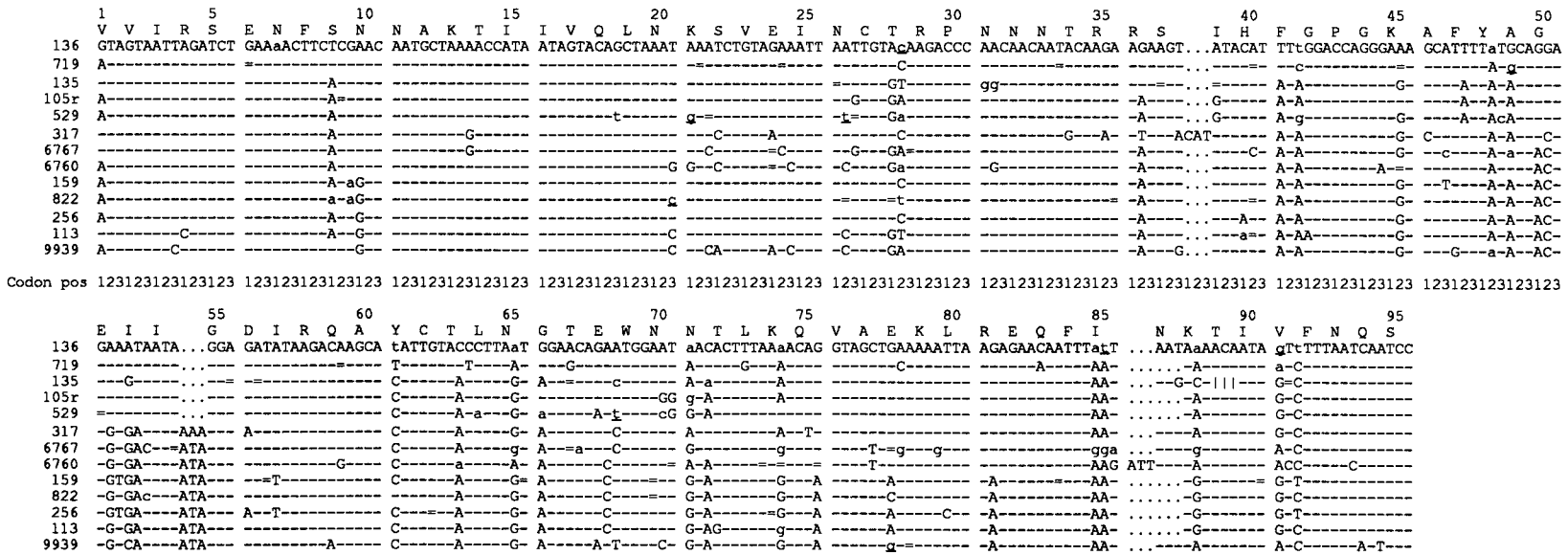

FIG. 3. Nucleotide alignment of the investigated *env* V3 fragment. The first available sequence from the index case (p1.136) is used as master sequence, to which the others are compared. To simplify identification of the positions, the translated amino acid sequence of the master sequence is also indicated, as well as codon positions. A dash denotes identity with the master sequence, while a letter indicates a mismatch. Lowercase letters indicate that the nucleotide is the major type in a multistate position. An equal sign indicates that the nucleotide is the major type in a multistate position with identity to the master sequence. An underlined lowercase letter indicates that the nucleotide is derived from a multistate position were the nucleotides were present in equal frequencies (see Materials and Methods). Gaps were introduced to align the sequences and are indicated by dots. The columns that included gaps were omitted in subsequent analyses. ||| in sequence 135 indicates that the codon was inserted as a repeat in a fraction of the population, thereby creating a length polymorphism; however, the major form had the indicated length.

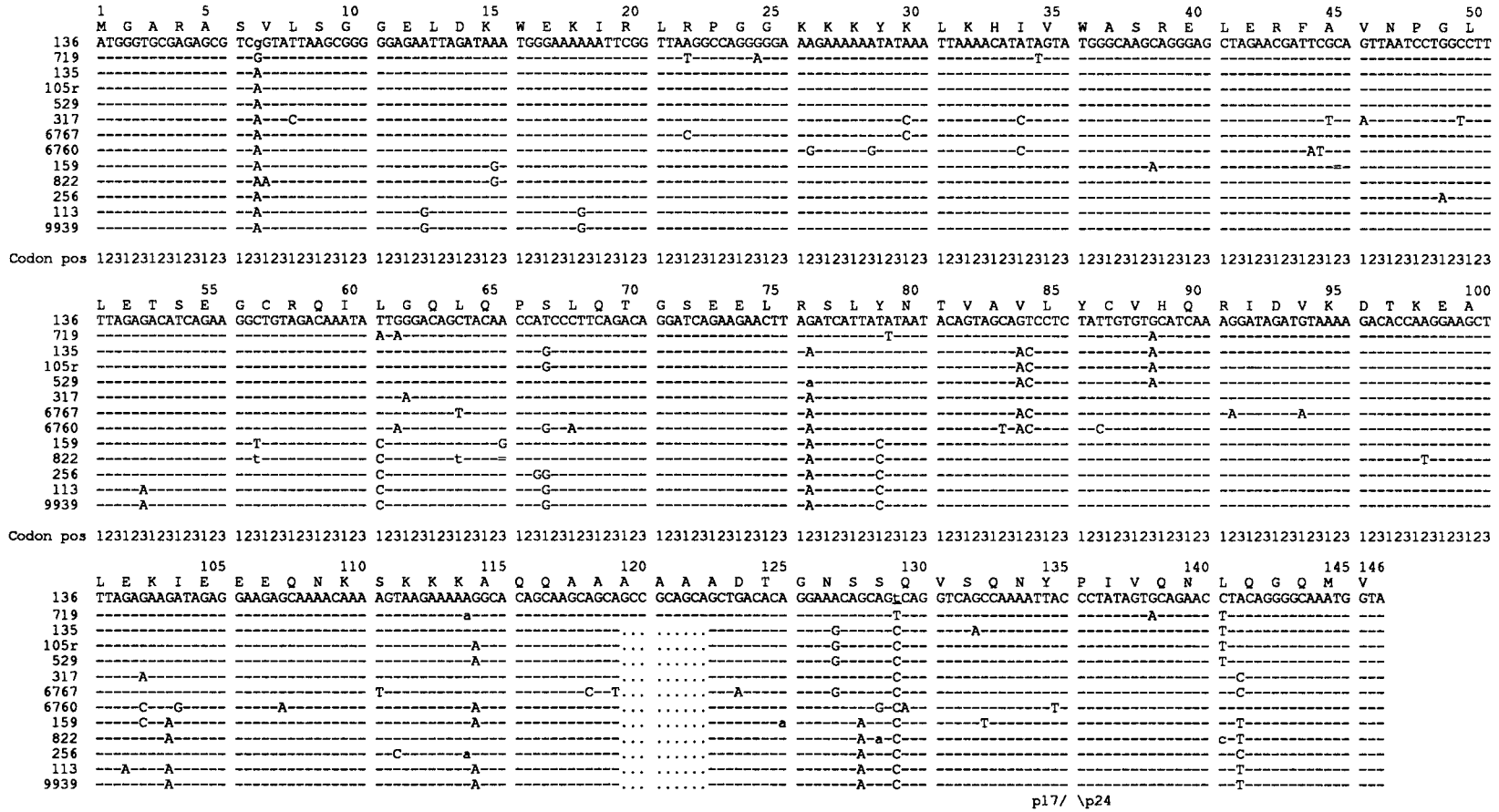

FIG. 4. Nucleotide alignment of the investigated *gag* fragment. The fragment includes the complete p17 part and the start of the p24 part of this gene, as indicated. Alignment strategy and symbols are as in the legend to Fig. 3.

TABLE 2. Effects of incorporation of the gamma distribution on different codon positions under the JC model measured with log likelihood values, assuming the true tree topology

Codon position	Uniform ($\alpha = \infty$)	Gamma		Uniform vs gamma, P^a
	l	α	l	
env V3				
1	−450.05	0.74	−441.53	3.7×10^{-5}
2	−385.94	0.32	−367.68	1.5×10^{-9}
3	−360.29	0.39	−350.81	1.3×10^{-5}
1+2	−849.12	0.49	−824.41	2.1×10^{-12}
p17*gag*				
1	−391.76	0.25	−383.47	4.6×10^{-5}
2	−312.47	0.18	−308.73	6.2×10^{-3}
3	−407.38	0.35	−400.72	2.6×10^{-4}
1+2	−715.02	0.20	−702.38	4.9×10^{-7}

[a] Gamma distribution was significantly better at the P level according to the χ^2 distribution with df = 1.

it is expected to produce more reliable estimates. However, this was not always possible because of the errors involved in the analysis of short sequences with use of a complex model. Therefore, we estimated α by an approximate method of Yang and Kumar (52) and conducted comparable maximum-likelihood computations under the JC model (16). The estimates of α obtained in this manner are given in Table 2. To test the null hypothesis that the sites are evolving with uniform rates, we conducted a likelihood ratio test (Table 2). Using the log likelihood ratio test, we concluded that the null hypothesis (i.e., that there is no nucleotide substitution rate variation across sites) was rejected at $<<1\%$ level for all codon positions in both V3 and p17. For the complete p17 fragment, the α value was estimated to be 0.257, and for the complete V3 fragment, α was estimated to be 0.384. Thus, nucleotide substitution methods which do not include rate variation across sites are inadequate descriptions of HIV-1 evolution.

Differences in substitution rates between nucleotides. Using the maximum-likelihood methods, we also compared the fits of different nucleotide substitution models to the V3 data. We used the REV model as a reference because it is the most general model available and since all simpler models are special cases of the REV model (see Materials and Methods). These comparisons allowed us to identify parameters that are of importance when one is estimating genetic distances and branch lengths of a phylogenetic tree. We used uniform (single-rate) as well as variable-rate models. In the latter case, the discrete gamma distribution was used. The results from these analyses are shown in Table 3. It was clear that the fit of the REV model was significantly better than that of all simpler models and that the fit of the REV model could be further improved by allowing gamma-distributed rates ($P < 10^{-17}$ with $\alpha = 0.384 \pm 0.082$). Clearly, the JC and K2 models provided the worst fit to the data, and the improvement in the fit of a model which accounts for unequal base frequencies (e.g., F81) was substantial. This means that a model that does not account for the nucleotide compositional bias in the HIV-1 genome is simply inadequate. When the fits of the JC and K2 models were compared, it became evident that also the differences in rates of transitions and transversions should be considered. Thus, the K2 model, which allows different rates for transitions and transversions, was significantly better than the simple JC model. Similarly, among models which account for differences in nucleotide frequencies (F81, F84, HKY, and TN), the models which also allow different rates for transitions and transversions (F84, HKY, and TN) were significantly better than the F81 model, which does not. There were no statistically significant differences in the fits of the F84, HKY, and TN models, which differ somewhat even though they all account for differences in both nucleotide frequencies and rates of transitions and transversions. The similar result from the F84 and HKY models meant that presetting the transition/transversion ratio to a suitable value was not inferior to estimating it for each comparison, as long as the value is fairly correct. Importantly, the relative fit of all models improved when the discrete gamma model was used ($P \approx 10^{-18}$). The estimates of the α values from these computations were quite similar to that obtained by using the REV model (see also reference 20).

The p17 fragment did not allow comparative studies of the different models because the maximum-likelihood computations did not converge when we used the known tree. However, by using the maximum-likelihood tree obtained by using DNAML (7), which differs slightly from the known tree (23), it was possible to carry out some computations. These results showed that it was important to account for unequal nucleotide frequencies and the transition/transversion rate bias also in p17 (results not shown).

Evolutionary patterns of HIV-1 *env* sequences. The estimated rate matrix (Q) for the whole V3 fragment, calculated by maximum likelihood assuming the REV model with variable rates across sites, is presented in Table 4. The transition/transversion rate bias averaged over base frequencies (R) was about 1.42. The estimate of R has been reported to be seriously underestimated when rate variation across sites is neglected, and the extent of this underestimation increases with the amount of rate variation among sites (49). However, in this material, the estimate of R was only slightly lower ($R = 1.37$) if a single rate (uniform) is assumed among sites. Large differences were seen among the substitution rates among nucle-

TABLE 3. Comparison of different substitution models and the effect of inclusion of a gamma distribution for the *env* V3 fragment measured with log likelihood calculations

Model	Uniform ($\alpha = \infty$)			Gamma				Uniform vs gamma, P
	l	df	P^a	α	l	df	P^a	
REV	−1,156.63			0.38380	−1,119.57			7.4×10^{-18}
TN	−1,164.12	2	5.5×10^{-4}	0.35912	−1,127.95	2	2.3×10^{-4}	1.8×10^{-17}
HKY	−1,165.70	3	4.1×10^{-4}	0.37471	−1,127.23	3	1.6×10^{-3}	1.8×10^{-18}
F84	−1,169.26	3	1.4×10^{-5}	0.35948	−1,128.58	3	4.3×10^{-4}	1.9×10^{-19}
F81	−1,187.55	4	1.2×10^{-12}	0.36018	−1,147.51	4	2.1×10^{-11}	3.6×10^{-19}
K2	−1,210.49	6	6.1×10^{-21}	0.37361	−1,173.29	6	7.0×10^{-21}	6.3×10^{-18}
JC	−1,227.45	7	2.2×10^{-27}	0.38575	−1,191.74	7	6.2×10^{-28}	2.9×10^{-17}

[a] Comparison between the tested model and the REV model, with the degrees of freedom indicated for each comparison.

TABLE 4. Estimate of the rate matrix Q for the *env* V3 fragment, using the REV model with a discrete gamma distribution to account for rate variation over sites[a]

From	To			
	T	C	A	G
T	−0.704671	0.348807	0.247731	0.108133
C	0.544886	−1.393152	0.790871	0.057395
A	0.123262	0.251903	−0.835056	0.459891
G	0.155810	0.052941	1.331818	−1.540570

[a] The element of the matrix, Q_{ij} ($i \neq j$), is the rate of substitution from nucleotide i to j. The matrix is scaled so that the average rate in equilibrium is 1.

otides. The largest substitution rate was found for G-to-A changes, and the V3 sequences were found to drift toward A richness by a factor of 1.5 from the other nucleotides, T, C, and G. A slight drift into T was also observed, balancing for the drift from C and G. The lowest rates were seen in both directions of transversions between C and G, while other transversional substitution rates (e.g., C to A) were more than 15 times higher. This may explain why the TN model did not provide significantly better fit than the F84 and HKY models and why the REV model fit significantly better.

As an alternative approach to estimate the different substitution types, substitution steps were counted according to the parsimony procedure and inverted into a substitution matrix. Also, this parsimony-based strategy recognizes that rates between different types of substitutions are highly uneven, thereby creating an asymmetrical weighting matrix (Table 5). The created weighting matrix down-weights the most common type of substitution and up-weights the most rare events. The resulting length of a branch thereby does not attempt to describe the number of events as the maximum-likelihood procedure used here does but rather described the weighted steps required along it.

The total amount of evolution, measured by the sum of estimated branch lengths in the true tree, was underestimated by about 12% if variable rates across sites were not accounted for in all of the examined models. The effect of allowing for rate variation among sites is greater on long branches than on short, making the inference more realistic since superimposed events are more likely to have occurred on sites with high rates after longer times. As a result, a distance of 0.1 calculated by the simple JC model may almost be doubled by using the REV

TABLE 5. Counts of substitution steps and the inverted weighting matrix for asymmetric parsimony weighting

From	To (substitution steps)				To (relative weighting matrix)[a]			
	T	C	A	G	T	C	A	G
env V3								
T		9	5	8		11	17	13
C	7		9	1	14		11	15
A	14	19		27	7	5		4
G	1	2	23		11	9	4	
p17^{gag}								
T		4	3	0		25	33	40
C	10		3	0	10		33	40
A	7	6		15	14	17		7
G	2	2	17		20	23	6	

[a] Corrected for triangle inequality.

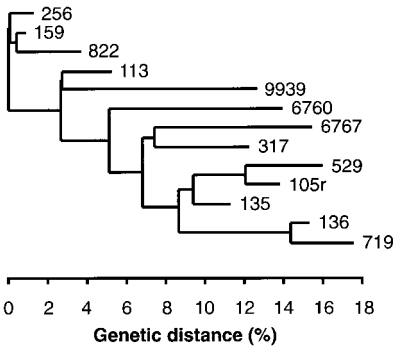

FIG. 5. Maximum-likelihood estimates of branch lengths for the *env* V3 sequences in the true topology of transmitted virus populations. A discrete gamma distribution model was used to account for rate variation over sites, while the substitution pattern was assumed to follow the REV model.

model with gamma-distributed rates. Figure 5 shows the inferred branch lengths in the true topology, using the V3 sequences and the REV model with variable rates across sites. The rate parameters were $\hat{a} = 0.822 \pm 0.257$, $\hat{b} = 0.186 \pm 0.059$, $\hat{c} = 0.235 \pm 0.097$, $\hat{d} = 0.594 \pm 0.149$, and $\hat{e} = 0.125 \pm 0.096$, and the shape parameter of the gamma distribution was $\alpha = 0.384 \pm 0.082$, with the standard errors of the estimates. When the branch lengths of the tree with estimated genetic distances (Fig. 5) was compared with the time distances in the true transmission history (Fig. 1), it became clear that not even the best model (i.e., the REV model with gamma-distributed rates across sites) was able to accurately recover the true branch lengths. In fact, although there was a large difference in absolute fit, the relative branch lengths calculated with all tested models, including both maximum-likelihood and maximum-parsimony approaches, had a correlation of about 0.7 to the true branch lengths measured in time. Branches expected to be short were generally estimated too long relative to the longer branches. Ancestral divergence, substitutional fluctuations over time, and changes in evolutionary rates in different lineages may explain this discrepancy. We are currently investigating these factors in HIV-1 evolution.

DISCUSSION

In this study, we investigated the fits of different models of nucleotide substitution to the true pattern of evolution of the *env* V3 and p17^{gag} fragments of the HIV-1 genome. Analyses of HIV nucleotide sequences are increasingly used to answer a variety of epidemiological questions ranging from macroscopic issues (e.g., the global spread of different genetic subtypes) to microscopic issues (e.g., forensic investigations of single transmission events). While there has been considerable debate about which genes are best suited to elucidate phylogenetic history and about the efficiency of different phylogenetic methods, little attention has been given to the problem of understanding the nucleotide substitution patterns. Because the use of inadequate models in phylogenetic inferences may lead to incorrect phylogenies, it is important to accurately describe the pattern of nucleotide substitution. Estimates of branch lengths are particularly sensitive to the choice of model, and evaluations of the reliability of an estimated tree may be misleading if oversimplified models are used.

In this study, we used a unique data set from a known HIV-1 transmission history. These data were previously used to investigate the abilities of different tree-building methods to reconstruct the true tree topology (23). We found that most

tree-building methods could accurately reconstruct the correct topology. In contrast, branch lengths did not correspond well to the amount of time elapsed on each branch. In the present study, we show that branch length estimates were inaccurate because oversimplified and unrealistic models of evolution were used. We show that the REV model with a gamma distribution is significantly more accurate than simpler nucleotide substitution models. It should be pointed out that the models which we found inaccurate have been used in virtually all previous molecular studies of HIV-1 and other viruses. Earlier attempts to correct for asymmetries in the substitution pattern of HIV-1 have applied weighted parsimony (11, 32), but the decisive parameters and rate variation across sites were not addressed. Thus, the REV model with gamma distribution has never previously been used in analyses of viral sequences. The main objective of all nucleotide substitution models is to correct for superimposed mutational events in individual nucleotide positions (i.e., multiple hit). The effects of these corrections will be moderate when closely related sequences are analyzed but dramatic when the genetic distances are larger (50–52). Here, an uncorrected distance of 0.10 will become almost twice as large (0.20) in the REV model with a gamma distribution, while it would increase to only about 0.11 in the simple JC model. Failures to correct for multiple hit will have particularly serious effects when attempts are made to extrapolate the age of distant viral ancestors from contemporary and closely related viral sequences. For instance, preliminary analyses indicate that the age of HIV-1 group M is likely to be pushed back considerably simply by the application of a more realistic model of evolution (i.e., the REV model with gamma distribution) (22a). It is likely that similar effects will be seen in analyses of other organisms, including other viruses.

Biologically, there are a number of reasons for differences in substitution rates among sites. First, we know that some characters cannot change without destroying the function of the protein, while changes in other positions may be favored. Thus, some positions will change rapidly and others will change more slowly. Second, alignments of HIV-1 sequences clearly show localized regions of high and low variability (32). For example, the beginning of the investigated V3 fragment displays considerably less variation than the region downstream of the second cysteine of the V3 loop (Fig. 3). This is also indicative of the differences in substitution rates among sites. In our study, we provide the first formal evidence of that there are significant differences in rates among sites ($\alpha < 1$).

Our p17 data contained too little information to allow estimation of the gamma distribution by the maximum-likelihood method. The shape parameter of the discrete gamma distribution (α) was therefore estimated by the YK approximate method (52), a method suggested by Sullivan et al. (41), and the traditional parsimony method (the method of moments). These methods are expected to overestimate α and thus underestimate rate variation among sites (52). This is also clear from the analysis of the V3 data. However, the YK method seems to give a reasonable estimate of α. By introducing the YK estimates of the α value in maximum-likelihood calculations, we found that the fit of different models to p17 also improved significantly.

The two HIV-1 genes analyzed have biased nucleotide compositions (Table 1), and clearly models that assume equality of base frequencies are unacceptable. As expected, the two models making this assumption (JC and K2) provided the worst fit to the data. By further also accounting for transition/transversion rate bias the fit of the model improved dramatically ($P = 10^{-25}$ for $2\Delta l_{\text{JC-HKY}} = 123.5$). The high rate of A-to-G transitions and the fact that large differences among transversion rates prevailed were taken into consideration only by the REV model, which therefore gave the best result. The next step of generalization of a substitution model is to remove the reversibility restriction, increasing the number of free parameters from 8 to 11. On short sequences, this model is therefore difficult to fit, since it needs more information on substitutions of all types. In an earlier study on primate mitochondrial DNA (mtDNA) sequences, the unrestricted model appeared not to be significantly better than the REV model (50).

Weighted parsimony has been reported to increase the accuracy in phylogenetic analyses when unequal rates between different types of substitutions prevail (11). As we have shown here, also this method picks up the asymmetries in the substitution pattern of HIV-1. The parsimony approach is simple and relatively fast; however, the method cannot include rate variation across sites. The weighting strategy of parsimony down-weights frequent types of changes, while in maximum-likelihood calculations, superimposed events are attempted to be corrected for so that distances estimate number of events.

Interestingly, in the above-mentioned mtDNA data set of primates, G-to-A substitutions were found to have the highest rate also (20, 50). Furthermore, Moriyama et al. (31) compared HIV mutations with those of nuclear pseudogenes and found high rates of substitutions from A to G of HIV, but this rate was also high in pseudogenes, as were rates of C-to-T transitions. While they concluded that the rates of substitutions between A and G were HIV specific, this finding suggests that preferred G-to-A mutation may not be an HIV-specific sign but rather something found in primate hosts. As a consequence of the similar substitution patterns of HIV, cellular pseudogenes, and primate mtDNA, the favored theory of a recent introduction of HIV into humans may not be supported by the G-to-A drift, which has been taken as a sign of an unadapted virus-to-host relationship. However, it does not provide evidence for the antithesis, that primate immunodeficiency viruses have evolved in step with their hosts, but rather shows that the issue still is open and needs further attention.

Some of the available computer programs for molecular evolutionary analyses and phylogenetic inference give researchers the possibility of including the findings from this study in their analyses. The shape parameter (α) of the gamma distribution estimated to be 0.38 and 0.25 for *env* V3 and p17^{gag}, respectively, can be used in, for instance, programs DNADIST in PHYLIP (under the Jin-Nei model), MEGA (21), and the new PAUP*. These programs, and many others including the widely used maximum-likelihood program DNAML, can also be given a transition/transversion rate ratio. When there is no transition/transversion rate bias, then the uncorrected mean rate ratio (κ) is 0.5. In *env* V3, we found the average transition/transversion rate ratio to be 1.42 when rate variation across sites was considered. However, when this value is used in DNAML, it will be corrected accordingly for the empirical purine and pyrimidine pools, since that program is based on the F84 model ($\kappa \neq \kappa_{F84}$, where κ_{F84} is the transition/transversion parameter [7], and here $\kappa_{F84} \approx 1.1$). The program mainly used in this study, PAML, can estimate these parameters under several models by maximum-likelihood iterations. Although the program is mainly constructed to test substitution models on a given tree, it can also search for a maximum-likelihood tree, but on comparative runs, DNAML appeared to be more effective in finding the correct topology (data not shown). These two programs use different search algorithms; however, neither algorithm performs an exhaustive tree search, and neither is guaranteed to find the best tree under the given criterion. Furthermore, it is possible that better estimates of topology can be obtained by using simple models. Subsequently, given the derived topology,

a complex and more realistic model can be used to calculate accurate estimates of branch lengths. However, it should be pointed out that not even the REV model with a gamma distribution succeeded in accurately reconstructing the true branch lengths of the known phylogeny. The apparent substitution rates may be influenced by rate differences in different lineages or individuals (9, 39) or by transmissions and other bottleneck effects (for instance, drug treatment). Preliminary data show that ancestral divergence and rapid substitution fluctuations also seem to influence the apparent rates and branch length estimates (to be published elsewhere).

In summary, we found that the REV model with gamma-distributed rates across sites was the best available description of the true nucleotide substitution pattern of HIV-1 *gag* and *env* gene fragments. Many characteristics of the genetic variation and evolution of HIV-1 could be accounted for by this model, thereby making phylogenetic inferences more realistic. The application of a more realistic substitution model will have a greater effect on estimates of genetic distances and branch lengths than on tree topologies. Thus, the use of these models is especially important when attempts are made to estimate the age of distant viral ancestors from contemporary viral sequences.

ACKNOWLEDGMENTS

This study was supported by grants from the Swedish Medical Research Council and the Swedish National Board for Industrial and Technical Development and by NSF and NIH grants to Masatoshi Nei.

REFERENCES

1. **Albert, J., J. Wahlberg, T. Leitner, D. Escanilla, and M. Uhlén.** 1994. Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 *pol* and *gag* genes. J. Virol. **68:**5918–5924.
2. **Baltimore, D.** 1970. RNA-dependent DNA polymerase in virions of RNA tumor viruses. Nature **226:**1209–1211.
3. **Berkhout, B., and F. J. van Hemert.** 1994. The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. Nucleic Acids Res. **22:**1705–1711.
4. **Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Menunier-Rotival, and F. Rodier.** 1985. The mosaic genome of warm-blooded vertebrates. Science **228:**953–958.
5. **Coffin, J.** 1992. Genetic diversity and evolution of retroviruses. Curr. Top. Microbiol. Immunol. **176:**143–164.
6. **Felsenstein, J.** 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:**368–376.
7. **Felsenstein, J.** 1993. PHYLIP: phylogeny inference package, 3.52c ed. University of Washington, Seattle, Wash.
8. **Gojobori, T., Y. Yamaguchi, K. Ikeo, and M. Mizokami.** 1994. Evolution of pathogenic viruses with special reference to the rates of synonymous and nonsynonymous substitutions. Jpn. J. Genet. **69:**481–488.
9. **Halapi, E., T. Leitner, M. Jansson, A. Plebani, G. Scarlatti, P. A. Tovo, P. Orlandi, J. Albert, H. Wigzell, and P. Rossi.** HIV-1 sequence evolution and specific immune response in children with distinct clinical courses. Submitted for publication.
10. **Hasegawa, M., H. Kishino, and T. Yano.** 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:**160–174.
11. **Hillis, D. M., J. P. Huelsenbeck, and C. W. Cunningham.** 1994. Application and accuracy of molecular phylogenies. Science **264:**671–677.
12. **Holland, J. J., J. C. De La Torre, and D. A. Steinhauer.** 1992. RNA virus populations as quasispecies. Curr. Top. Microbiol. Immunol. **176:**1–20.
13. **Hu, W. S., and H. M. Temin.** 1990. Retroviral recombination and reverse transcription. Science **250:**1227–1233.
14. **Ikemura, T., and S. Aota.** 1988. Global variation in G+C content along vertebrate genome DNA. Possible correlation with chromosome band structures. J. Mol. Biol. **203:**1–13.
15. **IUPAC-IUB Combined Commission on Biochemical Nomenclature.** 1966. Abbreviations and symbols for chemical names of special interest in biological chemistry, revised tentative rules. Biochemistry **5:**1445–1453.
16. **Jukes, T. H., and C. R. Cantor.** 1969. Evolution of protein molecules, p. 21–132. *In* H. N. Munro (ed.), Mammalian protein metabolism, vol. 3. Academic Press, New York, N.Y.
17. **Katz, R. A., and A. M. Skalka.** 1990. Generation of diversity in retroviruses. Annu. Rev. Genet. **24:**409–445.
18. **Kimura, M.** 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:**111–120.
19. **Korber, B. T. M., E. E. Allen, A. D. Farmer, and G. L. Myers.** 1995. Heterogeneity of HIV-1 and HIV-2. AIDS **9**(Suppl. A)**:**S5–S18.
20. **Kumar, S.** 1994. Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. Genetics **143:**537–548.
21. **Kumar, S., K. Tamura, and M. Nei.** 1993. MEGA: molecular evolutionary genetic analysis, 1.01 ed. The Pennsylvania State University, University Park, Pa.
22. **Leigh Brown, A., and P. Monaghan.** 1988. Evolution of the structural proteins of human immunodeficiency virus: selective constraints on nucleotide substitution. AIDS Res. Hum. Retroviruses **4:**399–407.

22a. **Leitner, T., and J. Albert.** Unpublished data.

23. **Leitner, T., D. Ecanilla, C. Franzén, M. Uhlén, and J. Albert.** 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc. Natl. Acad. Sci. USA **93:**10864–10869.
24. **Leitner, T., D. Escanilla, S. Marquina, J. Wahlberg, C. Broström, H. B. Hansson, M. Uhlén, and J. Albert.** 1995. Biological and molecular characterization of subtype D, G and A/D recombinant HIV-1 transmissions in Sweden. Virology **209:**136–146.
25. **Leitner, T., E. Halapi, G. Scarlatti, P. Rossi, J. Albert, E. M. Fenyö, and M. Uhlén.** 1993. Analysis of heterogeneous viral populations by direct DNA sequencing. BioTechniques **15:**120–126.
26. **Leonard, C. K., M. W. Spellman, L. Riddle, R. J. Harris, J. N. Thomas, and T. J. Gregory.** 1990. Assignment of intrachain disulfide bonds and characterization of potential glycosylation sites of the type I and recombinant human immunodeficiency virus envelope glycoprotein (gp 120) expressed in Chinese hamster ovary cells. J. Biol. Chem. **265:**10373–10381.
27. **Li, W.-H., M. Tanimura, and P. M. Sharp.** 1988. Rates and dates of divergence between AIDS virus nucleotide sequences. Mol. Biol. Evol. **5:**313–330.
28. **Louwagie, J., F. E. McCutchan, M. Peeters, T. P. Brennan, B. E. Sanders, G. A. Eddy, G. van der Groen, K. Fransen, G.-M. Gershy-Damet, R. Deleys, and D. S. Burke.** 1993. Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. AIDS **7:**769–780.
29. **Maddison, W. P., and D. R. Maddison.** 1992. MacClade: analysis of phylogeny and character evolution, 3.06 ed. Sinauer, Sunderland, Mass.
30. **Mansky, L. M., and H. M. Temin.** 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J. Virol. **69:**5087–5094.
31. **Moriyama, E. N., Y. Ina, K. Ikeo, N. Shimizu, and T. Gojobori.** 1991. Mutation pattern of human immunodeficiency virus genes. J. Mol. Evol. **32:**360–363.
32. **Myers, G., B. Korber, B. H. Hahn, K.-T. Jeang, J. W. Mellors, F. E. McCutchan, L. E. Henderson, and G. N. Pavlakis.** 1995. Human retroviruses and AIDS: a compilation and analysis of nucleic acid and amino acid sequences. Los Alamos National Laboratory, Los Alamos, N.Mex.
33. **Myers, G., K. MacInnes, and B. Korber.** 1992. The emergence of simian/human immunodeficiency viruses. AIDS Res. Hum. Retroviruses **8:**373–386.
34. **Ou, C. Y., C. A. Ciesielski, G. Myers, C. I. Bandea, C.-C. Luo, B. T. M. Korber, J. I. Mullins, G. Schochetman, R. L. Berkelman, A. N. Economou, J. J. Witte, L. J. Furman, G. A. Satten, K. A. MacInnes, J. W. Curran, H. W. Jaffe, Laboratory Investigation Group, and Epidemiologic Investigation Group.** 1992. Molecular epidemiology of HIV transmission in a dental practice. Science **256:**1165–1171.
35. **Robertson, D. L., P. M. Sharp, F. E. McCutchan, and B. H. Hahn.** 1995. Recombination in HIV-1. Nature **374:**124–126.
36. **Scarlatti, G., T. Leitner, E. Halapi, J. Wahlberg, P. Marchisio, M. A. Clerici-Schoeller, H. Wigzell, E. M. Fenyö, J. Albert, M. Uhlén, and P. Rossi.** 1993. Comparison of variable region 3 sequences of human immunodeficiency virus type 1 from infected children with the RNA and DNA sequences of the virus populations of their mothers. Proc. Natl. Acad. Sci. USA **90:**1721–1725.
37. **Shpaer, E. G., and J. I. Mullins.** 1993. Rates of amino acid change in the envelope protein correlate with pathogenicity of primate lentiviruses. J. Mol. Evol. **37:**57–65.
38. **Starcich, B. R., B. H. Hahn, G. M. Shaw, P. D. McNeely, S. Modrow, H. Wolf, E. S. Parks, W. P. Parks, S. F. Josephs, R. C. Gallo, and F. Wong-Staal.** 1986. Identification and characterization of conserved and variable regions in the envelope gene of HTLVIII/LAV, the retrovirus of AIDS. Cell **45:**637–648.
39. **Strunnikova, N., S. C. Ray, R. A. Livingston, E. Rubalcaba, and R. P. Viscidi.** 1995. Convergent evolution within the V3 loop domain of human immunodeficiency virus type 1 in association with disease progression. J. Virol. **69:**7548–7558.
40. **Sueka, N.** 1959. Dependence of the density of deoxyribonucleic acids on guanine-cytosine content. Nature **183:**1429–1431.
41. **Sullivan, J., K. E. Holsinger, and C. Simon.** 1996. Among-site rate variation and phylogenetic analysis of 12S rRNA in Sigmontine rodents. Mol. Biol. Evol. **12:**988–1001.
42. **Swofford, D. L.** 1991. PAUP: phylogenetic analysis using parsimony, 3.1.1 ed. Illinois Natural History Survey, Champaign, Ill.
43. **Tamura, K., and M. Nei.** 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and

chimpanzees. Mol. Biol. Evol. **10:**512–526.
44. **Tavare, S.** 1986. Some probabilistic and statistical problems on the analysis of DNA sequences, p. 57–86. *In* Lectures in mathematics in the life sciences, vol. 17.
45. **Temin, H. M.** 1993. Retrovirus variation and reverse transcription: abnormal strand transfers result in genetic variation. Proc. Natl. Acad. Sci. USA **90:** 6900–6903.
46. **Temin, H. M., and S. Mitzutani.** 1970. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature **226:**1211–1213.
47. **Vartanian, J.-P., A. Meyerhans, M. Sala, and S. Wain-Hobson.** 1994. G-A hypermutation of the human immunodeficiency virus type 1 genome: evidence for dCTP pool imbalance during reverse transcription. Proc. Natl. Acad. Sci. USA **91:**3092–3096.
48. **Vartanian, J.-P., A. Meyerhans, B. Åsjö, and S. Wain-Hobson.** 1991. Selection, recombination, and A-G hypermutation of human immunodeficiency virus type 1 genomes. J. Virol. **65:**1779–1788.
49. **Wakely, J.** 1994. Substitution-rate variation amoung sites and the estimation of transition bias. Mol. Biol. Evol. **11:**436–442.
50. **Yang, Z.** 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39:**105–111.
51. **Yang, Z.** 1995. PAML: phylogenetic analysis by maximum likelihood, 1.1 ed. Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, Philadelphia, Pa.
52. **Yang, Z., and S. Kumar.** 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. **13:**650–659.
53. **Yokoyama, S., L. Chung, and T. Gojobori.** 1988. Molecular evolution of the human immunodeficiency and related virus. Mol. Biol. Evol. **5:**237–251.