JOURNAL OF VIROLOGY, Dec. 1990, p. 6221–6233
0022-538X/90/126221-13$02.00/0

Vol. 64, No. 12

Concurrent Evolution of Human Immunodeficiency Virus Type 1 in Patients Infected from the Same Source: Rate of Sequence Change and Low Frequency of Inactivating Mutations

PETER BALFE,[1] PETER SIMMONDS,[1] CHRISTOPHER A. LUDLAM,[2] JOHN O. BISHOP,[1,3] AND ANDREW J. LEIGH BROWN[1]*

Department of Genetics, University of Edinburgh, King's Buildings, Edinburgh EH9 3JN,[1] and Department of Haematology, Royal Infirmary, Edinburgh EH3 9YW,[2] Scotland, and Department of Biological Sciences, University of Maryland Baltimore Campus, Baltimore, Maryland 21228[3]

Received 11 July 1990/Accepted 19 September 1990

Direct sequencing of segments of the envelope gene of human immunodeficiency virus type 1 proviruses in peripheral blood mononuclear cells has revealed that a cohort of hemophiliacs who were infected after exposure to a single common batch of factor VIII share closely related virus strains. Seventy-four sequences extending from hypervariable regions V4 through V5 from nine patients yielded a mean intrapatient nucleotide distance of 5.5%, while a mean of 4.2% was observed in 39 sequences of the V3 loop (six patients). Phylogenetic analysis revealed that sequences of six Edinburgh patients were particularly closely related and those from a patient infected in the United States were very distinct. The mean nucleotide distance among these six was 8.3%, while the mean distance from the U.S.-derived sequences was 25.5% in the V4-V5 region. The rate of sequence change across this patient group has been estimated to be 0.4% per year in the V4-V5 region and 0.5% per year in the V3 region, with at least a twofold range across patients. Only two inactivating nucleotide substitutions have been observed in a total of 42 kb of sequence obtained from the *env* and *gag* genes during this study.

Human immunodeficiency virus type 1 (HIV-1) is characterized by extensive genomic heterogeneity. Independent isolates from North America differ in almost 10% of amino acid residues in *env*, while isolates from Africa can differ in up to 25% of these sites (1, 30). In common with visna virus and equine infectious anemia virus (3, 23), HIV-1 also shows variability within infected individuals. Genetic differences have been observed among consecutive isolates and, concurrently, within the same isolate established from the same patient (9, 26). It has previously been suggested in the case of the other lentiviruses that sequence change may be immunologically driven, natural selection favoring the spread of variants that escape from immune surveillance during the course of the infection (2, 20). If sequence changes in HIV occur under the influence of selection by the host immune response, their identity and rate of occurrence will be important for the effort to develop an effective vaccine.

Previous attempts to estimate the rate of evolution of the HIV-1 genome have been based on analyses of sequence data obtained from viral isolates made at different times. One group (29) suggested from an analysis of *env* sequences that HIV-1 and HIV-2 may have diverged as recently as 40 years ago. Others (34; P. M. Sharp and W.-H. Li, Nature [London] **336**:315, 1988), using data on *gag* and *pol* as well, suggested that the two viruses diverged from 150 to 300 years ago. Hahn et al. (9) obtained three isolates from the same patient over an 18-month period and analyzed the *env* gene sequences obtained in a similar way. Their estimate of the rate of sequence change is compatible with all of these.

We have been studying HIV sequence change in a group of patients who became infected during 1984 after exposure to factor VIII concentrate prepared from donated blood by the Scottish National Blood Transfusion Service (SNBTS). A single batch of factor VIII was implicated in the seroconversion of these patients (17). Of a total of 32 patients exposed to the batch, 18 seroconverted. These patients received, on average, significantly more units of the putative infected batch than those who did not (17). This batch was prepared from plasma collected at a time when HIV-1 was only beginning to enter the donor population, as indicated by the low incidence of anti-HIV-1-positive sera among samples collected at that time (24). For both these reasons it has been considered that very few HIV-infected individuals could have contributed to the infected plasma pool.

We have analyzed HIV sequences in samples donated in the summer of 1989 by eight members of this cohort and one other HIV-infected hemophiliac. We have used a double polymerase chain reaction (PCR) procedure to obtain the nucleotide sequence of individual provirus molecules (27). This approach is not subject to any bias due to virus culture in vitro nor to possible errors introduced by *Taq* polymerase (31). Between 5 and 14 sequences from each patient studied were obtained from one region of *gag* encoding part of p24 and from two regions of *env*; one, V4-V5, which includes the CD4-binding domain (4, 12) and the flanking hypervariable regions (V4 and V5; 19), and another which spans the immunodominant or V3 loop (19, 25). We have investigated the evolutionary relationships among sequences from different patients and found that they are much more closely related to each other than are sequences from independent isolates or sequences obtained from the noncohort control. We will present evidence that at least six of the cohort patients were infected by virus originating from a single individual donor. On this basis we have estimated the extent of divergence in these regions of the HIV-1 genome among this group of patients and the rate of sequence change.

* Corresponding author.

MATERIALS AND METHODS

Patients studied. Heparinized whole blood (20 ml) was obtained from each of eight patients treated solely with SNBTS-prepared factor VIII who seroconverted during 1984 and from 1 hemophiliac infected in the United States. The circumstances of infection and details of the HIV-1 antigen and antibody status of the cohort have been described previously (17, 28). The current clinical status is described elsewhere (R. Cuthbert, C. A. Ludlam, C. M. Steel, J. Tucker, D. Beatson, S. Rebus, and J. F. Peutherer, Br. Med. J., in press).

HIV primers. Oligonucleotides were synthesized by the Oswel DNA Service, Department of Chemistry, University of Edinburgh, and were purified by high-pressure liquid chromatography. The sequences of the primers are given, along with their positions in HIV HXB2 in parentheses (+, sense; −, antisense). V3: -a-, TACAATGTACACATGGA ATT (+, 6957); -b-, TGGCAGTCTAGCAGAAGAAG (+, 7009); -c-, CTGGGTCCCCTCCTGAGG (−, 7331); and -d-, ATTACAGTAGAAAAATTCCCC (−, 7381). V4-V5: -e-, TCAGGAGGGGACCCAGAAATT (+, 7316); -f-, GGGG AATTTTTCTACTGTAAT (+, 7360); -g-, CTTCTCCAATT GTCCCTCATA (−, 7665); and -h-, CCATAGTGCTTCCTG CTGCT (−, 7814).

PCR amplification. DNA samples were prepared from at least 5×10^6 peripheral blood mononuclear cells (PMBCs). DNA preparation, amplification by double PCR, and quantification by limiting dilution were carried as described previously (27). All DNA extractions and amplification reactions carried appropriate negative controls (blood from seronegative, low-risk group blood donors) to detect contamination at every stage in the procedure. To permit the same sample to be amplified in two regions, the first reaction used the positive-sense outer V3 primer -a- with the antisense outer V4-V5 primer -h- to amplify an 858-bp fragment. The second reaction amplified an aliquot of the product using the inner V3 sense -b- and inner V4-V5 antisense -g- primers to give a 657-bp fragment. The final amplifications used the inner V3 and V4-V5 primers in separate reactions. V3 and V4-V5 DNAs amplified by this modified procedure are equivalent to those prepared by conventional amplification of separate samples by the double PCR.

Determination of sequence. The double PCR procedure allows the amplification of segments of single isolated molecules of HIV-1 provirus to yield quantities sufficient for direct sequencing (>500 ng of DNA). The DNA product of PCR amplification was purified by using Gene-Clean (Bio 101 Inc.). From 50 μl of PCR amplification a yield of 500 ng was routinely achieved, of which 100 ng was used for each set of sequencing reactions. Sequencing was carried out by using Sequenase protocols (U.S. Biochemicals) except as follows. The standard annealing mix was adjusted to 10% dimethyl sulfoxide (33), denatured by boiling (5 min), and chilled to 0°C. The labeling reaction was performed at room temperature for 5 min with only two of the usual three unlabeled dNTPs present at 0.25 μM and 5 μCi of [^{35}S]dATP or [^{35}S]dCTP (1,000 Ci/mM) (32). The choice of labeled nucleotide was based on nucleotide usage 3′ to the priming site used. Termination reaction mixtures were adjusted to 10% dimethyl sulfoxide, and the termination was performed at 37°C for 5 min. The sequences were analyzed on 6% wedge sequencing gels.

Sequences were collected and aligned using the University of Wisconsin GCG package (5) and the Needleman and Wunsch alignment algorithm (21). Final alignments were adjusted by eye, and their quality was assessed by the reduction of the total branch lengths of the phylograms generated under the assumption of maximum parsimony.

Because of the volume of data available it was not practicable to perform a maximum likelihood analysis of the sequence data itself. The data were therefore transformed into a distance matrix, using a modification of the distance algorithm of Kimura (10) which takes into account the base composition of the DNA under study (11). Comparison of results obtained by maximum likelihood analysis of subsets of the data showed good agreement with those given by using the distance matrix approach.

Phylogenetic trees were generated from the distance matrix by the method of Fitch and Margoliash (7), using the Phylip package (6) distributed by J. Felsenstein. Final tree topologies were permuted by using a branch swapping algorithm to minimize overall length.

Nucleotide sequence accession number. The sequences presented in this report have been submitted to GenBank under accession number M36997.

RESULTS

Nucleotide sequences from PBMCs. In Fig. 1 we present 74 nucleotide sequences of the V4-V5 region obtained directly from double PCR-amplified PBMC DNAs of eight members of the Edinburgh hemophiliac cohort and one control patient, also a hemophiliac, who was infected in the United States. We include for comparison published sequences from eight independent HIV isolates. In Fig. 2, 39 sequences of the V3 region obtained similarly from five of the cohort members and the control patient are given. Each sequence represents an individual provirus molecule, isolated by limit dilution of the PBMC DNA. Because it is possible to amplify more than one region of the same molecule using this procedure, in many cases we have been able to obtain sequences of both V4-V5 and V3 from the same provirus. The sequences were aligned by the introduction of gaps as necessary. As the positions and lengths of gaps were constrained to preserve the reading frame, the alignment was initially carried out on the deduced amino acid sequences.

Absence of inactivating substitutions. The sequences of the V4-V5 region exhibit extensive variation due both to nucleotide substitutions and to insertion-deletion events, as indicated by the need to introduce a large number of gaps in the alignment (Fig. 1). Despite this, only two inactivating substitutions were observed in the 74 sequences (each approximately 300 bp in length). There are no inactivating mutations among the 39 V3 region sequences shown in Fig. 2 (325 bp in length). We have also determined 33 sequences of a 220-bp segment of the *gag* (p24) gene (coordinates 1407 to 1646 in the HXB2 sequence) from five cohort and three noncohort patients (data submitted to Los Alamos AIDS data base). No inactivating mutations were observed in these sequences. In this data set, therefore, the frequency of inactivating substitutions is approximately 1 in 20 kb.

These results suggest that in these regions of *gag* and *env* there is strong selection for maintenance of a continuous reading frame. The strength of selection favoring the existing amino acid sequence is indicated by the ratio of synonymous to nonsynonymous substitutions (13, 14), which we have estimated by using the procedure of Li et al. (16). For *gag* the ratio is 6.7, indicating that the amino acid sequence is under significant constraints. For the V4-V5 and V3 regions the ratios are 1.3 and 0.92, respectively. These extremely

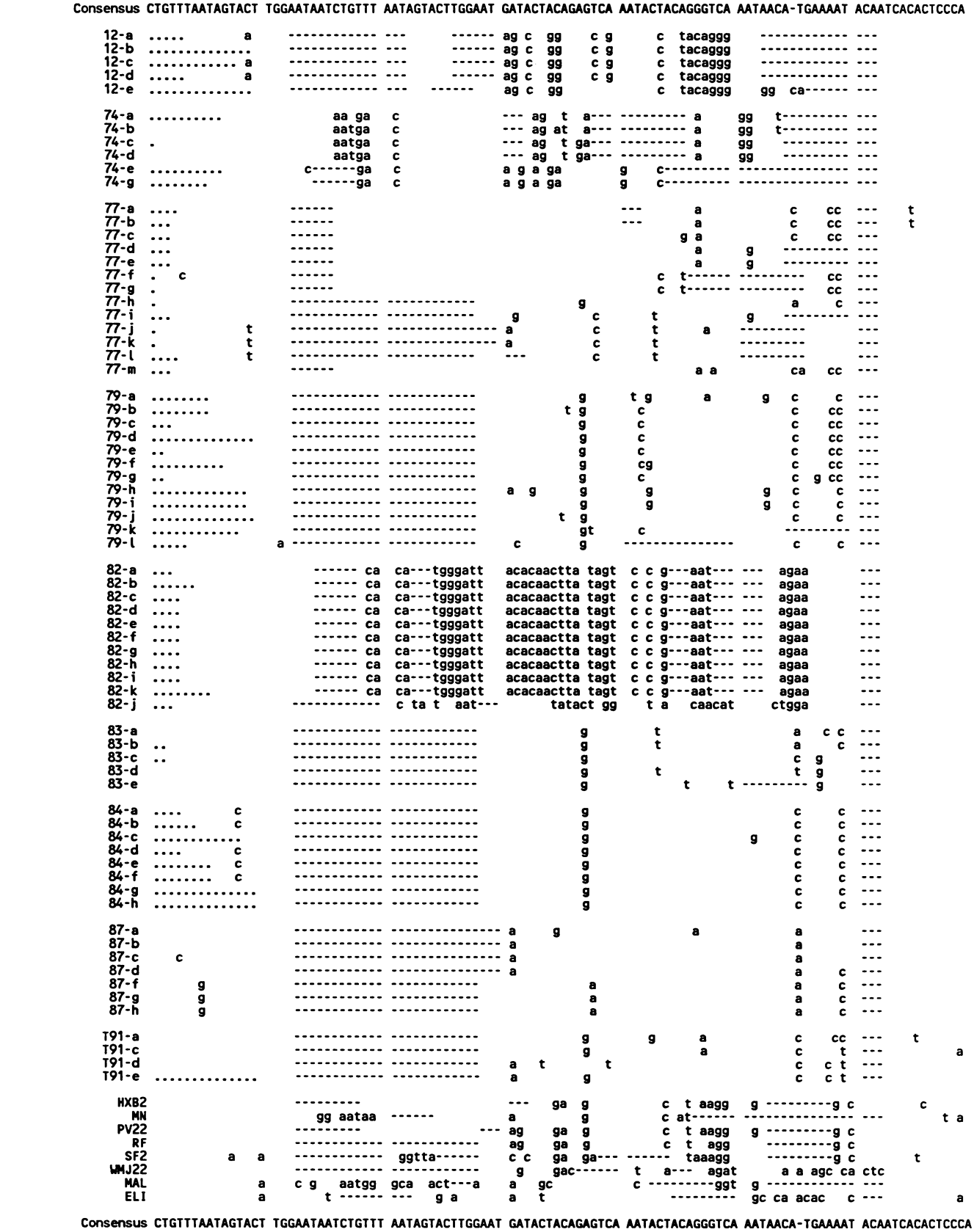

FIG. 1. Nucleotide sequence of the V4-V5 region of gp120 in 74 HIV-1 provirus molecules from Edinburgh hemophiliacs. Sequences are grouped by patient number and an identifying letter. Sequences of the equivalent region from eight independent HIV isolates obtained from the Los Alamos AIDS data base are shown below for comparison. Gaps inserted during alignment are indicated by dashes. Regions for which the sequence has not been determined for a particular provirus are shown by dots. See Materials and Methods for experimental procedures. Sequence continues on the following pages.

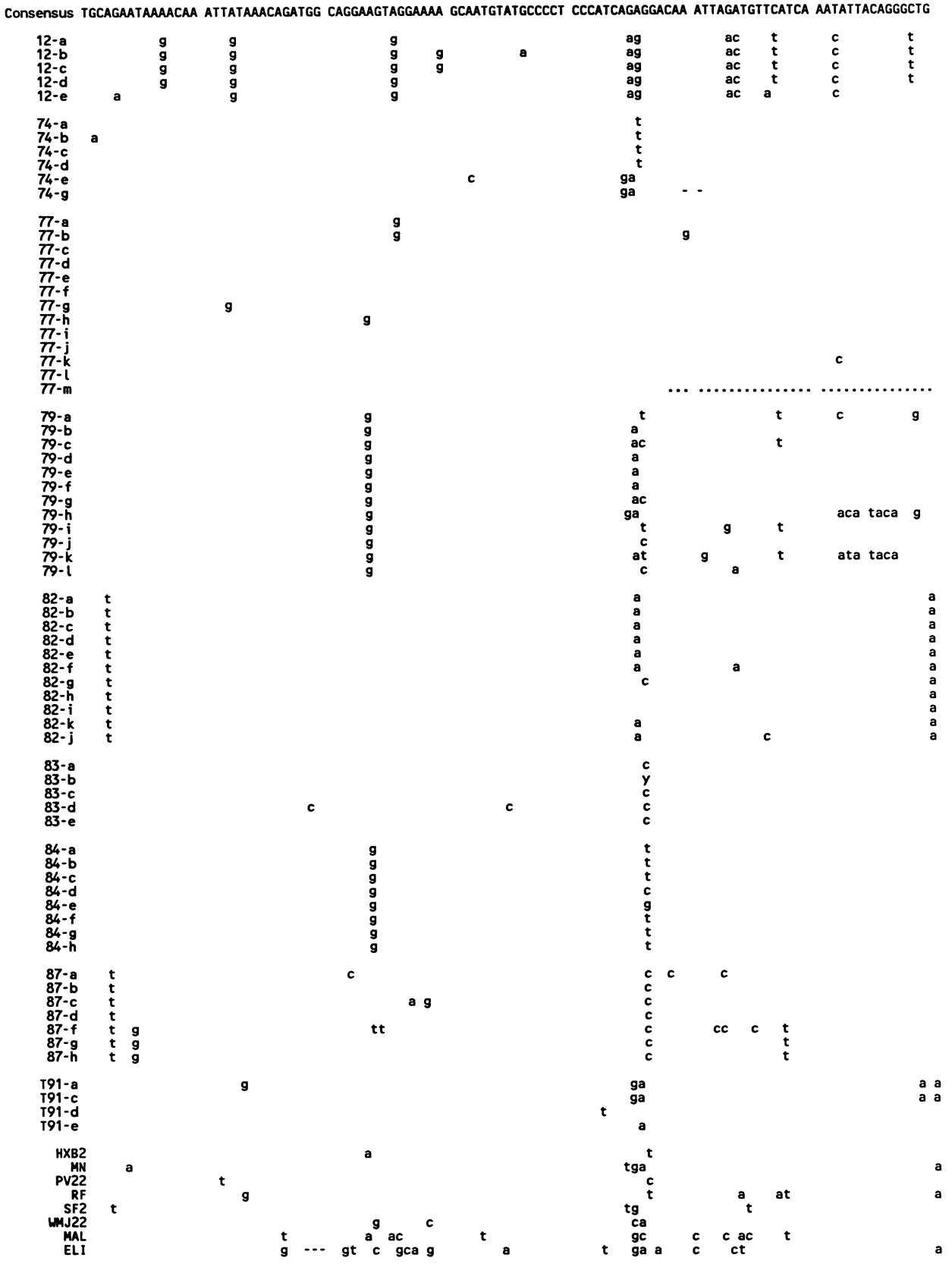

FIG. 1—*Continued.*

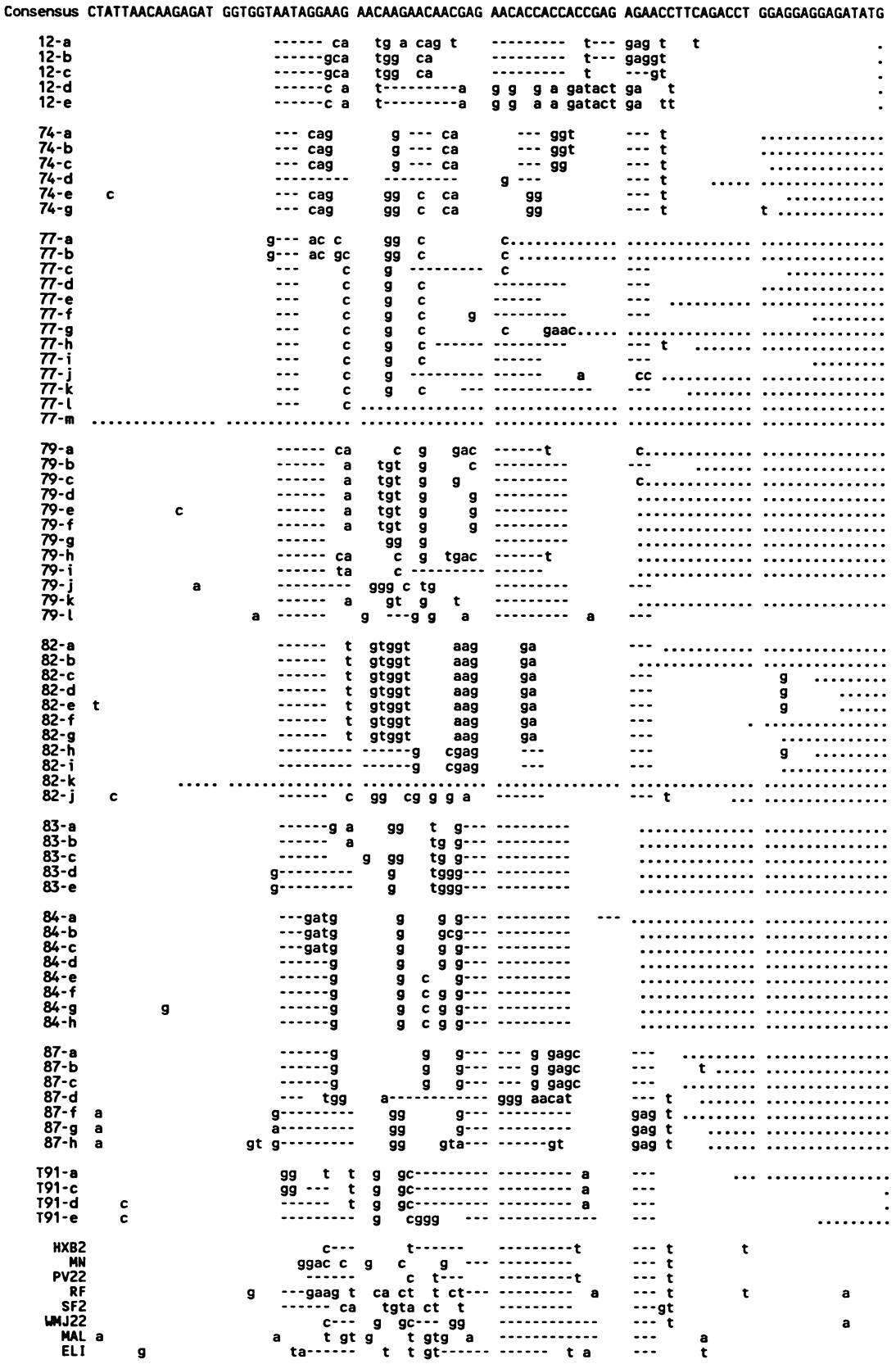

FIG. 1—*Continued.*

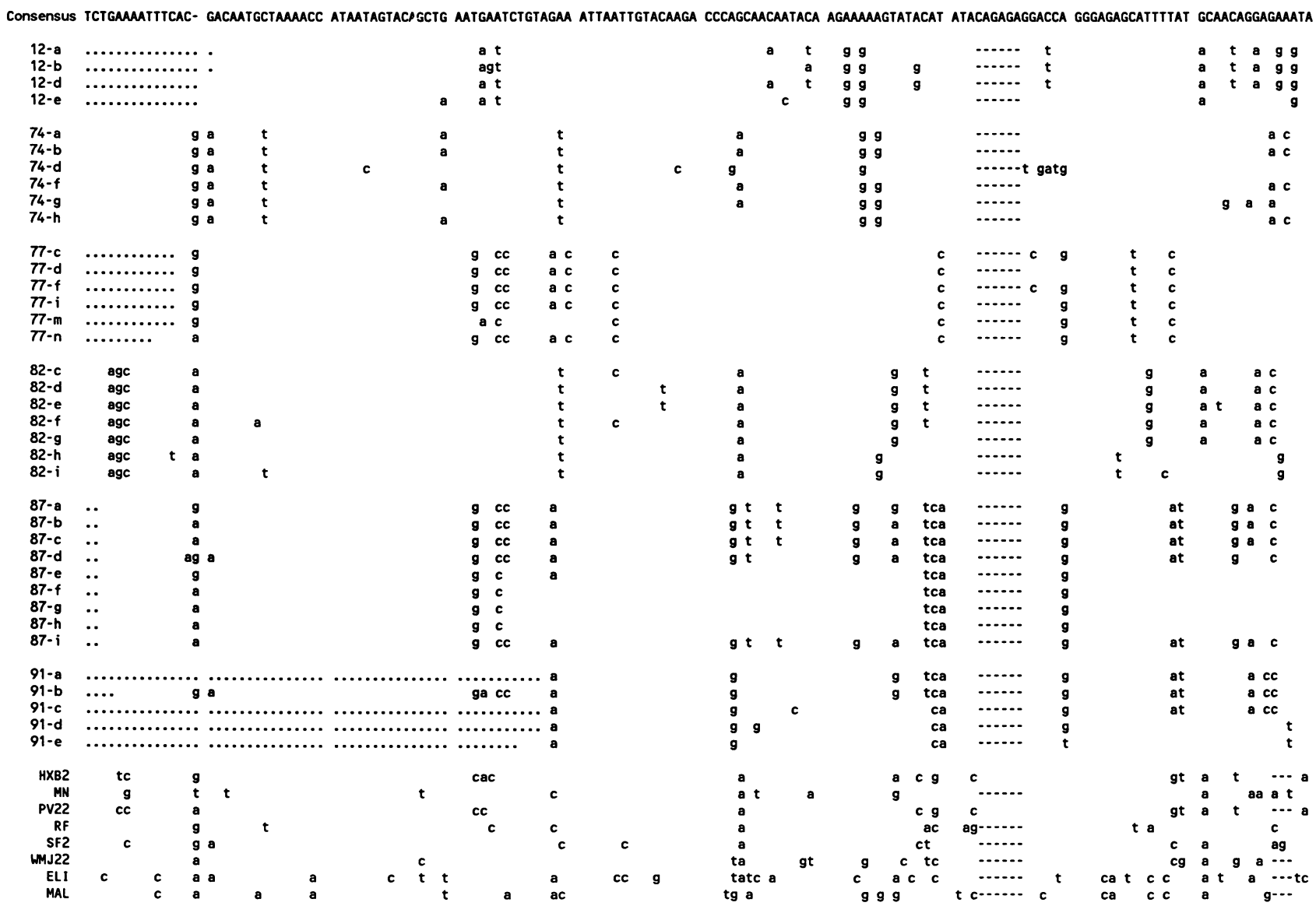

FIG. 2. Nucleotide sequence of the V3 region of gp120 in 39 HIV provirus molecules from Edinburgh hemophiliacs. Sequences are identified as described in the legend to Fig. 1. The same number-letter combination indicates that the sequences were obtained from the same provirus molecule.

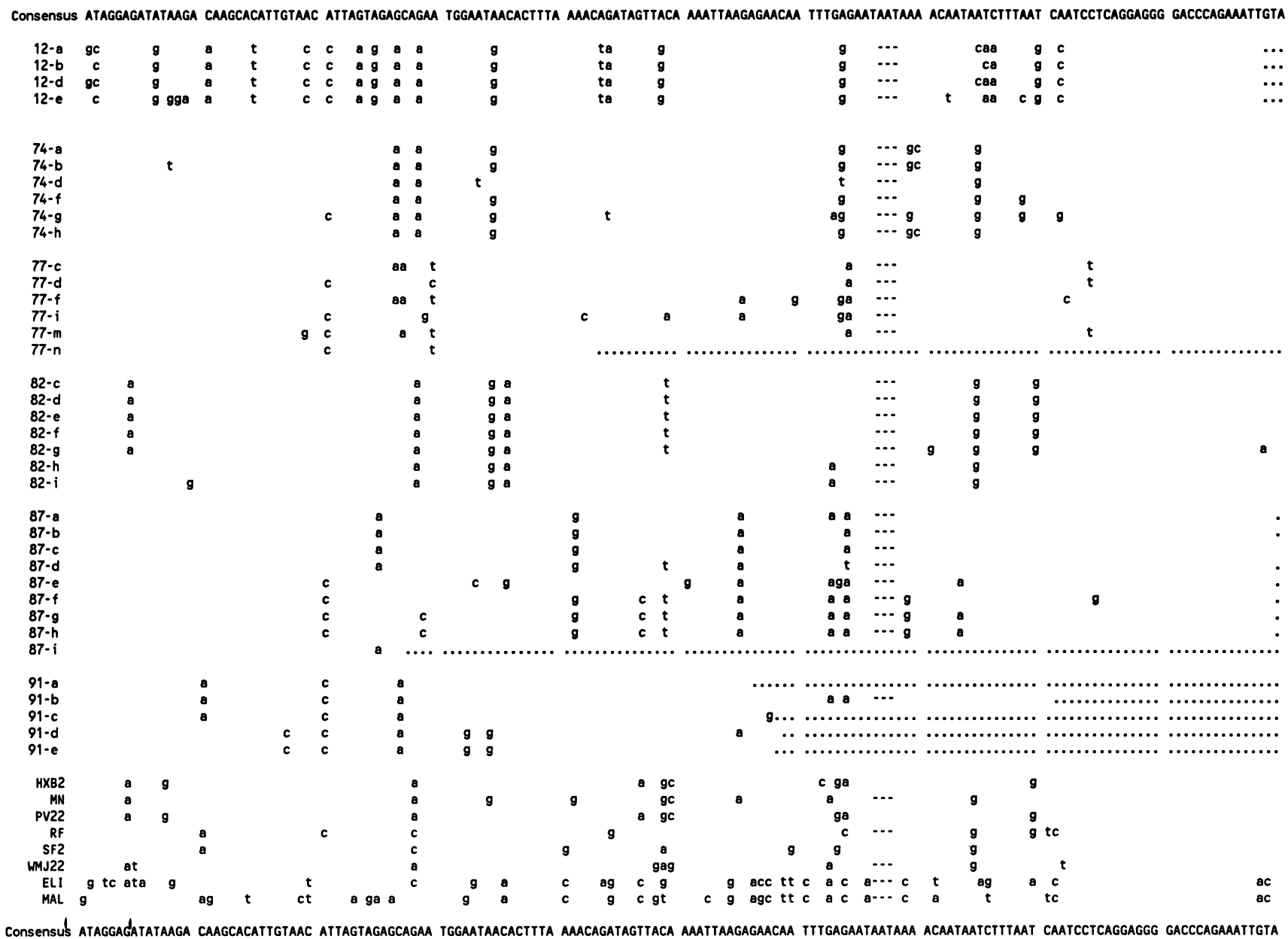

FIG. 2—*Continued*.

low ratios indicate that the amino acid as well as the nucleotide sequence of gp120 is evolving very rapidly (13). The limit to the rate of sequence evolution appears to be the requirement to maintain an intact reading frame.

We have calculated the overall frequency of transitions and transversions for all pairwise comparisons among these sequences. The ratio of transition to transversion substitutions was 1.79 for the V4-V5 region and 1.50 in the *gag* sequence, comparable to ratios observed for mammalian genes (22).

Evolutionary relationships among HIV sequences. Examination of Fig. 1 reveals, as might be expected, that sequences from the same patients are often more similar to each other than are sequences from different patients. In order to quantify the extent of evolutionary divergence between sequences we have estimated nucleotide distances from the observed percent nucleotide sequence identity, using the model of Kimura (10), which allows for different rates of transition and transversion substitutions. The mean intrapatient and interpatient nucleotide distances have then been used to estimate the phylogenetic relationships among all 74 V4-V5 and 39 V3 sequences (see Materials and Methods). Representative maximum parsimony trees depicting these relationships are given in Fig. 3 (V4-V5) and 4 (V3).

Several points emerged directly from the phylogenetic analysis. (i) For both regions, sequences from any particular patient group most closely with other sequences from the same patient. (ii) The sequences from the noncohort control (patient 12) are unrelated to those of any other patient and group more closely with those from North American isolates. (iii) Sequences from six patients (no. 77, 79, 83, 84, 87, and 91) group together very closely and in some cases are interspersed. (iv) Among the Edinburgh sequences, those from patients 74 and 82 are much less closely related to those from the other patients in both regions of *env*. Patient 82 is as distant from all other Edinburgh sequences in V4-V5 as the U.S.-infected control (no. 12). On the other hand, sequences from patient 74 which cluster together with those from the central six Edinburgh cohort samples (see above) in V4-V5 do not do so in V3.

Nucleotide distances of HIV sequences within and between patients. We estimated the evolutionary divergence for all pairwise comparisons among the sequences in Fig. 1 and 2. We then grouped the estimates into (i) comparisons between sequences derived from a single patient and (ii) comparisons between sequences taken from each pairwise combination of patients, and we examined the distribution of distances for each. For most groups the distribution of estimates was relatively narrow and unimodal (data not shown), but for some, specifically for patient 74 and for the group of HIV isolates, a wider spread was observed. The group means are given in Tables 1 and 2.

The mean nucleotide distance in V4-V5 between HIV isolates (17.0%) was considerably higher than that for the intrapatient comparisons. The mean intrapatient distance among all patients was 5.5% (range 1 to 9.2%). Among the Edinburgh patients, the mean nucleotide distances of several interpatient pairs, e.g., 5.6% for patients 83 and 84, were within this range. However, the U.S.-infected control (patient 12) had very distinct HIV sequences, as did patient 82. In V3, for which the mean distance within patients was similar (4.2%), all the nucleotide distances between patients were much greater. It should be noted that our estimates of divergence are based solely on the extent of nucleotide substitution. In regions such as V4, for which DNA insertion and deletion events are also frequent, these will be underestimates of the overall mutation distance.

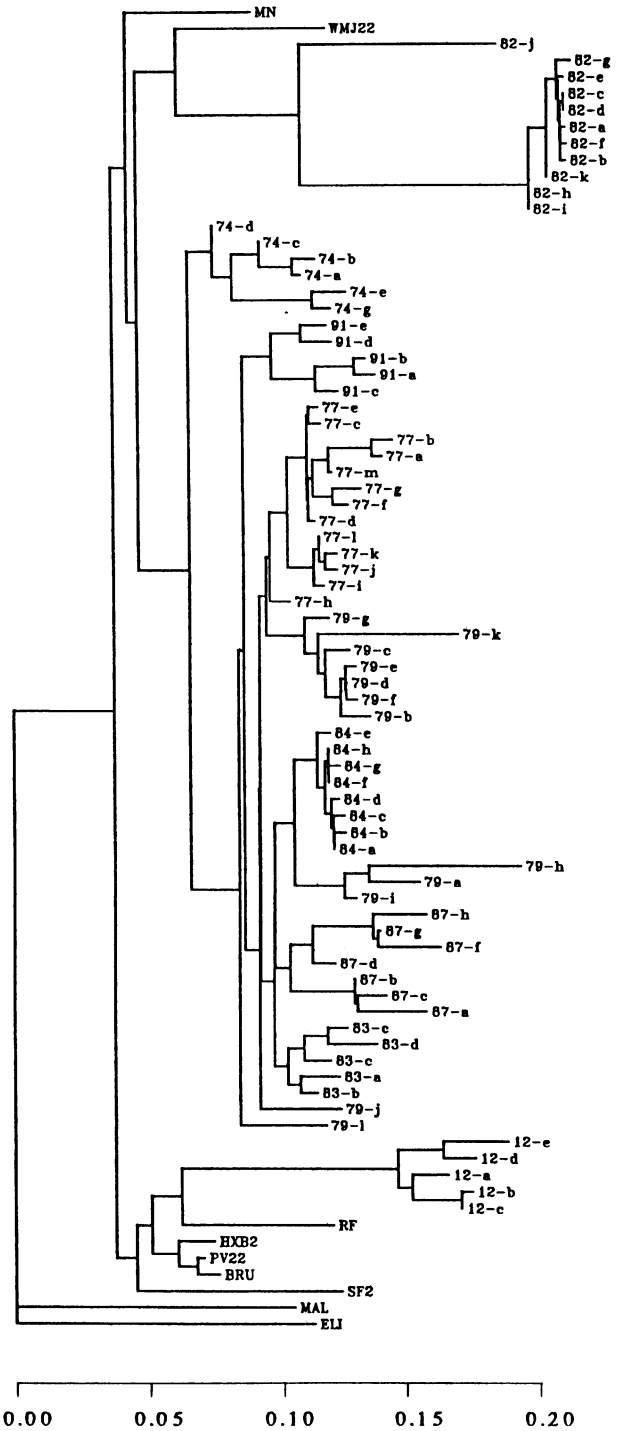

FIG. 3. Phylogenetic relationships of the V4-V5 region of the HIV-1 *env* gene from Edinburgh hemophiliacs. The tree was obtained by using a maximum parsimony procedure on the basis of the nucleotide distances summarized in Table 1. The scale bar shows the proportion of nucleotides substituted for a given horizontal branch length. See Materials and Methods for procedures.

In Table 3 we present estimates obtained in a similar fashion for the region sequenced in *gag*. As expected from

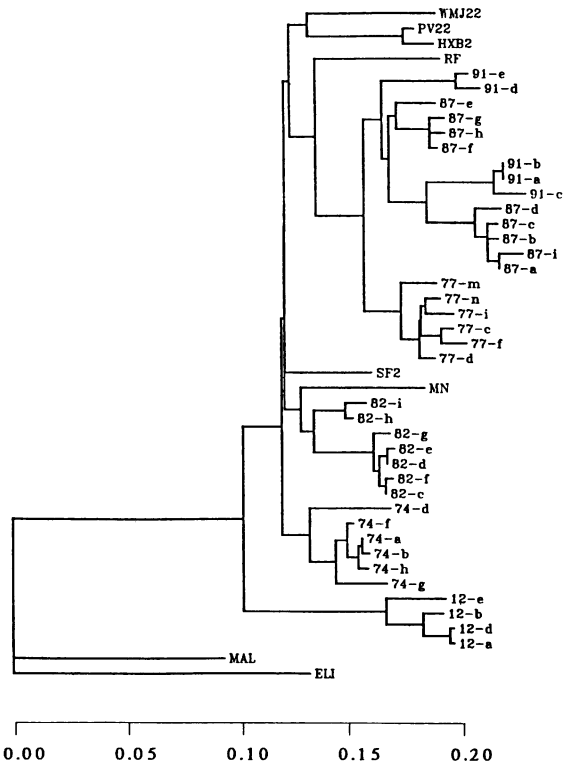

FIG. 4. Phylogenetic relationships of the V3 region of the HIV-1 *env* gene from Edinburgh hemophiliacs. The tree was obtained in the same way as that shown in Fig. 3, using the data summarized in Table 2.

previous work, the overall mean within-patient diversity is much lower, at only 1.65%. It is interesting to note that the patient showing the greatest diversity in this region also had relatively high diversity in *env*, but the figures for *gag* are too low to contain much information on phylogenetic relationships.

Divergence of HIV-1 sequences within the hemophiliac cohort. The phylogenetic analyses presented in Fig. 2 and 3 indicate that the sequences obtained from most of these patients are distinct from those of any other, despite the close relationships between those from several patients. Thus, the evolutionary divergence that is represented by the within-patient nucleotide distances has occurred during the period of infection, which is approximately 5 years, although we cannot know exactly when divergence occurred within that time. If we assume that divergence occurred immediately after infection, then we might estimate the rate of sequence change in these regions of gp120 by dividing the mean nucleotide distance between sequences from the same patient by the length of time those sequences might have been evolving independently (twice the time since infection). Using the data in Table 1, we would obtain a mean estimate of 0.55% per annum.

An estimate of sequence change in HIV-1 obtained in this way depends also on the assumptions (i) that the rate of change is constant and (ii) that it is the same in each patient. An estimate which did not depend on these assumptions would be preferable. To avoid the need to make these assumptions, we have adopted a model of HIV evolution in the cohort with which we can use information on the nucleotide distances between patients as well as within them (Fig. 5). As indicated above, six of the patients studied show interpatient mean distances which overlap the within-patient diversity. In addition, these six patients share a specific

TABLE 2. Nucleotide distances in the V3 region of HIV-1 among Edinburgh hemophiliacs[a]

Patient no.	Nucleotide distance						
	Patient no.						HIV isolates
	12	74	77	82	87	91	
12	*0.0378*						
74	0.1572	*0.0354*					
77	0.1858	0.1217	*0.0321*				
82	0.1618	0.0912	0.1228	*0.0328*			
87	0.2183	0.1375	0.0943	0.1371	*0.0498*		
91	0.2183	0.1335	0.1005	0.1480	0.0939	*0.0614*	
HIV isolates	0.2192	0.1608	0.1894	0.1525	0.1868	0.2125	*0.2000*

[a] Calculated from the data given in Fig. 2. See Materials and Methods and Table 1, footnotes *a* through *c*.

TABLE 1. Nucleotide distances in the V4-V5 region of HIV-1 among Edinburgh hemophiliacs[a]

Patient no.	Nucleotide distance										Provirus abundance[d]	Clinical status CDC stage[e]
	Patient no.									HIV isolates[c]		
	12	74	77	79	82	83	84	87	91			
12	*0.0529*										1/400	II
74	0.2429	*0.0920*									1/2,000	IVC2
77	0.2205	0.1115	*0.0379*								1/2,500	IVA
79	0.2665	0.1381	0.0820	*0.0728*							1/3,300	IVC2
82	0.3160	0.2254	0.2599	0.2852	*0.0729*						1/700	II
83	0.2581	0.1052	0.0669	0.0882	0.2454	*0.0380*					1/14,000	II
84	0.2643	0.1075	0.0720	0.0701	0.2691	0.0556	*0.0095*				1/10,000	IVC2[f]
87	0.2776	0.1278	0.0866	0.1098	0.2516	0.0779	0.0731	*0.0652*			1/600	IVC2
91	0.2454	0.1196	0.0787	0.0960	0.2376	0.0898	0.0808	0.1106	*0.0493*		1/2,700	IVC2
HIV isolates	0.2080	0.1766	0.1781	0.2030	0.2748	0.1687	0.1893	0.1950	0.1731	*0.1701*		

[a] Distances calculated from the data given in Fig. 1 as described in Materials and Methods. On diagonal, mean distances from all pairwise intrapatient comparisons. Below diagonal, mean distances from all pairwise interpatient comparisons.
[b] Patient infected in the United States after exposure to commercial factor VIII.
[c] Data from the group of eight independent HIV-1 isolates whose sequences are included in Fig. 1.
[d] HIV-1 provirus abundance in each PBMC DNA sample was estimated by limit dilution (27).
[e] Clinical status of each patient at the time of sampling.
[f] Given in error as CDC stage II previously (27).

TABLE 3. Nucleotide distances in *gag* p24 among Edinburgh hemophiliacs[a]

Patient no.	Nucleotide distance						
	Patient no.						HIV isolate
	74	77	79	82	83	84	
74	*0.0151*						
77	0.0261	*0.0101*					
79	0.0177	0.0178	*0.0150*				
82	0.0419	0.0376	0.0370	*0.0407*			
83	0.0359	0.0153	0.0271	0.0391	*0.0051*		
84	0.0278	0.0120	0.0145	0.0324	0.0214	*0.0133*	
HIV isolate	0.0320	0.0434	0.0361	0.0567	0.0489	0.0442	*0.0489*

[a] See Materials and Methods and Table 1, footnotes *a* through *c*, for details. Calculated from data submitted to the Los Alamos AIDS data base.

sequence motif in V4 which is readily recognizable in the amino acid sequence (26a). We interpret this result as indicating that they were infected with HIV sequences which originated from a single donor. On the other hand, patients 74 and 82 show much larger nucleotide distances from other Edinburgh patients and have done since seroconversion (L.-Q. Zhang and P. Simmonds, unpublished data). On this basis they are not strong candidates for infection from the same donor. We have therefore analyzed the data excluding those from patients 74 and 82.

Two components comprise the total sequence change in each patient (Fig. 5), one representing the within-patient divergence and a second comprising the sequence change that occurred prior to the time when the sequences present within the patient diverged from each other. In the case of patient A (Fig. 5), these are represented by *a* and *d*. Sets of simultaneous equations allow the unknown components *d*, *e*, and *f* to be estimated, using data for each of the inter- and intrapatient comparisons, represented by *a*, *b*, and *c* and AB, AC, and BC, respectively. The unknown branch length *d* can be estimated from any of 10 sets of three-way comparisons. We have taken the mean of all 10 and give these as our estimates for V4-V5 and V3 (Table 4). The branch components, e.g., *a* and *d* for patient A (Fig. 5), can then be summed for each patient to give the total HIV sequence evolution in these patients. We have also calculated the branch lengths for the *gag* sequences for five of the patients in the same way (Table 4).

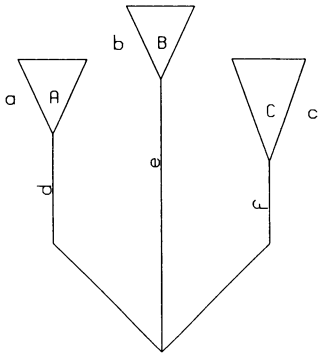

FIG. 5. Model of HIV sequence evolution in the hemophiliac cohort. In this model patients are indicated by A, B, and C, and the branch lengths of the individual components of the tree are indicated by *a* through *f*. The upper parts of the branches in each case represent sequence divergence among the samples obtained, and the branch length is taken as 1/2 the mean sequence distance within each sample. Then, if the mean interpatient nucleotide distances are indicated by AB, AC, BC, we obtain:

$$f = \frac{AC - AB + BC}{2} - c$$

and $e = BC - b - c - f$ and $d = AC - a - c - f$. The total sequence change that has occurred in patient A is then given by $a + d$ and likewise for the other patients. Estimates of these parameters for six hemophiliac patients are given in Table 4.

TABLE 4. Evolutionary divergence in the V4-V5 region, the V3 region, and *gag* among Edinburgh hemophiliacs[a]

Patient no.	Branch length (*d*)	Branch length (*a*)[b]	Total branch (*d* + *a*)
V4-V5 region			
77	0.0157	0.0190	0.0347
79	0.0132	0.0364	0.0492
83	0.0137	0.019	0.0327
84	0.0212	0.005	0.0262
87	0.0200	0.0326	0.0526
91	0.0274	0.025	0.0524
V3 region			
77	0.0343	0.0161	0.0506
87	0.0190	0.0250	0.044
91	0.0193	0.0308	0.0501
gag			
77	0.002	0.0051	0.0071
79	0.0008	0.0075	0.0083
83	0.012	0.0026	0.0146
84	0.0002	0.0067	0.0069

[a] Estimated from the model described in the legend to Fig. 5, using the data from Tables 1, 2, and 3.
[b] Calculated by dividing the mean intrapatient nucleotide distance by 2.

By this approach we estimate the mean branch length separating the sequences from these patients to be 4.12% (V4-V5) and 4.82% (V3). We have used both synonymous and nonsynonymous substitutions to obtain this estimate, because their ratio is close to unity in these regions. The range of branch lengths varies approximately twofold over patients in V4-V5 but less so in V3; however, fewer patients can be compared in that region. Comparable branch lengths in *gag* are much shorter because of the low frequency of amino-acid replacement substitutions (Table 4).

The extent of HIV sequence evolution in each patient is depicted in Fig. 6. Triangles represent the divergence observed between sequences in the same sample (Fig. 5, *a*). The vertical distance below the triangles represents prior divergence (Fig. 5, *d*) from the sequence in the infected donor (represented by the origin). Thus, patients 79 and 87, who show the greatest intrapatient nucleotide distance of this group, also contain sequences that appear to be most distant from the origin. In contrast, patient 84 shows the lowest intrapatient nucleotide distance but a greater change from the origin than does patient 79. All three patients were Centers for Disease Control (CDC) stage IVC2 when sampled, and all patients had been infected for approximately 5 years.

DISCUSSION

Low frequencies of inactivating mutations. It has recently been suggested that a high proportion, even a majority, of HIV proviral genomes must be defective. In a 216-nucleotide

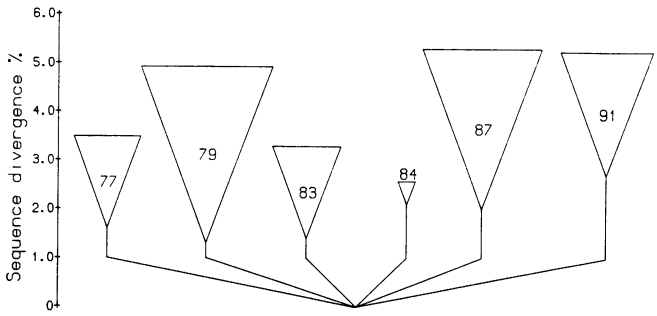

FIG. 6. HIV-1 sequence evolution among six hemophiliacs infected by exposure to a common batch of factor VIII. The estimates of mean branch length in Table 4 are plotted as a set of independently evolving virus populations which shared a common ancestor in the infected donor. Vertical distance is directly proportional to the estimated mean sequence change in each patient. The inverted triangles indicate the observed intrapatient sequence divergence for each patient (Fig. 5, *a*). The vertical distance from the root to the (inverted) apex of each triangle indicates the inferred evolutionary change (Fig. 5, *d*).

(72-amino-acid) segment of the first exon of the *tat* gene, Meyerhans et al. (18) observed four inactivating mutations in 35 sequences (total of 7,000 nucleotides, approximately). This gives a frequency of 5×10^{-4} per nucleotide, which is about 10-fold higher than the 5×10^{-5} we have observed in 42,000 nucleotides sequenced from *gag* and *env*. A number of defective *gag* and *env* sequences have been described by the same group, but these were obtained from isolates propagated in tissue culture (8). Meyerhans et al. (18) also observed a further eight variants which gave no Tat activity in vitro. It was suggested that up to a third of *tat* genes, and perhaps a majority of HIV-1 proviruses are inactive. However, the data we have presented indicate that the extension to the rest of the HIV genome is not justified. In the regions of *gag*- and *env*-encoding structural proteins that we have sequenced, the frequency of defective sequences is much lower than that found in *tat*.

Our earlier work (27) led us to conclude that the copy number of HIV-1 provirus in circulating PBMCs was close to one. We argued that functional complementation would not be possible within an infected cell and so we would not expect a high proportion of proviruses to be defective because they would continually be exposed to a strong selection for function. The results we have presented here confirm that even in the most variable regions of the HIV-1 genome there is only a very low frequency of inactivating mutations. We conclude that, in the peripheral circulation, proviruses defective in *env* or *gag* do not play an important role in the pathogenesis of HIV-1.

Source of infection. The particular value in studying the rate and pattern of sequence change in HIV-1 among these patients arises from two features. First, as for many hemophiliacs and other patients, the dates of seroconversion and likely dates of infection are known. Second, a single batch of factor VIII which was prepared by the SNBTS has been implicated in the infection of these patients (17), who had no other known risk factors. This batch was given to all but 1 of 16 patients who seroconverted between March and May 1984 and to 2 others who seroconverted later. In contrast, the next most likely batch was only given to nine of the patients who seroconverted at that time. None of the patients had received commercial factor VIII. There were, however, two other patients who seroconverted at that time who had not received either commercial factor VIII or the implicated batch.

A total of 32 patients were exposed to this batch, and the probability of seroconversion was significantly associated with the number of vials used (17). Whereas 8 of 9 patients who used more than 40 vials became infected, only 7 of 23 who used less did so. In addition, it seems likely that only a very few HIV-infected individuals contributed to that batch as only small numbers of HIV seropositives have been identified among sera obtained from the major risk groups in the collection area during 1983, when the plasma pool was being collected (24). On both counts, the implicated batch of SNBTS factor VIII does not appear to have been highly infectious.

As documented here and elsewhere (8, 9, 18, 26), HIV-infected individuals harbor many different viral sequences, but those present at any one time usually bear a close relationship to each other. In the case of the implicated factor VIII batch, the patients who seroconverted may have received a minimum infectious dose. Sampling small numbers of HIV sequence variants in this way from those present in the infected donor(s) could cause them to be unevenly distributed among the recipients. The hemophiliac patients studied might therefore have become infected with viruses whose genomes were already differentiated. However, if members of the cohort were infected with virus which originated from a single individual, they would possess HIV sequences that, although not identical, would be much more similar to each other than to those from unrelated individuals. If more than one infected donor contributed to the infected batch, then two or more groups of sequences might be distributed among the recipients.

Pattern of HIV sequence relationships within the Edinburgh hemophiliac cohort. We have analyzed a total of 74 sequences from the CD4-binding domain of *env* and flanking hypervariable regions V4 and V5 (19), which we obtained by limit dilution, amplification, and direct sequencing of the product (27). The phylogenies given in Fig. 3 and 4, which were obtained from nucleotide distance matrices, were similar to others produced by a maximum-likelihood procedure (see Materials and Methods). All the phylogenetic analyses showed, first, that sequences from the same patient are more similar to each other than to those from different patients. Second, the noncohort patient, who was infected in the United States from commercial factor VIII, has sequences which are very distinct from those of the cohort patients, as are all the published sequences from viral isolates. Third, in the V4-V5 region, six of the eight Edinburgh-infected hemophiliacs have sequences which are particularly closely related to each other, while two others (no. 82 and 74) have sequences which are less so.

Diversity within patients. The nucleotide diversity observed within individual patients varied over a wide range. Within the V4-V5 region the mean nucleotide distance between sequences obtained from patient 84 was 1%, while that from patient 74 was 9.2%. The overall mean sequence divergence within patients was 5%. We have examined the distribution of nucleotide distances among sequences from the same patients by plotting histograms of the results of all pairwise sequence comparisons (data not shown). We would in this way be able to detect whether there was evidence for any patient, and patient 74 in particular, having been infected with virus originating from more than one donor, which would be revealed as a bimodal distribution of nucleotide distances (see above). We did not detect such bimodality in any patient. There is therefore no evidence from the

distribution of nucleotide distances that any of these patients have received virus from more than one donor.

It has been suggested that the diversity of HIV sequences to be found within a sample might increase as the patient progresses into acquired immunodeficiency syndrome (M. Nowak, R. M. May, and R. M. Anderson, personal communication). For several of these patients, we earlier determined proviral abundance in terms of the number of PBMCs calculated to contain one HIV provirus, and this also might influence sequence diversity. In Table 1 the clinical status and proviral abundance of each patient is given. There is no apparent association with either variable. The samples at the extremes of the range of diversity are both CDC stage IVC2. The patient with the lowest mean diversity did have a low proviral abundance, but was not the lowest, and the two CDC stage II patients in the data set were as variable as most others. We conclude that no relationship between diversity and these variables can be established on the basis of these data.

Origin of HIV infection of the Edinburgh hemophiliac cohort. We have identified a group of six patients (no. 77, 79, 83, 84, 87, and 91) among whose HIV sequences there is a very close relationship (Tables 1 and 2). The mean of the interpatient nucleotide distances among these six is 8.25%. This is close to the intrapatient diversity observed for patients 79, 82, and 87, less than that found for patient 74, and about half that found in this region for independent HIV isolates (Table 1). All of the Edinburgh-infected patients are very divergent from the U.S.-infected control patient, although, as mentioned above, two of them grouped less closely to the other six. The clustering on the basis of nucleotide distance is supported by the absence in patients 74 and 82 of a specific oligopeptide motif in the V4 region which is common to the six most closely related patients (26a). We interpret the sequence data as indicating that these six patients were indeed infected with virus that originated from a single individual.

The two more-divergent Edinburgh-infected patients were also unusual in their seroconversion data. One (no. 82) seroconverted after exposure to a much lower dose than that received by most of those who became HIV seropositive after exposure to the implicated batch of factor VIII. The other (no. 74) seroconverted a very long time after exposure. In both cases, we have recently obtained evidence that their current distinctiveness is not due to very rapid sequence change; sequences from free virus in stored plasma from the time of seroconversion are similarly distinct (L.-Q. Zhang, unpublished data). We have recently found sequences similar to those of patient 82 in two more Edinburgh hemophiliacs (data not shown), one of whom also seroconverted in 1984 but did not receive any of the common batch. We therefore suggest that three patients may have been infected independently of the others at about the same time. The origin of the virus present in patient 74 is unclear and will require wider analysis of samples from patients who seroconverted in Scotland in 1983-1984.

Rate of HIV sequence evolution. The unusual circumstances surrounding the infection of these patients have permitted a direct interpretation of the divergence of HIV sequences from different patients in the three regions studied. Our model of sequence divergence among the six patients who contain the most closely related sequences (Fig. 5) has permitted us to use all the data to estimate directly the branch lengths in both the V4-V5 and V3 regions for each patient (Table 4; Fig. 6). These branch lengths are obtained from the sum of two components, the first of which is obtained from the diversity within the sample. The second component estimates the sequence change which had occurred since the original divergence of these viral genomes within the infected donor, up to the time the sequences within each of the current samples began to diverge from each other.

From this analysis we can show that the earlier component in fact varies relatively little in the V4-V5 region, by less than a factor of two (Table 4), and most of the difference between patients in terms of the overall lengths is contributed by the sevenfold range in nucleotide diversity among sequences within each sample. This could have come about through differences between patients in the rate of nucleotide substitution per virus replication cycle or the rate of virus replication, or both. In either case, in view of the genetic similarity of the virus which infected these patients, the similarity in initial dose, and the duration of the infection, the difference is likely to reflect differences in the response of the host to the virus.

The value of the approach we have taken is that, while it is impossible to determine exactly how long ago the sequences within each sample diverged from each other and it may be different for different patients, the overall divergence across patients infected by a single donor has accumulated over approximately 5 years. The HIV sequences received by each of the six patients shown in Fig. 6, although recently diverged, would not have been identical at the time of infection. The extent of this prior divergence, and consequently the time since divergence from the common ancestral sequence, is estimated by the difference between the mean inter- and intrapatient divergence for this group, 8.3 and 4.6%, respectively. The difference of 3.7% is lower than the mean intrapatient diversity, which implies that at least half (56%) of the divergence between sequences from different patients, indicated by the branch lengths in Fig. 6 and Table 4, has occurred in the 5 years since infection, the remainder having occurred in the infected donor. From this we estimate the mean rate of nucleotide sequence change in the V4-V5 and V3 regions to be 0.41 and 0.48% per annum, respectively. The total time since the cohort sequences diverged is estimated to be approximately 9 years.

Two earlier estimates of the rate of sequence change in the *env* gene of HIV-1, 0.92% (15) and 1.2% (29) on the basis of sequences from virus isolates, are higher than the estimate obtained here. Closer comparisons cannot be drawn however, as there are considerable differences between conserved and variable regions of the gene (15). Two aspects of our data suggest that caution should be exercised in extrapolating to obtain times of divergence between virus sequences from different sources. First, the twofold range over patients indicates that estimates based on individual patients or isolates (29) will be subject to large errors. Second, and perhaps more important, the extreme localization of much of the variability to very small regions implies that the nucleotide distance between highly diverged sequences will be consistently underestimated because of saturation effects which will not be adequately taken into account in whole-gene analyses.

Nevertheless, taking only the patients within our group, we can estimate three divergence times. First, the time since the divergence of the cohort sequence from the related Edinburgh sequence found in patient 74 would be approximately 14 years (against 9 years for the cohort). Second, the time since divergence of the cohort sequence from the U.S.-derived sequence found in patient 12 could be as much as 30 years, and third, the time since the divergence from the

sequence found in patient 82 (and two others, as indicated above) could be about 24 years. A reasonable interpretation would be that two HIV-1 lineages which were transmitted independently from North America are currently represented in Scotland and have been transmitted to these hemophiliacs.

ACKNOWLEDGMENTS

We thank Fiona McOmish and Alexander Cleland for technical assistance, the staff of the Haemophilia Centre, Royal Infirmary, for obtaining samples, and Paul Sharp for comments on the manuscript.

This work was supported by the Medical Research Council AIDS Directed Programme.

LITERATURE CITED

1. **Alizon, M., S. Wain-Hobson, L. Montagnier, and P. Sonigo.** 1986. Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients. Cell **46**:63–74.
2. **Carpenter, S., L. H. Evans, M. Sevoian, and B. Chesebro.** 1987. Role of the host immune response in selection of equine infectious anemia virus variants. J. Virol. **61**:3783–3789.
3. **Clements, J. E., F. S. Pedersen, O. Narayan, and W. A. Haseltine.** 1980. Genomic changes associated with antigenic variation of visna virus during persistent infection. Proc. Natl. Acad. Sci. USA **77**:4454–4458.
4. **Cordonnier, A., L. Montagnier, and M. Emmerman.** 1989. Single amino-acid changes in HIV envelope affect viral tropism and receptor binding. Nature (London) **340**:571–574.
5. **Devereux, J., P. Haeberli, and O. Smithies.** 1984. A comprehensive set of sequence analysis programs for the Vax. Nucleic Acids Res. **12**:387–395.
6. **Felsenstein, J.** 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22**:521–565.
7. **Fitch, W. M., and E. Margoliash.** 1967. Construction of phylogenetic trees. Science **155**:279–284.
8. **Goodenow, M., T. Huet, W. Saurin, S. Kwok, J. Sninsky, and S. Wain-Hobson.** 1989. HIV-1 isolates are rapidly evolving quasispecies: evidence for viral mixtures and preferred nucleotide substitutions. J. Acquired Immune Defic. Syndr. **2**:344–352.
9. **Hahn, B., G. M. Shaw, M. E. Taylor, R. R. Redfield, P. D. Markham, S. Z. Salahuddin, F. Wong-Staal, R. C. Gallo, E. S. Parks, and W. P. Parks.** 1986. Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. Science **232**:1548–1553.
10. **Kimura, M.** 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Molec. Evol. **16**:111–120.
11. **Kishino, H., and M. Hasegawa.** 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Molec. Evol. **29**:170–179.
12. **Lasky, L. A., G. Nakamura, D. H. Smith, C. Fennie, C. Shimasaki, E. Patzer, P. Berman, T. Gregory, and D. J. Capon.** 1987. Delineation of a region of the human immunodeficiency virus type 1 gp120 glycoprotein critical for interaction with the CD4 receptor. Cell **50**:975–985.
13. **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.
14. **Li, W.-H., C.-C. Luo, and C.-I. Wu.** 1985. Evolution of DNA sequences, p. 1–94. *In* R. J. MacIntyre (ed.), Molecular evolutionary genetics. Plenum Publishing Corp., New York.
15. **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.
16. **Li, W.-H., C.-I. Wu, and C.-C. Luo.** 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.
17. **Ludlam, C. A., J. Tucker, C. M. Steel, R. S. Tedder, R. Cheingsong-Popov, R. A. Weiss, D. B. McClelland, I. Philp, and R. J. Prescott.** 1985. Human T-lymphotropic virus type III (HTLV-III) infection in seronegative haemophiliacs after transfusion of factor VIII. Lancet **ii**:233–236.
18. **Meyerhans, A., R. Cheynier, J. Albert, M. Seth, S. Kwok, J. Sninsky, L. Morfeldt-Manson, B. Asjo, and S. Wain-Hobson.** 1989. Temporal fluctuations in HIV quasispecies in vivo are not reflected by sequential HIV isolations. Cell **58**:901–910.
19. **Modrow, S., B. H. Hahn, G. M. Shaw, R. C. Gallo, F. Wong-Staal, and H. Wolf.** 1987. Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates: prediction of antigenic epitopes in conserved and variable regions. J. Virol. **61**:570–578.
20. **Narayan, O., D. E. Griffin, and J. Chase.** 1977. Antigenic shift of visna virus in persistently infected sheep. Science **197**:376–378.
21. **Needleman, S. B., and C. D. Wunsch.** 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**:443–453.
22. **Nei, M.** 1987. Molecular evolutionary genetics. Columbia University Press, New York.
23. **Payne, S. L., O. Salinovich, S. M. Nauman, C. J. Issel, and R. C. Montelaro.** 1987. Course and extent of variation of equine infectious anemia virus during parallel persistent infections. J. Virol. **61**:1266–1270.
24. **Robertson, J. R., A. B. V. Bucknall, P. D. Welsby, J. J. K. Roberts, J. M. Inglis, J. F. Peutherer, and R. P. Brettle.** 1986. Epidemic of AIDS related virus (HTLV-III/LAV) infection among intravenous drug abusers. Brit. Med. J. **292**:527–529.
25. **Rusche, J. R., K. Jahaverian, C. McDanal, J. Petro, D. L. Lynn, R. Grimaila, A. Langlois, R. C. Gallo, L. O. Arthur, P. J. Fischinger, D. P. Bolognesi, S. D. Putney, and T. J. Matthews.** 1988. Antibodies that inhibit fusion of human immunodeficiency virus-infected cells bind a 24-amino acid sequence of the viral envelope, gp120. Proc. Natl. Acad. Sci. USA **85**:3198–3202.
26. **Saag, M. S., B. H. Hahn, J. Gibbons, Y. Li, E. S. Parks, W. P. Parks, and G. M. Shaw.** 1988. Extensive variation of human immunodeficiency virus type-1 *in vivo*. Nature (London) **334**: 440–444.

26a. **Simmonds, P., P. Balfe, C. A. Ludlam, J. O. Bishop, and A. J. Leigh Brown.** 1990. Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1. J. Virol. **64**:5840–5850.

27. **Simmonds, P., P. Balfe, J. F. Peutherer, C. A. Ludlam, J. O. Bishop, and A. J. Leigh Brown.** 1990. Human immunodeficiency virus-infected individuals contain provirus in small numbers of peripheral mononuclear cells and at low copy numbers. J. Virol. **64**:864–872.
28. **Simmonds, P., F. A. L. Lainson, R. Cuthbert, C. M. Steel, J. F. Peutherer, and C. A. Ludlam.** 1988. HIV antigen and antibody detection: variable responses to infection in the Edinburgh haemophiliac cohort. Br. Med. J. **296**:593–598.
29. **Smith, T. F., A. Srinavasan, G. Schochetman, M. Marcus, and G. Myers.** 1988. The phylogenetic history of immunodeficiency viruses. Nature (London) **333**:573–575.
30. **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.
31. **Tindall, K. R., and T. A. Kunkel.** 1988. Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase. Biochemistry **28**:6008–6013.
32. **Tsang, T. C., and D. R. Bentley.** 1989. An improved method using Sequenase™ that is independent of template concentration. Nucleic Acids Res. **16**:6238.
33. **Winship, P. R.** 1989. An improved method for directly sequencing PCR amplified material using dimethyl sulphoxide. Nucleic Acids Res. **17**:1266.
34. **Yokoyama, S., L. Chung, and T. Gojobori.** 1988. Molecular evolution of the human immunodeficiency viruses. Mol. Biol. Evol. **5**:237–251.