JOURNAL OF VIROLOGY, Dec. 1990, p. 5840–5850
0022-538X/90/125840-11$02.00/0

Vol. 64, No. 12

Analysis of Sequence Diversity in Hypervariable Regions of the External Glycoprotein of Human Immunodeficiency Virus Type 1

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

Department of Genetics, University of Edinburgh, Edinburgh EH9 3JN,[1] and Department of Haematology, Royal Infirmary, Edinburgh EH3 9YW,[2] Scotland, and Department of Biological Sciences, University of Maryland–Baltimore County, Baltimore, Maryland 21228[3]

Received 13 July 1990/Accepted 5 September 1990

Nucleotide sequences in three hypervariable regions of the human immunodeficiency virus type 1 (HIV-1) *env* gene were obtained by sequencing provirus present in peripheral blood mononuclear cells of HIV-infected individuals. Single molecules of target sequences were isolated by limiting dilution and amplified in two stages by the polymerase chain reaction, using nested primers. The product was directly sequenced to avoid errors introduced by *Taq* polymerase during the amplification process. There was extensive variation between sequences from the same individual as well as between sequences from different individuals. Interpatient variability was markedly less in individuals infected from a common source. A high proportion of amino acid substitutions in the hypervariable regions altered the number and positions of potential N-linked glycosylation sites. Sequences in two hypervariable regions frequently contained short (3- to 15-bp) duplications or deletions, and by amplifying peripheral blood mononuclear cell DNA containing 10^2 or 10^3 proviral molecules and analyzing the product by high-resolution electrophoresis, the total number and abundance of distinct length variants within an individual could be estimated, providing a more comprehensive analysis of the variants present than would be obtained by sequencing alone. Sequences from many individuals showed frequent amino acid substitutions at certain key positions for neutralizing-antibody and cytotoxic T-cell recognition in the immunodominant loop. The rates of synonymous and nonsynonymous nucleotide substitution in the region of this and flanking regions indicate that strong positive selection for amino acid change is operating in the generation of antigenic diversity.

Understanding the nature of sequence change in the human immunodeficiency virus type 1 (HIV-1) genome is central to current theories of viral pathogenesis and the immune response to infection. In common with other eucaryotic viruses with RNA genomes, HIV-1 shows considerable sequence diversity between different isolations, particularly those from geographically distinct regions, where divergence has taken place over a number of years. Sequence diversity is seen within isolations from the same individual as well as between HIV strains infecting different individuals (11, 29). The rate of HIV sequence variation is not uniform throughout the genome. Comparison of published sequences has shown that the *gag* and *pol* genes are more conserved than *env*. Furthermore, within *env* the pattern of variation is unusual (3, 34, 38). Five regions in gp120, V1 to V5, have been designated as hypervariable (23). These were defined as regions with 25% or less conservation of amino acids between a number of published sequences. In addition to high rates of amino acid substitution, the published sequences of some hypervariable regions have been previously reported to contain short deletions and insertions (3, 11, 23, 34). Several actual or potential gp120 epitopes are located in these regions (23, 34). Antigenic variation consequent to sequence differences in V3 (the immunodominant loop) has been reported (19, 36). This sequence corresponds to the predominant neutralization epitope of gp120 (28), and it has been argued that the observed variability represents an adaptive response by HIV to evade the immune system, as proposed for equine infectious anemia virus and visna virus (4, 5, 24). Whether other variable areas of gp120 are important in neutralizing-antibody or T-cell recognition is not known.

In this work, variation in the gp120 sequences of provirus present in circulating peripheral blood mononuclear cells (PBMCs) was studied. A number of individuals included in this study were infected from a common source (20), allowing variation between patients to be assessed. The hypervariable regions of *env* not only are polymorphic in sequence but in many cases also differ in length (3, 11, 23, 34). Thus, the spacing of the conserved regions of gp120 shows considerable variation. Using the polymerase chain reaction (PCR), we have investigated length variation of amplified DNA in different regions of gp120 and have visualized distinctive patterns of coexisting variants in each infected individual. The data on length variation together with extensive data obtained by sequencing hypervariable regions of single isolated molecules (30) indicate that the number of variants present in each patient is extremely large. Despite this heterogeneity, HIV sequences from individuals in the hemophiliac cohort, who were infected from a common source, were clearly more closely related to each other than to published sequences of HIV-1 and to those of a hemophiliac infected in the United States. In particular, the V4 sequence of the cohort members was distinctive and served to distinguish them from other hemophiliacs.

* Corresponding author.
† Present address: Department of Medical Microbiology, Medical School, University of Edinburgh, Teviot Place, Edinburgh EH3 9AG, Scotland.

MATERIALS AND METHODS

Clinical samples. Blood samples were obtained from 11 hemophiliacs infected with HIV-contaminated factor VIII

approximately 5 years ago (p12 to p95). For sequencing studies, samples were used from eight haemophiliacs who were exposed a batch of factor VIII that was implicated in infection by HIV-1 of 18 out of 32 recipients (20). All eight individuals seroconverted for antibody between 3 and 10 months after receiving the factor VIII (31). p82 is currently asymptomatic, and p83 has thrombocytopenia but is otherwise well; all others have been classified as IVc and suffer from a range of opportunistic infections and constitutional symptoms of HIV infection. Samples from individuals infected from other sources include p12, who was infected in the United States from commercial factor VIII, and 10 drug abuse-related and pediatric seropositive individuals (n1 to n10) from J. Mok (Infectious Diseases Unit, City Hospital, Edinburgh, Scotland).

HIV primers. Oligonucleotides were synthesized by the Oswel DNA Service, Department of Chemistry, University of Edinburgh, and were purified by high-performance liquid chromatography. The primers were based on the consensus of the following published HIV sequences: $HIV_{HTLVIIIB}$ (clones HXB2, BH102, and BH8), HIV_{BRU}, HIV_{CDC42}, HIV_{ELI}, HIV_{MAL}, HIV_{MN}, HIV_{PV22}, HIV_{RF}, HIV_{SC}, HIV_{SF2}, HIV_{WMJ22}, and HIV_{Z6}. The primer-binding sites in gp120 were chosen to be as highly conserved as possible between published sequences of geographical variants of HIV-1. No more than one mismatch with any of the North American, Haitian, or African sequences was present in any one of the primers. The sequences of the primers are given, with their positions in HIV clone HXB2 indicated in parentheses (+, sense; −, antisense): V3 (a) TACAATGTACACA TGGAATT (+, 6957), (b) TGGCAGTCTAGCAGAAGAAG (+, 7009), (c) CTGGGTCCCCTCCTGAGG (−, 7331), and (d) ATTACAGTAGAAAAATTCCCC (−, 7381); V4-V5 (e) TCAGGAGGGGACCCAGAAATT (+, 7316), (f) GGGGAA TTTTTCTACTGTAAT (+, 7360), (g) CTTCTCCAATTGT CCCTCATA (−, 7665), and (h) CCATAGTGCTTCCTGCT GCT (−, 7814).

Double-PCR method. DNA from at least 5×10^6 PBMCs was prepared as described by Simmonds et al. (30). Amplification of DNA by double PCR and quantification by limiting dilution were carried out as described by Simmonds et al. (30). All DNA extractions and amplification reactions carried appropriate parallel negative controls (blood from seronegative, low-risk group blood donors) to detect contamination at any stage in the procedure. In principle, at least five outer primer pairs may be used simultaneously in the first amplification reaction. However, DNA sequences amplified by the outer V3 and V4-V5 primers overlap. To permit the same sample to be amplified in the two regions, the following combinations of primers were used. In the first reaction, positive-sense, outer V3 primer (primer a; see above) was used with the antisense, outer V4-V5 primer (primer h), to amplify a 858-bp fragment; in the second reaction, a sample of the product was amplified by using the inner V3 (primer b) and inner V4-V5 (primer g) primers (657 bp); in the third reaction, the final amplification used the inner V3 (b plus c) and V4-V5 (f and g) primers in separate reactions. This modified procedure is equivalent in sensitivity and yield to separate amplification in two stages by double PCR (data not shown).

Length analysis of PCR products. To investigate length variation in product DNA amplified in the three *env* regions, the final PCR reaction was carried out in the presence of 0.05 to 0.1 μCi of [α-^{35}S]thio-dATP (1,000 Ci/mmol; Amersham) and 8.3 mM each of unlabeled deoxynucleoside triphosphates (dNTPs). A 1-μl sample of the PCR product was heated to 95°C in 50% formamide for 3 min and electrophoresed on a denaturing polyacrylamide gel (6% acrylamide, 0.3% *N',N* bisacrylamide, 8 M urea, 0.089 M Tris, 0.089 M boric acid, 0.002 M EDTA, pH 8.3). Standard DNA size markers (1,635, 1,018, 516, 506, 394, 344, 298, 220, 200, 154, and 142 bp; BCL) were end labeled with [γ-^{32}P]dATP. Gels were dried and exposed overnight on X-ray film. Analysis of migration distances and intensity of bands was assisted by the use of a densitometer.

DNA sequencing of PCR products. The double-PCR procedure produces sufficient DNA from single isolated molecules of target sequence to permit direct sequencing of the PCR product (30). Unreacted dNTPs and primers were removed from product DNA by using Gene-Clean (Bio 101, Inc.). Direct sequencing was then carried out by using one of the primers used for the final amplification by the Sequenase protocol (U.S. Biochemical Corp.), with the following modifications: the standard annealing mix was adjusted to 10% dimethyl sulfoxide (40), and the sample was denatured by boiling for 5 min and immediately chilled on ice. The labeling reaction was performed at room temperature for 5 min with 0.25 μM two unlabeled dNTPs and 5 μCi of α-^{35}S-labeled dATP or dCTP (1,000 Ci/mmol; 37). The choice of unlabeled and labeled dNTPs was based on the nucleotides present 3′ to the priming site used. Termination reaction mixtures contained a final concentration of 10% dimethyl sulfoxide. Sequences were analyzed on a 6% polyacrylamide wedge sequencing gel.

Nucleotide sequence accession number. All sequences used in this study have been submitted to GenBank (accession number M36997).

RESULTS

Amplification of HIV provirus by PCR. Two different approaches to PCR amplification were used. In the first, a sample of DNA large enough to contain a representative sample of provirus was amplified. In this case, the product should be representative of the provirus population in the sample (see below). In the second method (30), single molecules of provirus were isolated by dilution before being amplified. In this case, the product of each amplification derives from a single provirus. In many cases, a double-PCR reaction using nested primers was used. In this method, a small sample of the first PCR reaction is used to prime a second (30). Consequently, the products of a first amplification can be used to prime more than one amplification, for example, one with labeled precursor for visualization by autoradiography and a second without labeled precursor for direct sequencing (see below).

In published HIV-1 sequences, the length of the region amplified by the V4-V5 primers (spanning V4, C3, and V5; 23) varies by as much as 30 of around 300 nucleotides. Substantial size variation in this region is also found within the provirus population present in a single infected individual. For example, a sample of DNA from hemophiliac p79, containing about 60 molecules of provirus (30), was amplified with the V4-V5 primers and labeled with [α-^{35}S]thio-dATP during the second amplification step. When analyzed by gel electrophoresis, the product showed a complex set of size variants (Fig. 1, lane U). To confirm that these were natural variants present in the DNA sample rather than artifacts of the PCR reaction, single provirus molecules were isolated by dilution before amplification (30). A total of 31 replicate tubes, each containing 625 cell-equivalents of DNA, were amplified in the double PCR, with [α-^{35}S]thio-

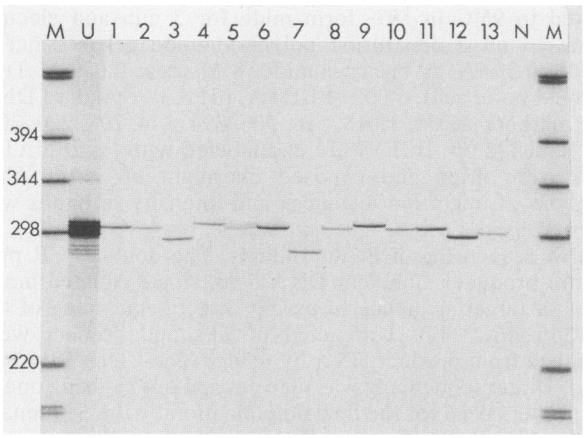

FIG. 1. Size analysis of amplified proviral DNA in the V4 and V5 regions from PBMCs of an HIV-infected individual (p79). Positive reactions at limiting dilution of PBMC DNA from an HIV-infected hemophiliac (p79) were amplified with V4-V5 primers and electrophoresed on a polyacrylamide gel (lanes 1 to 13). Lane U, 1 μg of PBMC DNA from p79; lane M, product size estimated with DNA markers of sizes (in nucleotides) indicated (see Materials and Methods); lane N, 1 μg of PBMC DNA from an uninfected individual.

dATP added to the second reaction. Thirteen of the reactions were positive. From the 18 negative reactions, the Poisson formula predicts that approximately 75% of the positive reactions are due to the amplification of a single provirus molecule. Electrophoretic analysis of the positive reactions showed that the length variants found by amplifying 60 provirus molecules together were also found when the molecules were isolated before amplification (Fig. 1, lanes 1 to 13). A single lane (lane 5) clearly contained two length variants, consistent with the prediction of the Poisson formula.

To determine the basis of the length variation, the products of 12 of the 13 first PCR reactions (omitting reaction 5) were again amplified in the second PCR but without labeled substrate and directly sequenced (30). The 12 nucleotide sequences are shown in Fig. 2. The lengths of the amplified sequences calculated from electrophoretic analysis agree well with the lengths determined by sequencing (Fig. 2, last two columns). Thus, the differences in electrophoretic mobility reliably reflect differences in length.

The isolated molecules appeared to be a random sample of the sequences visualized in the bulk amplification of 60 provirus molecules (Fig. 1, lane U). Thus, the four more intense bands in lane U with lengths of 296, 299, 302 and 305 nucleotides were represented by one, one, six, and two isolated sequences, respectively. Faint bands with lengths of 284 and 293 nucleotides were represented by one isolated sequence, and others with lengths of 287 and 281 nucleotides were not represented. Thus, the analysis of bulk samples provides a convenient and rapid assessment of the spectrum of size variation within the provirus population present in a DNA sample.

Amino acid sequence variation in HIV-infected individuals. PBMC DNA from eight other infected individuals was diluted and sequenced in the V4 and V5 regions as described for p79, in all cases using low frequencies of positive reactions at limiting dilution to avoid multiple positives. To simplify the presentation, the nucleotide sequences were translated. Figure 3 shows alignments of such amino acid sequences obtained from each individual. Considerable diversity in the V4 and V5 regions was seen within and between individuals. In many individuals, it was necessary to insert notional gaps to preserve alignment of the sequences. These gaps were concentrated exclusively in the V4 and V5 regions, where there was also a high rate of amino acid substitution. Each gap was a multiple of 3 nucleotides, maintaining the reading frame of the sequence. The degree of sequence variation within an individual varied considerably, the sequences from p77 and p79 showing the most heterogeneity in this group and those from p84 showing the least.

A number of sequences showed short repeats of 3 to 6 amino acids in V4 and V5 (Fig. 3). The same repeated sequences were found in different individuals. In V4, sequences containing (N)STW were repeated in p74, p77, and p82, while p77, p79, p83, p84, p87, and p91 showed repeats of sequences TTGSN and TTESN. In V5, repeats of NET were found in two individuals, p77 and p12. The number of repeated sequences varied both between and within samples. This duplication of sequence motifs accounts for some of the observed length variation. Sites for potential N-linked glycosylation sites were concentrated in the hypervariable regions. In each individual, there was a concentration of potential N-linked glycosylation sites in the V4 and V5 regions. The number and positions of such sites were variable between individuals and within each patient sample (Fig. 3). Variation in the number of repeats changed the overall potential for carbohydrate addition in this region of gp120 (Fig. 3).

Length polymorphism of DNA amplified in the V4 and V5 regions. With the exception of p82, few identical sequences

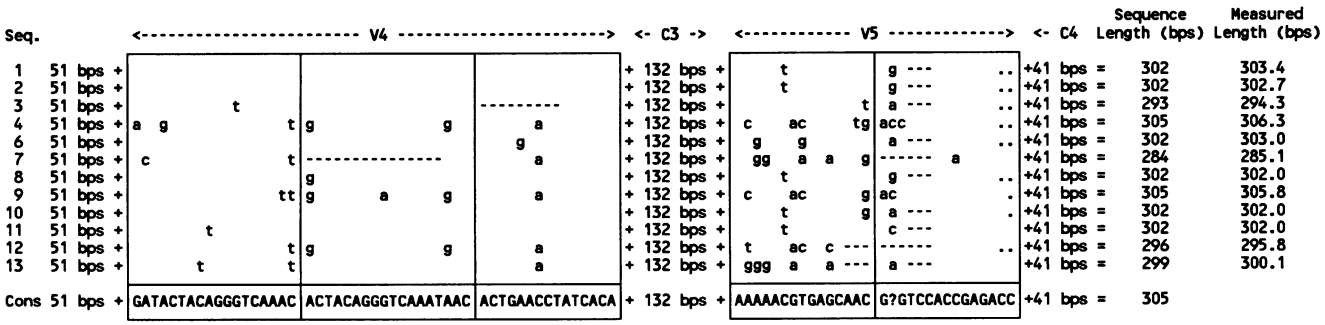

FIG. 2. Alignment of sequences obtained by limiting dilution of PBMC DNA from p79. Sequence labels V4, C3, V5, and C4 follow previous usage (23). Symbols: ., unsequenced; −, gap introduced to preserve alignment; ?, no majority consensus. Sequences 1 to 13 correspond to lanes 1 to 13 in Fig. 1. Differences from the consensus are indicated in lowercase.

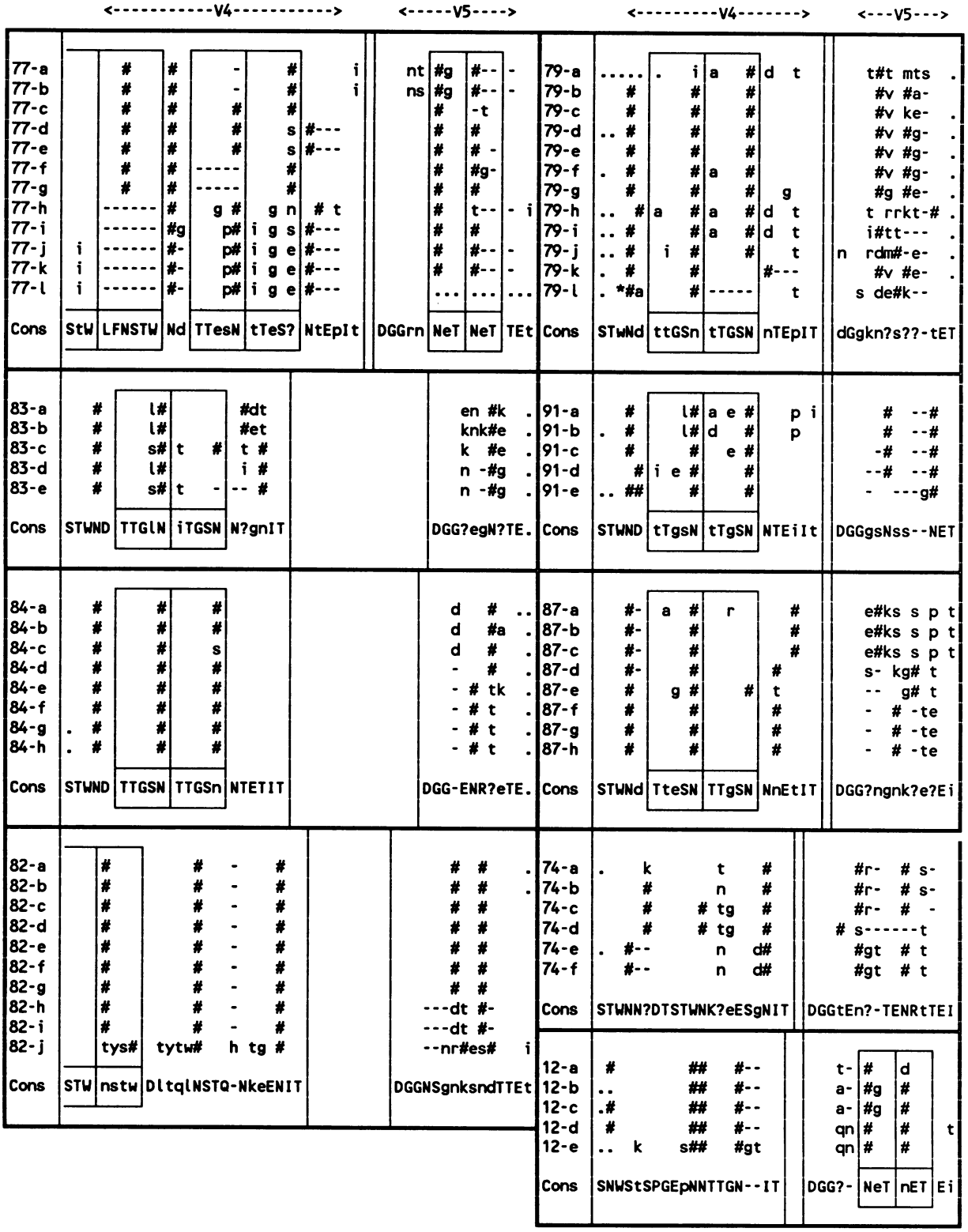

FIG. 3. Alignment of sequences from nine hemophiliacs in the V4 and V5 regions. Individual sequences from each individual are indicated as a to z. Symbols: ., not sequenced; −, gap introduced to preserve alignment; #, asparagine residue site of potential N-linked glycosylation; ?, no majority consensus at this position; *, stop codon. Conserved amino acids are shown in uppercase; nonconserved amino acids are shown in lowercase. Tandem repeated sequences are indicated by boxes. Potential glycosylation sites shown whether conserved or not.

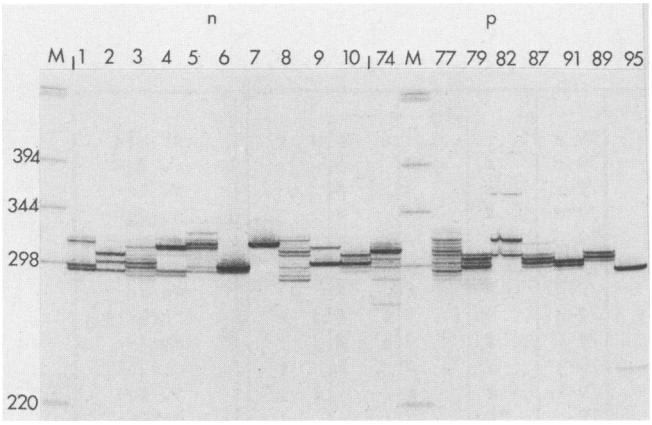

FIG. 4. Size analysis of amplified proviral DNA in the V4 and V5 regions from PBMCs of HIV-infected individuals. Samples (1 μg) of PBMC DNA from HIV-infected nonhemophiliacs (n1 to n10) and hemophiliacs (p74 to p95) were amplified with V4-V5 primers and electrophoresed on a polyacrylamide gel. Lane M, Product size estimated with DNA markers of sizes (in nucleotides) indicated (see Materials and Methods).

were isolated from the same individual (Fig. 3). No two identical sequences were isolated from p12, p79, p83, p87, or p91; p74 and p77 each had two identical sequences, p84 had three, and only p82 and p87 had multiple incidences of two sequences. Clearly, the sequences determined do not represent the full range of variation present in most of the individuals studied. It would be possible in principle to sequence each DNA sample exhaustively. However, a simpler approach to the question of variation within and between individuals is to use the PCR technique described above to assess the profile of length variation in each sample. The length analysis can be carried out with a standard amount of PBMC DNA, so that many samples can be dealt with rapidly.

To demonstrate the range of length variation within PBMC samples, 1-μg samples of DNA from 18 HIV-infected individuals were amplified by using the V4-V5 primers as described above (Fig. 4). The samples compared were from p74, p77, p79, and p82, and the number of provirus molecules present in the 1-μg samples ranged from 60 (p79) to 200 (p82; 30). A close concordance was observed between the sizes of the bands measured by gel electrophoresis and sizes determined by sequencing isolated molecules from the same sample of DNA (Table 1). For example, the prominent bands in the p74 and p82 samples with estimated sizes of 308.0 and 316.5 bp were paralleled by a predominance of sequences with overall lengths of 308 and 317 bp. Similarly, the wide range of band lengths observed upon amplication of the p77 sample was paralleled by a wide range of lengths among the individually determined sequences. These data show that an analysis of length variation serves as an approximate measure of the overall sequence variation in a provirus population. The sequence data (Fig. 3) show that a length variant may comprise several different sequence variants. For example, sequences 74-a, -b, and -c had the same overall length as 74-e and -f, although the two groups differed in the positions of the gaps and a number of amino acid substitutions. An impression of the large range of variation which can occur within a sample may be gained by considering the range of size variation and the number of sequences of a given length together. In several cases, this must represent a tremendous amount of variation in the V4-V5 region alone.

Band lengths overved in the V4 and V5 regions with samples from eight hemophiliac patients are summarized in Table 2. Data from 10 drug abuse and pediatric patients are also listed for comparison. The overall range of lengths and patterns were similar in the two groups. The majority of samples contained bands in the range of 293 to 317 bp, but there was considerable variation in the diversity of bands present in a sample. At one extreme, n3 and p77 had 10 and 13 bands, respectively, with ranges of 287 to 317 and 287 to 326 bp. At the other extreme, n6, n7, and p95 had single bands of 296, 314, and 293 bp. The range of lengths of the V4 and V5 regions of geographically diverse isolated HIV-1 sequence variants, from 293 (HIV_{SF2}) to 311 (HIV_{CDC42}), lies within the range observed in this study and suggests possible selective constraints on the size of these regions.

Sequence variation in the V3 hypervariable region. To investigate sequence variability of the immunodominant loop (28), sequences of individual provirus molecules in the PBMC DNA samples were determined with the V3 primers

TABLE 1. Comparison of size of amplified DNA with overall lengths of sequences obtained by limiting dilution[a]

p74		p77		p79		p82	
Amplified DNA	No. of sequences	Amplified DNA	No. of sequences	Amplified DNA	No. of sequences	Amplified DNA	No. of sequences
		325.1					
		322.5	1 at 323				
		316.6	1 at 317			**316.5**	8 at 317
		313.7	2 at 314			313.7	1 at 314
311.2		311.7	1 at 311				
310.0	5 at 308	**308.3**	2 at 308			306.9	
		305.1		**305.1**	2 at 305	**304.0**	2 at 305
301.2		302.0		**301.65**	6 at 302	301.3	
		299.2	1 at 299	**299.9**	1 at 299		
295.5	1 at 296	296.5		**298.1**	1 at 296		
		293.8			1 at 293		
		290.7					
288.0		288.0	2 at 287				
					1 at 284		

[a] Prominent bands are in bold type; all sizes are given in base pairs.

TABLE 2. Distribution of length variants in the V4 and V5 regions

Sample[a]	Intensity of band[b] at observed size (bp) of:													
	287	290	293	296	299	302	305	308	311	314	317	320	323	326
n1			3	3					1	2	1			
n2			3		3		3							
n3	1	2	1	3	3	1	1		2	1	1			
n4		2	2				1		5		1			
n5			1	1					3	3	1		2	
n6				5										
n7										5				
n8	2	2	2	1			2	2	1	1	2	1		
n9					5				2					
n10			1	1	5	1	3	1						
p74	1			2	1	2		4	1					
p77	1	1	2	1	2	1	2	2	1	1	2		1	1
p79[c]	1		1	3	2	3	2							
p82			1			1	2	1		1	4			
p87				2	3	3	2			1				
p91			1	3	3									
p89				1	2	3	3							
p95			5											

[a] n1 to n10, HIV-infected nonhemophiliac individuals; p74 to p95, hemophiliacs.
[b] Scored on a scale of 1 (weak) to 5 (strong).
[c] Bands (intensity = 1) also at 284 and 281 bp.

(Fig. 5). Sequences bearing the same identification letter in Fig. 3 (V4 and V5 sequences) and Fig. 5 (V3 sequence) were derived from the same provirus molecule (see Materials and Methods). Sequence variation accompanied by little length variation was the dominant feature of the sequences of V3 and flanking regions (Fig. 5). The sequence diversity within a sample was no less than in the V4 and V5 regions. However, in the V3 region, each sample showed particular sequence features, often common to all sequences within a sample, which distinguished it from other samples. Over the V3 region as a whole, only 23 of 83 amino acids were conserved in all of the sequences listed in Fig. 5. The immunodominant loop sequences all differed considerably from the prototype HIV_{MN} and $HIV_{HTLVIIIB}$ (clone HXB2) sequences and from those of other geographical HIV-1 variants (Fig. 5). The core sequence GPGR was well conserved between individuals, although a number of variants existed. However, in one individual (p12), the majority of sequences were GSGR.

The flanking regions of V3 contain a large number of potential N-linked glycosylation sites. There are six of these in the prototype $HIV_{HTLVIIIB}$ sequence in Fig. 5, and there are six potential sites for N-linked addition lying to either side of the V3 loop structure. No glycosylation sites are common to all of the sequences. By contrast, the two cysteine residues spanning the immunodominant loop are absolutely conserved (Fig. 5).

DISCUSSION

Diversity of *env* sequences within HIV-infected individuals. A striking feature of the proviral sequences reported here is the diversity of *env* sequences within a number of the individuals studied. In many cases, there is a complete lack of homology between variants over a number of amino acid residues (for example, the V5 region of p79). It is clear that a very considerable amount of sequencing would be necessary to describe fully the range of variation in proviral sequences within PBMCs. The visualization of length variants in the V4 and V5 regions provides a rapid method for the partial characterization of sequence variants within a sample. Analysis of the amplified DNA from samples containing large numbers of proviral sequences shows the range and relative abundance of sequences that differ in length. The method provides evidence for the existence of relatively scarce sequences that were not detected by conventional sequence analysis of a relatively small number of variants. The method does not provide comparative sequence information on the different length variants, and each length variant comprises an unknown number of distinct sequences. However, it provides a simple description of the provirus population which allows large numbers of individual samples to be compared. Applications include the analysis of sequence change over time in an individual (P. Simmonds, unpublished data) and comparisons between proviral variants in PBMCs and viral RNA in plasma (L.-Q. Zhang, personal communication). It also allows genuine positive PCR results from patient screening to be distinguished from those due to contamination by cloned sequences (39).

Phylogenetic significance of amino acid changes in *env*. Many hemophiliacs treated with commercial factor VIII in the early 1980s became infected with HIV-1. The rate of infection was considerably lower in those who were treated with factor VIII prepared from plasma of volunteer blood donors in a low-prevalence area for HIV-1 infection, such as Scotland (21). However, some hemophiliacs treated in Edinburgh solely with locally produced factor VIII were infected with HIV-1 in 1984 (20). A single batch of factor VIII has been implicated in the infection of 18 hemophiliacs, including hemophiliacs p74, p77, p79, p82, p83, p84, p87, and p91 studied here. p12 was infected in the United States from commercial blood products.

The V4 regions of six of the eight designated cohort members (p77, p79, p83, p84, p87, and p91) are very similar to each other, and all contain two repeats of a relatively well conserved 5-amino-acid sequence, TTGSN. p77 and p79 contain variant proviral sequences that lack one of the two copies of this sequence, while some variants from p77 have a second copy of the preceding LFNSTW sequence. However, p82 and p74, who were also considered to be infected

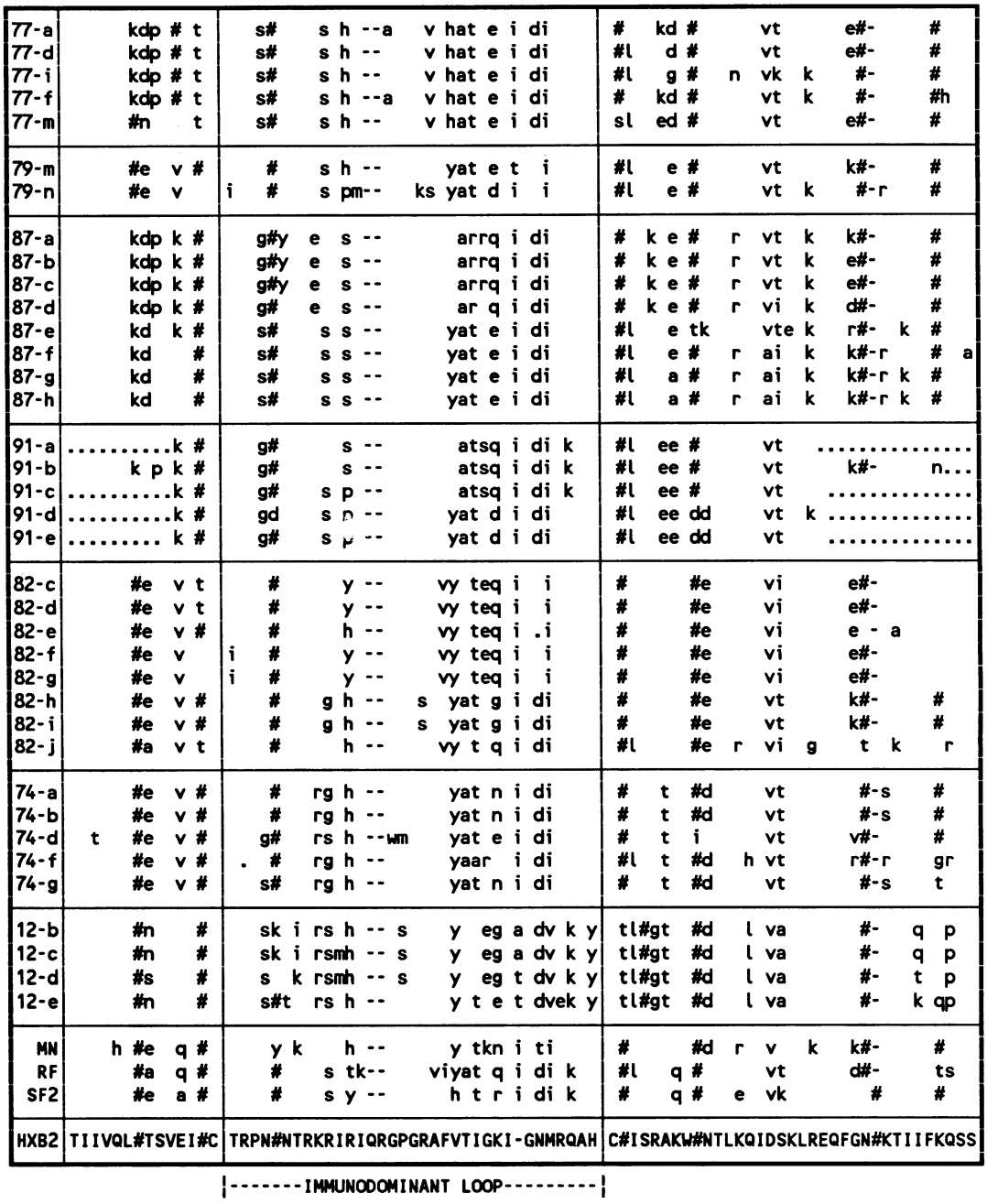

FIG. 5. Alignment of sequences obtained from seven hemophiliacs in the V3 and flanking regions. Individual sequences are indicated as a to z. Symbols are as for Fig. 3. Differences from the pHXB2 sequence are shown in lowercase. All potential N-linked glycosylation sites are shown.

from the same source, contained proviral sequences in this area that were distinct from each other, from those of the other cohort members and p12, and from published sequences (Fig. 6A). However, at the time of seroconversion, the V4 sequences of both individuals were similar to those reported here for the two individuals (Zhang and Simmonds, unpublished observations). This finding rules out the possibility that a higher rate of sequence change of HIV in these two individuals was responsible for the current dissimilarities in V4 sequences from the rest of the cohort. At present, it is not clear whether several strains of HIV-1 were present in the same contaminated batch of factor VIII or whether p82 and p74 were infected from a different source altogether. There is no evidence for other risk factors for HIV infection in either hemophiliac. The most likely sources are other infectious batches of locally prepared factor VIII concentrate. V4 sequences similar to those of p82 have been found in another hemophiliac (data not shown) who was not treated with the implicated batch of factor VIII but who did receive several other locally produced batches that were also given to p82.

In contrast to the similarities observed in the V4 region, no

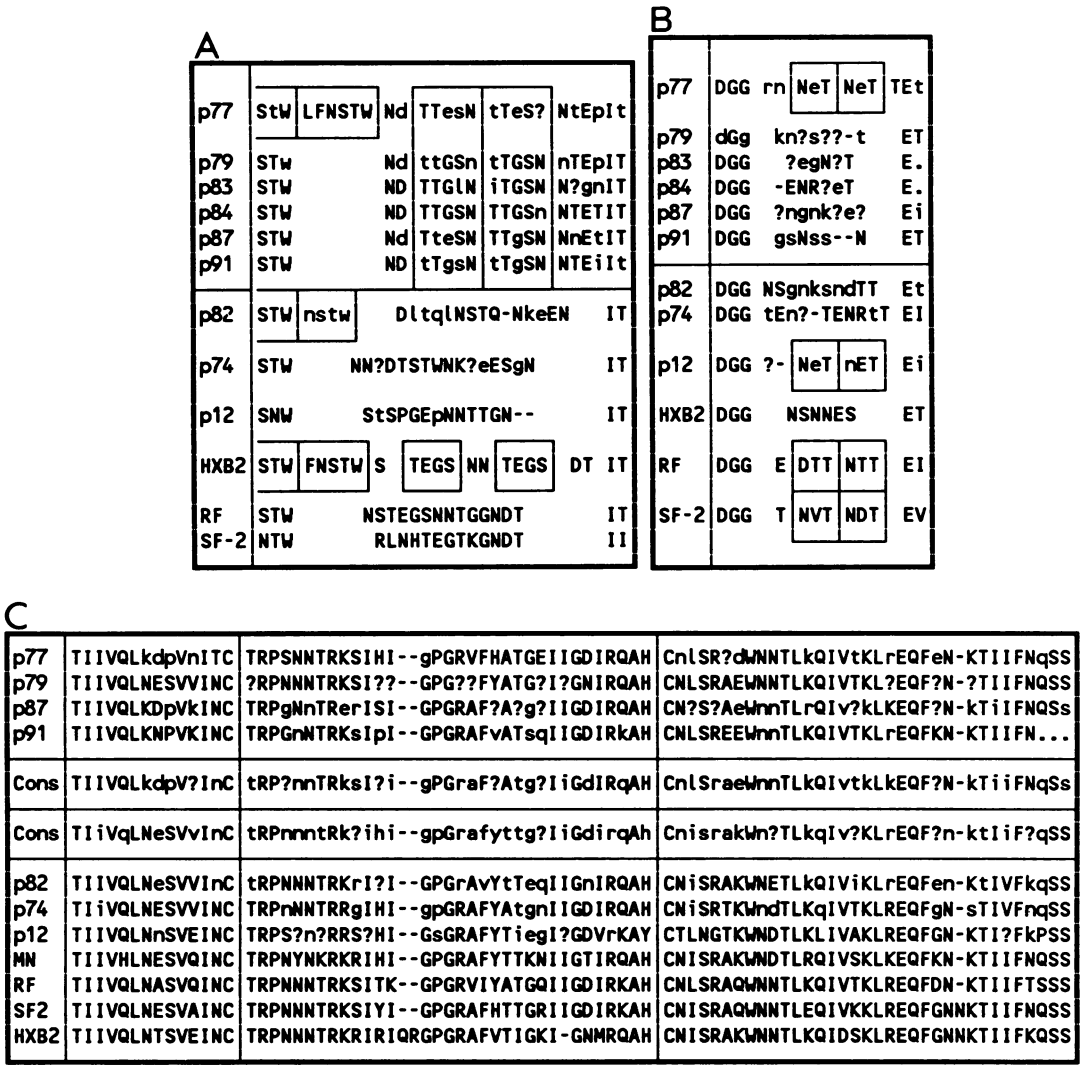

FIG. 6. Comparison of hemophiliac and published sequences in the *env* hypervariable regions V4 (A), V5 (B), and V3 (C). Consensus sequences from individuals believed to have been infected from a common source (p77, p79, p83, p83, p87, and p91) are compared with those of apparently unrelated sequences (p82 and p74), those of a hemophiliac infected from commercial factor VIII, and published sequences. Differences from the consensus within infected individual are indicated in lowercase. ?, No majority consensus at this position in an individual or group consensus sequence.

relationship between the cohort members is apparent in V5 (Fig. 6B), where sequence variation is much greater. The rate of sequence change in the short V5 region appears more rapid and does not indicate any relatedness between the six individuals who have similar V4 sequences. The V3 regions of the members of the hemophiliac cohort differ from the reference HXB2 sequence by a number of amino acids. However, no clear relationship distinguishes the cohort sequences. This is shown by the similarity in the consensus of the four confirmed cohort members and the combined consensus of noncohort members p12, p74, and p82 and published sequences (Fig. 6C). There are only five differences between the consensus sequences of the two groups, and at all but one of these sites, variants within the confirmed cohort members exist that match the noncohort sequences (Fig. 6C).

Nature of variation in the hypervariable regions of *env*. Both amino acid substitutions and gaps contribute to variation in the V4 and V5 regions. Many of the gaps in V4 involve repeated sequences such as TTGSN (p77 and p79). In such cases, some variants have one copy while others have two. Similar variation in the numbers of copies of the sequence (F)NSTW may also be found in p77. Indeed, the existence of repeats of these two sequences in most individuals indicates that some sort of duplication event has taken place. In many sequences, there are minor differences between the two copies, suggesting that some sequence change has taken place after duplication. The exact sequence involved in a duplication event may differ. For example, the block NSTW is repeated in p82, while in p77 there is often a repeat of the longer sequence LFNSTW. In published sequences of the viral isolated $HIV_{HTLVIIIB}$, the BH8 clone has one copy of the sequence FNSTW, while others (for example BH10) have two. The widespread occurrence of these repeated sequences (3, 11, 23, 34) and the likelihood that they occur independently in different HIV-infected

individuals suggest that this sequence is predisposed to duplication during either reverse transcription or RNA synthesis. Furthermore, if duplication is occurring repeatedly in the samples that we have examined, it is also likely that once formed, the duplications are predisposed to deletion.

All of the insertions or deletions in the V4 and V5 regions are multiples of 3 nucleotides, thus maintaining the reading frame downstream. Similarly, only one chain termination mutation (p79, sequence 1; Fig. 3) was found in 37 V3 sequences and 71 V4 and V5 sequences. The low rate of inactivating mutations is consistent an absence of phenotypic mixing. This finding may reflect the low copy number (close to one) of provirus in infected PBMCs (30).

Positive selection for sequence change in hypervariable regions of *env*. There has been no satisfactory explanation of the high rates of mutation in localized regions of the *env* gene. It could be argued that the cause is simply a lack of functional constraints which might limit the amount of variation in regions such as the CD4 binding site (6, 16). However, this view is not supported by a comparison of the rates of synonymous and nonsynonymous substitution rates in the different regions of *env*. Published data and data collected in this laboratory (P. Balfe et al., unpublished data) give a ratio of synonymous to replacement substitutions (K_s/K_a ratio; 18) for the CD4 binding site of 1.24. For comparison, the K_s/K_a ratio for *gag* sequences (17) and those of samples studied here is about 6.7, and the ratio for 42 eucaryotic genes is 5.28 (18). Thus, the CD4 binding site does not appear to be under stringent constraint to maintain its amino acid sequence, and the much higher substitution frequency of the hypervariable regions cannot simply be due to lack of constraint. The K_s/K_a ratio for the sequences reported here of the V3 loop and flanking regions is 0.67, lower than any previously reported. On average, the survival of a replacement mutation is almost twice as probable as the survival of a synonymous mutation. Overall, this signifies that selection favors change in this region, although the absence of stop codons and the conservation of a number of amino acids, including the cysteine residues spanning the V3 loop, indicate that the extent of change possible is limited. Although positive selection for change is unusual, it has been observed in major histocompatibility complex proteins (13) and in mammalian and avian serine protease inhibitors (12, 15). For the specific requirements of increased diversity of antigenic recognition by the major histocompatibility complex molecule, and for defense against a range of bacterial proteases in the latter example, positive selection appears to confer a selective advantage for the mutated sequences. Given the known involvement of V3 in virus neutralization (28), selection for change in V3 suggests that the selective force is the immune defense system.

Consequences of sequence change in the hypervariable regions of *env*. Several areas of the *env* gene product have been shown to be antigenic upon natural infection and upon vaccination (8, 9, 19, 26–28, 36). The V3 loop and regions in gp41 have been positively identified as targets of antibody-mediated neutralization or cytotoxic T-lymphocyte (CTL) killing. Both immune effector functions are sensitive to sequence variation around the crown of the V3 loop, i.e., either side of the relatively well conserved GPGR central sequence. The specificity of the CTL response can be determined in large part by the amino acids immediately downstream of GPGR (36), while residues immediately upstream of GPGR have been shown to be important in antibody-mediated neutralization (10, 19, 22). Considerable sequence variation is found in both of these sites in the sequences obtained in this study. Both tyrosine and valine are found at a position critical for CTL recognition (36), not only in different samples but also within the same sample (e.g., p87 and p91; Fig. 6C). A histidine substitution at this site (p77) would also be expected to modify the epitope. Other amino acid changes are concentrated at the site of B-cell recognition, with most samples containing at least two different sequences. There are also mutations in the central GPGR motif (p12, p74, and p77) that could disrupt the β turn at this site (22) and probably abolish immune recognition. This high level of variation in V3 predicts the existence of a wide range of neutralization serotypes in many of the patients. Sera and CTLs from such individuals may be expected to be reactive against a range of standard viral isolates. The broadening specificity of neutralizing antibodies with time seen in HIV-seropositve individuals (1) may be due to the de novo appearance of V3 serotypes upon long-term infection.

Neutralizing antibodies that bind to V4 and V5 regions have not been described, and the contribution of these two regions to the overall antigenicity of gp120 is uncertain. However, both regions have been identified as being potentially antigenic on the basis of surface probability and hydrophilicity (23). The absence of linear epitopes in either region is shown by the poor serological reactivity with synthetic oligopeptides containing V4 and V5 sequences and the low immunogenicity of such peptides upon vaccination (25). This does not rule out the possibility that V4 and V5 form conformational epitopes that are not mimicked by synthetic peptides. Furthermore, the antigenicity of both areas in vivo may be altered by posttranslational additions of N-linked oligosaccharide groups. gp120 is heavily glycosylated at N-linked but not at O-linked sites (14), and all 24 potential N-linked sites are glycosylated when recombinant $HIV_{HTLVIIIB}$ gp120 is expressed in mammalian cells (T. J. Gregory, C. K. Leonard, L. Riddle, J. R. Thomas, R. J. Harris, and M. W. Spellman, J. Cell. Biochem. **14D:**151, 1990). N-linked glycosylation can mask potential peptide epitopes (2, 32) or themselves form conformational epitopes (2, 7, 33, 35). Glycosylation of V4 and V5 may therefore serve to mask the relatively invariant intervening CD4 binding region. The absence of monoclonal antibodies to V4, C4, and V5 may thus be a reflection of the effectiveness of glycosylation in masking potential epitopes rather than a supposed low antigenicity of the underlying peptide sequences. The preponderance of glycosylation sites in the hypervariable regions, and the major alteration in the position and number of such sites by amino acid substitution and sequence reduplication, could therefore be interpreted as an evolutionary response by HIV to evade the immune system.

ACKNOWLEDGMENTS

We thank F. McOmish and A. Cleland for technical assistance. Samples were collected by the staff of the Haemophilia Centre, Edinburgh Royal Infirmary.

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

LITERATURE CITED

1. **Albert, J., B. Abrahamson, K. Nagy, E. Aurelius, H. Gaines, G. Nystrom, and E. M. Fenyo.** 1990. Rapid development of isolate-specific neutralising antibodies after primary HIV-1 infection and consequent emergence of virus variants which resist neutralisation by autologous sera. AIDS **4:**107–112.
2. **Alexander, S., and J. H. Elder.** 1984. Carbohydrate dramatically influences immune reactivity of antisera to viral glycoprotein antigens. Science **226:**1328–1330.

3. **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.
4. **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.
5. **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.
6. **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.
7. **Feizi, T., and R. A. Childs.** 1987. Carbohydrates as antigenic determinants. Biochem. J. **245**:1–12.
8. **Gnann, J. W., P. L. Schwimmbeck, J. A. Nelson, A. B. Truax, and M. B. A. Oldstone.** 1987. Diagnosis of AIDS by using a 12-amino acid peptide representing an immunodominant epitope of the human immunodeficiency virus. J. Infect. Dis. **156**:261–267.
9. **Goudsmit, J., C. A. B. Boucher, R. H. Meloen, L. G. Epstein, L. Smit, L. Van Der Hoek, and M. Bakker.** 1988. Human antibody response to a strain-specific HIV-1 gp120 epitope associated with cell fusion inhibition. AIDS **2**:157–164.
10. **Goudsmit, J., M. C. Debouck, R. H. Meloen, L. Smit, M. Bakker, D. M. Asher, A. V. Wolff, C. J. Gibbs, and D. C. Gajdusek.** 1987. Human immunodeficiency virus type 1 neutralisation epitope with conserved architecture elicits early type-specific antibodies in experimentally infected chimpanzees. Proc. Natl. Acad. Sci. USA **85**:4478–4482.
11. **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.
12. **Hill, R. E., and N. D. Hastie.** 1987. Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature (London) **326**:96–99.
13. **Hughes, A. L., and M. Nei.** 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc. Natl. Acad. Sci. USA **86**:958–962.
14. **Kozarsky, K., M. Penman, L. Basiripour, W. Haseltine, J. Sodroski, and M. Krieger.** 1989. Glycosylation and processing of the human immunodeficiency virus type 1 envelope protein. J. AIDS **2**:163–169.
15. **Laskowski, M., I. Kato, W. Ardelt, J. Cook, A. Denton, M. W. Empie, W. J. Kohr, S. J. Park, K. Parks, B. L. Schatzley, O. L. Schoenberger, M. Tashiro, G. Vichot, H. E. Whatley, A. Wieczorek, and M. Wieczorek.** 1987. Ovomucoid third domains from 100 avian species: isolation, sequences, and hypervariability of enzyme-inhibitor contact residues. Biochemistry **26**:202–221.
16. **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.
17. **Leigh Brown, A., and P. Monaghan.** 1988. Evolution of the structural proteins of human immunodeficiency virus: selective constraints on nucleotide substitution. AIDS Res. Human Retroviruses **4**:399–407.
18. **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.
19. **Looney, D. J., A. G. Fisher, S. D. Putney, J. R. Rusche, R. R. Redfield, D. S. Burke, R. C. Gallo, and F. Wong-Staal.** 1988. Type-restricted neutralization of molecular clones of human immunodeficiency virus. Science **241**:357–359.
20. **Ludlam, C. A., J. Tucker, C. M. Steel, R. S. Tedder, R. Cheingsong-Popov, R. A. Weiss, D. B. McClelland, I. Philip, 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.
21. **Melbye, M., K. S. Froebel, R. Madhok, R. J. Biggar, P. S. Sarin, S. Sternbjerg, G. D. O. Lowe, C. D. Forbes, J. J. Goedert, R. C. Gallo, and P. Ebbeson.** 1984. HTLV-III seropositivity in Europeon haemophiliacs exposed to Factor VIII concentrate imported from the USA. Lancet **ii**:1444–1446.
22. **Meloen, R. H., R. M. Liskamp, and J. Goudsmit.** 1989. Specificity and function of the individual amino acids of an important determinant of HIV-1 that induces neutralising antibody. J. Gen. Virol. **70**:1505–1512.
23. **Modrow, S., B. H. Hahn, G. M. Shaw, R. C. Gallo, F. Wong-Staal, and H. Wolf.** 1986. 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.
24. **Narayan, O., D. E. Griffin, and J. Chase.** 1977. Antigenic shift of visna virus in persistently infected sheep. Science **197**:376–378.
25. **Neurath, A. R., N. Strick, and E. S. Y. Lee.** 1990. B cell epitope mapping of human immunodeficiency virus envelope glycoproteins with long (19- to 36-residue) synthetic peptides. J. Gen. Virol. **71**:85–95.
26. **Palker, T. J., T. J. Matthews, M. E. Clark, G. J. Cianciole, R. R. Randall, A. J. Langlois, G. C. White, B. Safai, R. Snyderman, D. P. Bolognesi, and B. F. Haynes.** 1987. A conserved region at the COOH terminus of human immunodeficiency virus gp120 envelope protein contains an immunodominant epitope. Proc. Natl. Acad. Sci. USA **84**:2479–2483.
27. **Reitz, M. S., C. Wilson, C. Naugle, R. C. Gallo, and M. Robert-Guroff.** 1988. Generation of a neutralization-resistant variant of HIV-1 is due to selection for a point mutation in the envelope gene. Cell **54**:57–63.
28. **Rusche, J. R., K. Javaherian, 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.
29. **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.
30. **Simmonds, P., P. Balfe, J. F. Peutherer, C. A. Ludlam, J. O. Bishop, and A. 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.
31. **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.
32. **Skehel, J., D. J. Stevens, R. S. Daniels, A. R. Douglas, M. Knossow, I. A. Wilson, and D. C. Wiley.** 1984. A carbohydrate side chain on haemagglutinins of Hong Kong influenza viruses inhibits recognition by a monoclonal antibody. Proc. Natl. Acad. Sci. USA **81**:1779–1783.
33. **Sodora, D. L., G. H. Cohen, and R. J. Eisenberg.** 1989. Influence of asparagine-linked oligosaccharides on antigenicity, processing, and cell surface expression of herpes simplex virus type 1 glycoprotein D. J. Virol. **63**:5184–5193.
34. **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.
35. **Sugwara, K., F. Kitame, H. Nishimura, and K. Kakamura.** 1988. Operational and topological analyses of antigenic sites on influenza C virus glycoprotein and their dependence of glycosylation. J. Gen. Virol. **69**:537–547.
36. **Takahashi, H., S. Merli, S. D. Putney, R. Houghten, B. Moss, R. N. Germain, and J. A. Berzofsky.** 1989. A single amino acid interchange yields reciprocal CTL specificities for HIV-1 gp160. Science **246**:118–121.

37. **Tsang, T. C., and D. R. Bentley.** 1989. An improved method using Sequenase (tm) that is independent of template concentration. Nucleic Acids Res. **16**:6238.
38. **Willey, R. L., R. A. Rutledge, S. Dias, T. Folks, T. Theodore, C. E. Buckler, and M. A. Martin.** 1986. Identification of conserved and divergent domains within the envelope gene of the acquired immunodeficiency virus syndrome retrovirus. Proc. Natl. Acad. Sci. USA **83**:5038–5042.
39. **Williams, P., P. Simmonds, P. L. Yap, P. Balfe, J. O. Bishop, R. Brettle, R. Hague, D. Hargreaves, J. Inglis, A. Leigh Brown, J. Peutherer, S. Rebus, and J. Mok.** 1990. The polymerase chain reaction in the diagnosis of vertically transmitted HIV infection. AIDS **4**:393–398.
40. **Winship, P. R.** 1989. An improved method for directly sequencing PCR amplified material using dimethyl sulphoxide. Nucleic Acids Res. **17**:1266.