JOURNAL OF VIROLOGY, Oct. 1994, p. 6730–6744
0022-538X/94/$04.00+0

Vol. 68, No. 10

Mutational Trends in V3 Loop Protein Sequences Observed in Different Genetic Lineages of Human Immunodeficiency Virus Type 1

BETTE T. M. KORBER,[1,2]* KERSTI MACINNES,[1] RANDALL F. SMITH,[3] AND GERALD MYERS[1]

Theoretical Biology and Biophysics, Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545[1]; Santa Fe Institute, Santa Fe, New Mexico 87501[2]; and Department of Molecular and Human Genetics and Department of Cell Biology, W. M. Keck Center for Computational Biology, Baylor College of Medicine, Houston, Texas 77030[3]

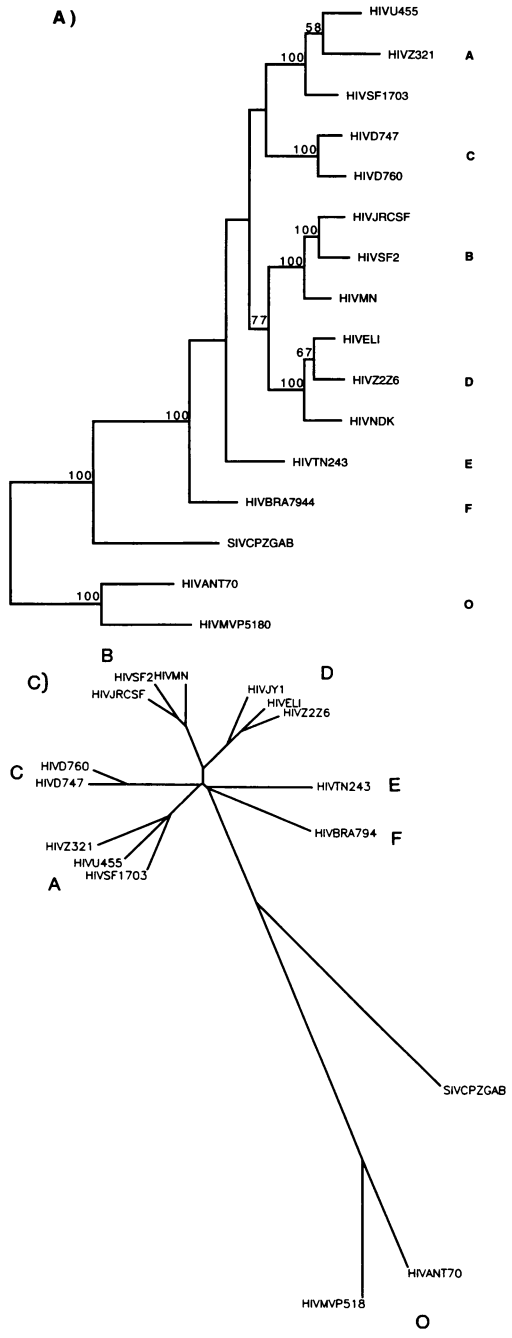
Received 24 March 1994/Accepted 25 July 1994

Highly variable international human immunodeficiency virus type 1 envelope sequences can be assigned to six major clades, or phylogenetically defined subtypes, designated A through F. These subtypes are approximately equidistant in terms of evolutionary distance measured by nucleotide sequences. This radiation from a common ancestral sequence may have been in step with the spread of the pandemic. In this study, V3 loop protein sequence relationships within these major clades are analyzed to determine how the different lineages might be evolving with respect to this biologically important domain. The V3 loop has been shown to influence viral phenotype and to elicit both humoral and cellular immune responses. To identify patterns in V3 loop amino acid evolution, we cluster the sequences by a phenetic principle which evaluates protein similarities on the basis of amino acid identities and similarities irrespective of evolutionary relationships. When phenetic clustering patterns are superimposed upon phylogenetic subtype classifications, two interesting mutational trends are revealed. First, a set of identical, or highly similar, V3 loop protein sequences can be identified within two otherwise dissimilar genetic subtypes, A and C. Second, the D subtype sequences are found to possess the most radically divergent set of V3 loop sequences. These and other patterns characteristic of the V3 loop reflect the acquisition of specific biological properties during the apparently recent evolution of the human immunodeficiency virus type 1 lineages.

The genome of human immunodeficiency virus type 1 (HIV-1) is rapidly evolving, and the resulting spectrum of genetic variation is being studied through the combined efforts of many groups worldwide. On the basis of studies of the viral envelope (*env*) gene, six nucleotide sequence subtypes, designated A through F, have been phylogenetically distinguished (51). There is an additional outlier category that includes two highly divergent HIV-1 sequences sampled from West Africa (28, 78). Furthermore, seven *gag* gene subtypes that correlate well with *env* gene subtypes have been identified (42). The major subtypes in both *gag* and *env* genes are approximately equidistant in terms of differences seen at the nucleotide level (51) (Fig. 1), with the exception of the B and D subtypes, which are slightly closer to each other than to the others (Fig. 1). Estimates of the rate of divergence of HIV-1 *env* sequences between infected individuals within a population suggest that they may be diverging at rates up to 1% per year (37, 50, 52). Using this estimate, one can hypothesize a look-back time to a common ancestral sequence for the six HIV-1 subtypes on the order of decades rather than centuries; this estimate is in accord with the epidemiologic history of the pandemic and with other estimates for the time of divergence between African and North American derived viral sequences (41, 82). Eigen and Neiselt-Struwe have traced the earliest node of the primate immunodeficiency viral sequences back 600 to 1,200 years (16), but this estimate does not address the temporal radiation of subtypes A through F, nor does it tell us what we might anticipate in terms of HIV-1 viral divergence over the next few decades.

One of the urgent questions of HIV research concerns the directions of change the virus will take in the course of its rapid evolution and whether distinct phenotypes are emerging among the global spectrum of variants. Hypotheses suggesting that HIV-1 will become either more or less pathogenic and transmissible as it evolves have been put forward (17, 18, 50, 76); it is possible that both potentials could be realized in different viral populations. Because much of the viral sequencing of international isolates has been carried out with viruses for which limited clinical and biological data have been available, it has not been possible to assess whether there are unique characteristics associated with viruses of different lineages in vivo. The duration of the asymptomatic phase of human infection can vary widely, making it difficult to know exactly how the clinical spectrum is influenced by the host, cofactors, and viral strains (77). One case study of a group of long-term asymptomatic Australians who were infected through a single contaminated blood sample has emphasized the importance of the viral strain (38). Nielsen and coworkers have reported a correlation between rapid clinical progression and the presence of syncytium-inducing (SI) HIV-1 strains at seroconversion, which also argues for the importance of the viral strain (54). It is difficult to determine not only whether there are differences in pathogenicity among different viral lineages but also whether transmissibility differs. In the Thai epidemic, heterosexual transmission rates per encounter were estimated to be far greater than transmission rates for the North American population; however, the authors of the analysis point out that the apparent efficiency in transmission could be explained by the timing of viral infection and cofac-

* Corresponding author. Mailing address: MS K710, Theoretical Biology and Biophysics (T10), Los Alamos National Laboratory, Los Alamos, NM 87545. Phone: (505) 665-4453. Fax: (505) 662-7517. Electronic mail address: btk@t10.lanl.gov.



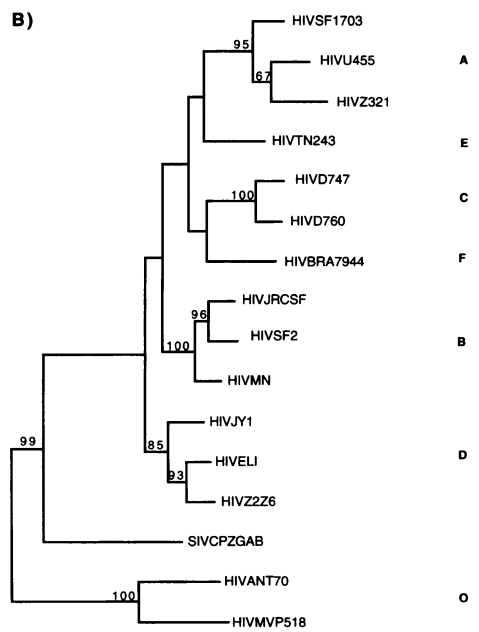


FIG. 1. Phylogenetic analysis of representative gp120 sequences from the six subtypes. The same input sequence alignment was used for the generation of phylogenetic trees by the weighted-parsimony (43, 74) (A), maximum likelihood (56) (B), and neighbor-joining (21) (C) methods. The weighted-parsimony tree incorporated $1/f$ of the relative frequencies of the character state changes shown in Table 1. The relative branch lengths shown in panels A and B were drawn from the most parsimonious (A) or the maximum likelihood (B) trees determined after 10 randomizations of the sequence input order. The numbers given at the branch points are the 50% threshold majority consensus values from 100 bootstrap replicates. Panels A and B were both drawn with the phylogenetic tree drawing tools of PAUP for direct comparison. The neighbor-joining tree (C) was drawn with the drawtree program in PHYLIP, since drawtree best illustrates the relative conservation of intersubtype branch lengths. These trees were based on an alignment of gp120 encoding nucleotide sequences from which columns containing gaps have been deleted, leaving 883 sites, 557 of which were varied.

tors and that there are insufficient data to implicate the viral genetic subtypes found in Thailand (44). Thus, clinical and epidemiological data are simply not yet sufficient to permit a full evaluation of the phenotypic trends among HIV-1 lineages. It is possible, however, to discern differences in mutational patterns in certain critical regions of the virus.

Because of the biological importance of the V3 region of the Env protein and, consequently, the large collection of coding sequences currently available for this region (51), V3 has been the topic of our twofold analysis of HIV-1 phylogenetic and phenetic relationships. With phylogenetics, the focus was upon the network of evolutionary relationships as manifested in V3 region nucleotide sequences. With phenetics, the focus was upon amino acid sequence similarities irrespective of evolutionary relatedness. For a discussion of cladistics versus phenetics, where cladistics is defined as the study of the pathways of evolution, see reference 40. Embedded in the V3 region is the V3 loop, which is approximately 35 amino acids long (30 to 40 amino acids, depending on the strain of HIV-1) and is bounded by a cysteine-to-cysteine disulfide bridge. V3 loop peptides are particularly immunogenic (25, 61, 66), and the loop structure also plays an essential role in virus-cell fusion (4, 5, 23, 24). Specific mutations on either side of the tip of the V3 loop can influence cellular tropism and SI capabilities of viruses in culture (12–14, 22, 67): SI, T-cell-tropic strains tend to have positively charged amino acids adjacent to the tip of the loop (13, 14, 22, 47). It is worth emphasizing that while mutations in the V3 loop can result in large phenotypic effects, the loop clearly functions in the context of the intact protein. Other regions in Env are also partly responsible for features of the viral phenotype influence by the V3 loop (2, 5, 8, 26). In particular, mutations in other regions of Env can affect ligand binding to the V3 loop and, similarly, mutations in the V3 loop can affect ligand binding to the other regions (46, 48, 62, 81). A growing body of evidence argues that non-SI, macrophage-tropic forms of the virus are the predominant forms detected

immediately postinfection and that SI strains emerge later during the course of the infection (11). These phenotypic properties appear to be dictated in part by the V3 loop sequences (83, 85).

The biological consequences of specific V3 loop mutations have generally been examined in North American and European isolates that are phylogenetically related members of the B subtype. It is now important to understand the range and implications of V3 loop mutations in the broader context of global variation. Hence, lineage-specific evolutionary patterns among the currently sequenced spectrum of HIV-1 variants were sought in order to define patterns of conservation and divergence among viral subtypes. When phylogenetic subtype designations based on longer nucleotide sequences were superimposed on shorter V3 protein clustering patterns to identify common sequence elements in genetically distant viruses, the phenetic clustering of V3 loop amino acid sequences did not always show a correspondence with the phylogenetic analysis, that is, the association of sequences with shared genetic lineages. There were more distinct kinds of V3 loop peptide clustering patterns (at least 14) than there were nucleotide sequence subtypes (6). This lack of correspondence has implications for HIV-1 serotyping and vaccine design. Specifically, the phenetic analyses led to the identification of a form of the loop which is highly similar among specific A and C subtype viruses despite 30% differences at the *env* gp120 nucleotide coding sequences. This shared form represents an apparently stable structure preserved through parallel evolution in the two separate lineages. Differences in the rates of V3 loop nonconservative amino acid substitutions in different monophyletic groupings were also observed: specifically, intrasubtype sequence comparisons show that D subtype V3 loop sequences are radically divergent relative to the other subtypes.

MATERIALS AND METHODS

Sequence sets. C2V3 sequences representing 302 individuals were derived from the collection of published sequences currently available in the Human Retroviruses and AIDS 1993 database (51). The database sequence locus names have been preserved, and detailed references and methods for generating the set can be found in the compendium (51). Only one sequence per individual was included in this set; consensus sequences were used when multiple viral sequences from the same individual were available. In studies of linked transmission cases, only the recipient was included. A minimum of 200 nucleotides was required for inclusion in the phylogenetic analyses. To construct the consensus, the most common amino acid in a given position was used; if there were equal numbers of two or more amino acids in a column, then the first one that appeared in the alignment was used. Phylogenetic analyses were performed by comparing well-characterized, complete or nearly complete gp120 sequences aligned with the sets of C2V3 region sequences. Generally, the individual trees used for subtype classifications were based on a set of unknown sequences from a single publication compared with standard reference sequences for which the subtypes were known.

This sequence set has several important limitations. (i) The time from seroconversion and health status of patients were not considered (this information often was not available for international sequences). (ii) A consensus sequence may not perfectly represent any of the actual sequences found in an infected individual. (iii) The sampling was not systematic; rather the pool of publicly available sequences was included. And (iv) no distinction is made between cultured viral sequences and those obtained directly from blood samples. On the other hand, this set has compensatory virtues when one is trying to comprehend the vast array of C2V3 regions sequences available to date. (i) The viral sequence of no single infected individual is weighted too heavily, since only one sequence per person is included. (ii) By using consensus sequences from individuals when possible, there is less opportunity for the inclusion of an inappropriate amino acid in a sequence due to sequencing errors or sequencing of nonviable virus. (iii) The set is as internally consistent as possible, given that many of the international sequences are a single direct sequence of PCR-amplified products of peripheral blood DNA (experimental consensus). And (iv) it provides a systematic overview of the spectrum of international sequences currently available. The entire C2V3 database includes thousands of sequences, with the number of HIV-1 sequences from a single individual ranging from 1 to more than 100.

TABLE 1. Character state changes[a]

Base	No. of observed changes[b] to:			
	A	C	G	T
A		144 (108–187)	283 (222–345)	117 (89–150)
C	85 (53–122)		41 (25–64)	141 (108–179)
G	197(144–258)	37 (19–60)		42 (22–67)
T	88 (57–120)	165 (127–206)	62 (40–92)	

[a] The character state changes are based on PAUP-generated parsimony trees, using the sequences included in the trees in Fig. 1. This table was based on the two most parsimonious trees found, using 10 randomizations of the input order. Weighting a subsequent parsimony run on the basis of these values and recalculating the frequencies of character state changes on the basis of the weighted tree did not alter the values significantly.

[b] The average number of observed changes of each type is followed by the minimum and maximum values shown in parentheses.

The *gag* gene sequence set and phylogenetic subtype designations are those determined by Louwagie et al. (42) and taken from the Human Retroviruses and AIDS database (51).

Phylogenetic analysis and distance measures. Alignments and similarity analyses (simple distance measurements) were generated by using MASE (19). Several approaches were taken for the phylogenetic analysis of HIV nucleotide sequences. Because of the highly skewed base composition of HIV and the asymmetrical substitutional frequencies of mutations from one base to another (Table 1), we are most confident of the trees determined by a weighted-parsimony approach (31, 34) that used PAUP 3.1.1 (74) in conjunction with MacClade version 3.03 (43). To generate weighted-parsimony trees, Table 1, which summarizes the frequencies of character state changes (f), was produced from a preliminary parsimony analysis using PAUP and MacClade. The conversion $1/f$ was used to weight the possible nucleotide changes, with truncation to avoid violations of the triangle inequality (43). The resulting matrix was then included as a character type assumption for subsequent phylogenetic reconstructions using PAUP. The construction of a new most parsimonious tree minimized the sum of the branch lengths calculated as the number of character changes multiplied by the weights of the respective character changes. The g_1 statistic for the phylogenetic analysis using parsimony was very low (-1.9), indicating that there is a strong signal in the data (33) and that parsimony is an appropriate phylogenetic tool for application to this sequence set. Furthermore, the phylogenetic reconstructions obtained by weighted parsimony had results very similar to those from other methods applied to this particular sequence set. The other phylogenetic methods used were neighbor joining, based on a Kimura

two-parameter distance matrix generated with the PHYLIP package (21), and maximum likelihood using fastDNAml (56). Bootstrap values for the maximum-likelihood and weighted-parsimony trees were calculated.

Synonymous and nonsynonymous substitution rates were calculated on the basis of the method of Nei and Gojobori (53). *Ps* is the number of observed divided by the number of possible synonymous substitutions; *Pn* is the number of observed divided by the number of possible nonsynonymous substitutions. The synonymous substitution rate, *ds*, is the Jukes-Cantor transformation of *Ps*: $ds = 3/4ln(1 - 4/3Ps)$; *dn* is analogously calculated from *Pn* (53).

Phenetic analysis. The phenetic sorting methods described herein are based on an adaptation of the PIMA program, originally designed for generation of protein alignments (69, 70). This adaptation provided the option of choosing different protein similarity matrices for the sake of comparison. Sequences were compared in a pairwise fashion, with gaps inserted as needed by PIMA to achieve the best score. The PIMA amino acid similarity matrix was based on a simple hierarchical scheme: perfect identity between two sequences in a position in an alignment is given a score of 5; chemically very conservative substitutions are given a score of 4, while less conservative substitutions are given a score of 3, etc. (69). Insertions are penalized, with a default value ($1X$) of -6.67 for the inserted gap and -1.33 for gap extensions. Different gap scoring schemes used either $0.0X$, $0.5X$, or $2X$ of the default values. The score at each position was totaled across the sequence; for example, a perfect identity between two V3 loop sequences of 35 amino acids yields a PIMA score of 175 (35×5). We also substituted a BLOSUM60 matrix (29) and an STR matrix (30) in the PIMA program for comparison of the phenetic clustering patterns from these schemes. The BLOSUM matrix was calculated by using observed amino acid frequencies in alignments generated from a large database of conserved protein sequences (29). The STR matrix (30) was based on observed amino acid frequencies in a structure-based alignment of otherwise variable proteins (60). The original PIMA program for sequence alignment uses maximal linkage to arrange the sequences; PIMA-generated similarity scores were also put into UPGMA and neighbor-joining clustering programs (21), basing the distance scores for the PHYLIP infile matrix on the PIMA score for a perfectly homologous pair of sequences minus the observed score. Phenograms based on the cluster output from PIMA were generated by using the programs drawtree-jt and XYPLOT.

RESULTS

Phylogenetic distances between major HIV-1 genetic subtypes. For *env* sequences, the average nucleotide branch lengths separating the various "leaves" or taxa of HIV-1 subtypes are remarkably similar (Fig. 1). All three phylogenetic reconstruction methods tested—weighted parsimony, neighbor joining, and maximum likelihood—give this result. The only exception to the conservation of intersubtype distances is found between B subtype sequences (associated with U.S., European, and some Asian samples) and D subtype sequences (associated with central African samples), which appear to be closer to each other in terms of genetic distance than to other subtypes (Fig. 1 and Table 2). The general conservation in distances between HIV-1 clades is also seen in trees based on *gag* gene sequences (phylogenetic trees not shown; Table 3); B and D sequences also group slightly closer together in *gag*.

Bootstrap tests of the relationships of gp120 sequences shown in Fig. 1 strongly support the notion of the subtype associations of the individual sequences (20, 32). A bootstrap test of the nodes generated by weighted-parsimony trees including 100 resamplings gave 100% recurrence of the nodes representing the branch points that define the major subtypes, A through D, for which multiple gp120 sequences were available (Fig. 1A). Similarly, a bootstrap test of the subtype-defining nodes using maximum likelihood trees gave bootstrap values of 85 to 100% for the subtype associations of the sequences. Therefore, the likely divisions between the major clades were strongly indicated by both phylogenetic reconstruction methods. The branching order of the clades relative to the outlying HIV-1 group O and chimpanzee viral sequences by the two methods differed, however, and the structure of the tree at this level was not supported by bootstrap analysis by either method. The equidistance between the different clades is clearly seen in the neighbor-joining method-based tree, shown in Fig. 1C.

Another way to measure relatedness and nucleotide distances among the subtypes is through examination of synonymous and nonsynonymous substitution frequencies, as summarized in Tables 2 and 3 (39, 53). The distances between clades, as measured by synonymous changes between each of the individual clade members and members of all other clades, are remarkably consistent, in both *env* and *gag*. The average number of synonymous substitutions which occur between sequences in different clades (excluding the O, or outlier, group) is 33 to 35% in *env* and 29 to 34% in *gag*. Increased similarity is observed between clades B and D in *gag*, with average synonymous substitutions of 21%. Intrasubtype distances are also shown in Tables 2 and 3; however, these values are heavily biased by the sample selection.

The nonsynonymous substitution rates are two- to threefold greater in *env* than they are in *gag*, as reflected in the ratios of synonymous-to-nonsynonymous substitution rates (*ds/dn*) for the two genes (39). As more distant *env* sequences are sampled, the prominence of nonsynonymous substitution diminishes: intracladal *ds/dn* ratios are between 1 and 2, and intercladal ratios are between 2 and 3. Further evidence for this effect was observed within E intrasubtype V3 region sequence comparisons. Thai E subtype sequences, which are phylogenetically tightly clustered as a result of the recent introduction of HIV-1 into Thailand (45, 59), manifest a very high rate of change in that the *ds/dn* ratio is quite low (0.5) because nonsynonymous substitutions predominate. For comparison, among the more highly diverged Central African Republic E subtype sequences (49), with interpatient distances comparable to interpatient distances found among B subtype sequences in the United States (58), the *ds/dn* ratio is about 1.5. Synonymous substitutions between the outlier HIV-1 and chimpanzee (CPZ) sequences, and between outliers and CPZs when compared with subtypes A through F, are clearly saturated; therefore, the *ds/dn* ratios at these genetic distances have little meaning. The distances between the two outlier sequences HIVANT70 and HIVMVP5180, however, are curiously comparable to the synonymous and nonsynonymous distances measured for intercladal relationships among subtypes A through F.

The B subtype comparisons revealed an uncannily high level of Gag protein conservation between two sequences, HIVCAM1, a British isolate, and HIVSF2, a U.S. isolate, which nevertheless had multiple synonymous substitutions (*Ps* = 0.10) (Table 3). The exceptional range of the B subtype sequences in Table 3 reflects this unusual relationship: the *ds/dn* ratio was 29.6 for this pair of sequences, and the value of *Pn* was 0.003. There were only three amino acid differences, all conservative changes, between the two sequences in p17 and

TABLE 2. Medians and ranges for pairwise comparisons of sequences showing inter- and intrasubtype similarity relationships between gp120 *env* sequences[a]

Compared value	Subtype	Interclade avg	Median (range) of pairwise comparisons for *env* subtype comparison:						
			A	B	C	D	E	F	O
ds/dn	A	2.6	2.1 (1.9–2.1)	2.6 (2.1–3.7)	2.5 (2.2–3.3)	2.6 (2.2–3.2)	2.3 (2.3–2.5)	3.0 (2.5–3.4)	6.4 (4.5–sat)
	B	2.4		1.4 (0.7–3.2)	2.4 (1.8–3.2)	1.9 (1.3–2.8)	2.6 (2.0–3.1)	2.7 (2.3–3.2)	4.5 (3.3–sat)
	C	2.6			1.2 (0.7–1.2)	2.7 (2.2–3.2)	2.7 (2.4–2.9)	2.7 (2.5–2.8)	3.9 (3.1–5.2)
	D	2.5				1.8 (1.1–2.8)	3.2 (2.3–3.4)	2.7 (2.2–2.7)	5.5 (5.1–sat)
	E	2.7						3.0	7.1 (5.0–9.2)
	F	2.8							2.7 (2.5–sat)
	O								5.5
Pn	A	0.16	0.11 (0.11–0.12)	0.16 (0.13–0.19)	0.16 (0.13–0.20)	0.17 (0.14–0.18)	0.17 (0.15–0.17)	0.16 (0.15–0.18)	0.31 (0.26–0.38)
	B	0.15		0.08 (0.04–0.12)	0.16 (0.14–0.18)	0.15 (0.13–0.17)	0.15 (0.14–0.17)	0.15 (0.13–0.16)	0.32 (0.26–0.38)
	C	0.15			0.08 (0.05–0.12)	0.16 (0.14–0.19)	0.15 (0.15–0.18)	0.14 (0.14–0.16)	0.33 (0.27–0.37)
	D	0.16				0.11 (0.07–0.15)	0.16 (0.14–0.17)	0.16 (0.14–0.17)	0.31 (0.25–0.37)
	E	0.16						0.14	0.31 (0.26–0.35)
	F	0.15							0.26 (0.18–0.35)
	O	0.31							0.38
Ps	A	0.35	0.22 (0.21–0.22)	0.35 (0.30–0.40)	0.35 (0.33–0.37)	0.35 (0.33–0.39)	0.34 (0.30–0.36)	0.38 (0.37–0.39)	0.74 (0.65–0.76)
	B	0.33		0.11 (0.05–0.22)	0.32 (0.27–0.41)	0.33 (0.30–0.36)	0.33 (0.29–0.36)	0.33 (0.30–0.36)	0.69 (0.62–0.75)
	C	0.34			0.09 (0.03–0.13)	0.36 (0.32–0.42)	0.35 (0.34–0.36)	0.34 (0.31–0.34)	0.67 (0.65–0.68)
	D	0.34				0.19 (0.14–0.23)	0.37 (0.33–0.41)	0.34 (0.32–0.38)	0.72 (0.66–0.79)
	E	0.34						0.35	0.70 (0.66–0.75)
	F	0.34							0.64 (0.39–0.75)
	O	0.69							0.73

[a] The matrices show the ratios of the synonymous substitution rates to nonsynonymous substitution rates (*ds/dn* values); the number of synonymous substitutions divided by potential synonymous substitutions (*Ps*); and the number of nonsynonymous substitutions divided by potential nonsynonymous substitutions (*Pn*). Potential synonymous substitutions in HIV-1 gp120 are about 21% of the possible substitutions, and nonsynonymous substitutions make up about 79%. sat indicates that the *Ps* value was too high to calculate a *ds* value and the synonymous substitutions are saturated. These values were all calculated by the methods described by Nei and Gojobori (53). The interclade averages were calculated excluding comparisons to outlier sequences. The alignments used were based on the alignments provided in the Human Retroviruses and AIDS database (51). Highly variable regions with multiple insertions and deletions were excised from the alignments because these regions are difficult to align with confidence. Therefore, the nonsynonymous substitutions are systematically underestimated and the numbers shown can be considered appropriate estimates for the relatively conserved regions of gp120. Codons that contained an ambiguous base or a deletion in one of the two sequences being compared were discounted. The boundaries of the gp120 region examined were dictated by the borders of the available Env F subtype sequence, HIVBR7944, which was kindly provided by Marcia Kalish of the Centers for Disease Control and Prevention, and included 245 codons, or 735 nucleotides. The region begins with amino acid 120 in the Env sequence HIVMN (51) and ends at amino acid 396 (bounded by the amino acid strings: KLTPLC through SPLFNS, with the first K at 120 and the last S at 396). Sequences included in the Env comparisons are the following: A subtype HIVU455, HIVZ321, and HIVSF170; B subtype HIV1D31, HIVADA, HIVALA1, HIVBAL1, HIVBRVA, HIVCAM1, HIVCDC42, HIVHAN, HIVJFL, HIVJH32, HIVJRFL, HIVLAI, HIVMN, HIVNY5CG, HIVOYI, HIVRF, HIVSC, HIVSF162, HIVSF2, HIVSF33, and HIVWMJ22; C subtype HIVD747, HIVD757, HIVNOF, and HIVD760; D subtype HIVELI, HIVJY1, HIVMAL, HIVNDK, and HIVZ2Z6; E subtype HIVTN243; F subtype HIVBR7944; and outliers (O) HIVANT70 and CPZGAB. See reference 51 for complete sequence references.

p24. This rare occurrence points to preservation of a viral protein that is probably significant in terms of structure and function, analogous to what we shall report for V3 loop conservation of A and C subtypes (as discussed below). The number of synonymous substitutions separating the two *gag* coding sequences makes linked transmission or cross-contamination of samples highly unlikely. Also, there was nothing extraordinary about the relative numbers of substitutions for the two sequences observed in the gp120 gene (*Ps* = 0.13, *Pn* = 0.07, and *ds/dn* = 1.9).

Linear correlation analysis to determine if the use of PCR gene fragments gives reasonable estimates of sequence distance relationships determined by intact genes. For phylogenetic analysis using genetic information, full-length coding sequences typically yield a more accurate analysis of sequence relationships than do PCR-amplified short gene fragments (75). Because of the greater cost and technical difficulty of obtaining complete gene sequences, many of the international HIV-1 *env* sequences available to date are PCR-amplified fragments 250 to 400 nucleotides long that cover the C2V3 region of *env* (51). To determine how well sequence relationships between such fragments represent the relationships between larger regions encoding envelope protein gp120, linear correlation analysis comparing pairwise distances between the C2V3 region, excised from intact gp120 sequences, to the cognate gp120 *env* sequences with the C2V3 region deleted was performed (Fig. 2). The boundaries of the C2V3 region were selected on the basis of being typical of sequences obtained by the Centers for Disease Control and Prevention in their efforts to survey international variation (63). The results shown in Fig. 2 indicate that the C2V3 region fragments are predictive of the sequence distances calculated for genes in which they are embedded. Because the C2V3 region fragment is evolving at a higher rate than the longer gp120 fragment, the relationship is linear but is not one to one. Similar results were obtained for a p17 fragment of the *gag* gene when compared with the rest of the *gag* gene (data not shown). Thus, for purposes of subtype classification, C2V3 and p17 gene fragments provide reasonably accurate accounts of the genetic lineages that would be derived for intact genes. In Fig. 2, the distinct cluster of points between 18 and 35% divergent in C2V3 represent intersubtype distances and the points less than 18% divergent represent intrasubtype distances, with the two clusters of points being quite distinctive. In a very limited number of cases, distance relationships determined on the basis of the fragments can be misleading, however, and this effect may be important in the range of sequence dissimilarity that borders the closest intersubtype distances and the greatest intrasubtype distances. For example, sequence distances of 15% obtained by Jukes-Cantor transformation in the C2V3

TABLE 3. Medians and ranges for pairwise comparisons of sequences showing inter- and intrasubtype similarity relationships for the p17 and p24 coding regions in *gag*[a]

Compared value	Subtype	Interclade avg	Median (range) of pairwise comparisons for p17 + p24 *gag* subtype comparison:							
			A	B	C	D	F	G	H	O
ds/dn	A	6.3	4.4 (1.4–8.5)	6.3 (4.1–9.1)	5.7 (3.8–8.5)	6.3 (4.0–9.4)	7.9 (5.3–11.1)	6.0 (4.2–7.7)	5.6 (3.5–7.6)	8.5 (6.0–sat)
	B	5.5		5.1 (2.43–29.6)	5.0 (3.7–7.2)	5.7 (3.3–8.5)	5.8 (4.8–7.3)	5.6 (4.4–7.3)	4.7 (3.6–5.3)	9.6 (7.1–sat)
	C	5.4			2.9 (2.2–4.1)	5.5 (4.0–7.2)	7.1 (5.4–8.2)	5.6 (5.1–6.8)	4.2 (3.4–5.4)	7.7 (6.2–sat)
	D	6.3				4.9 (2.5–9.1)	6.4 (4.6–9.1)	5.5 (4.2–7.3)	4.9 (3.2–5.9)	9.3 (6.5–sat)
	F	6.3					5.7 (5.4–6.6)	8.4 (7.1–10.9)	6.3 (5.8–6.9)	8.4 (8.8–11.8)
	G	6.2						3.3 (3.3–3.3)	5.8 (5.0–6.4)	10.8 (8.8–12.7)
	H	5.3							4.1 (4.1–4.1)	9.5 (7.3–sat)
	O									12.1 (7.3–16.2)
Pn	A	0.07	0.05 (0.01–0.10)	0.07 (0.04–0.10)	0.08 (0.06–0.11)	0.07 (0.04–0.11)	0.06 (0.05–0.10)	0.07 (0.06–0.10)	0.07 (0.05–0.11)	0.17 (0.14–0.23)
	B	0.07		0.02 (0.003–0.04)	0.07 (0.06–0.08)	0.04 (0.03–0.06)	0.06 (0.05–0.07)	0.08 (0.06–0.09)	0.07 (0.07–0.08)	0.18 (0.13–0.21)
	C	0.07			0.04 (0.03–0.06)	0.06 (0.05–0.09)	0.06 (0.06–0.07)	0.08 (0.06–0.08)	0.08 (0.06–0.09)	0.18 (0.13–0.21)
	D	0.06				0.04 (0.02–0.06)	0.06 (0.04–0.08)	0.07 (0.06–0.09)	0.07 (0.06–0.09)	0.17 (0.13–0.21)
	F	0.06					0.03 (0.02–0.03)	0.06 (0.05–0.07)	0.07 (0.06–0.07)	0.18 (0.05–0.23)
	G	0.07						0.05 (0.05–0.05)	0.07 (0.06–0.07)	0.16 (0.12–0.20)
	H	0.07							0.05 (0.05–0.05)	0.17 (0.14–0.20)
	O	0.17								0.21 (0.05–0.23)
Ps	A	0.34	0.20 (0.01–0.30)	0.33 (0.25–0.43)	0.35 (0.28–0.42)	0.34 (0.24–0.43)	0.38 (0.28–0.44)	0.33 (0.27–0.38)	0.31 (0.26–0.36)	0.68 (0.63–0.80)
	B	0.29		0.11 (0.04–0.17)	0.29 (0.24–0.34)	0.21 (0.14–0.27)	0.27 (0.23–0.32)	0.34 (0.31–0.39)	0.29 (0.25–0.32)	0.69 (0.64–0.77)
	C	0.32			0.12 (0.08–0.15)	0.29 (0.27–0.35)	0.35 (0.31–0.38)	0.34 (0.32–0.35)	0.28 (0.25–0.30)	0.68 (0.62–0.77)
	D	0.30				0.17 (0.06–0.23)	0.31 (0.28–0.34)	0.33 (0.29–0.38)	0.29 (0.25–0.32)	0.70 (0.62–0.78)
	F	0.32					0.13 (0.11–0.16)	0.38 (0.34–0.40)	0.32 (0.30–0.35)	0.73 (0.29–0.81)
	G	0.34						0.16 (0.16–0.16)	0.32 (0.29–0.34)	0.69 (0.66–0.72)
	H	0.30							0.18 (0.18–0.18)	0.68 (0.65–0.73)
	O	0.69								0.76 (0.29–0.81)

[a] The methodology was the same as the described in Table 2, footnote *a*. The included *gag* sequence was bounded by the coding regions of the p17 and the p24 proteins, excluding the *gag* sequence that shares an overlapping reading frame with the *pol* gene. An unusual feature of these results is the extraordinary range of the B intrasubtype comparisons. This was due to the conservation of the protein sequences HIVCAM1 and HIVSF2. Sequences included in the *gag* comparisons are the following: A subtype HIVU455, HIVMAL, HIVVI59, HIVVI310, HIVVI57, HIVK112, HIVK88, HIVK29, HIVK124, HIVK7, HIVK98, HIVK89, HIVVI32, HIVVI415, HIVCI4, HIVG141, HIVLBV23, HIVTN243, HIVTN245, HIVTN240, HIVCI20, HIVCI59, HIVLBV2310, HIVCI51, HIVLBV105, HIVCI32, HIVIC144, HIVDJ258, HIVCM238, HIVUG266, and HIVVI354; B subtype HIVSF2, HIVBZ167, HIVPH153, HIVPH136, HIVBZ200, HIVTB132, HIVBZ190, HIVLAI, HIVMN, HIVJH31, HIVJRCSF, HIVOYI, HIVNY5CG, HIVCDC41, HIVHAN, HIVCAM1, HIVRF, HIVD31, HIVUG280, and HIVYU2; C subtype HIVUG268, HIVSM145, HIVZAM18, HIVZAM19, HIVZAM20, HIVDJ259, and HIVVI313; D subtype HIVELI, HIVNDK, HIVZ2Z6, HIVNDK, HIVVI205, HIVG109, HIVK31, HIVUG274, HIVUG270, HIVSE365, and HIVVI203; F subtype HIVVI174, HIVVI69, and HIVBZ162; G subtype HIVLBV217 and HIVVI191; H subtype HIVVI525 and HIVVI557; and outliers (O) HIVANT70, HIVMVP5180, and CPZGAB. See reference 51 for complete sequence references.

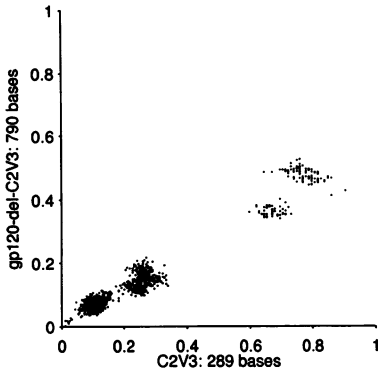


FIG. 2. Pairwise sequence distance comparisons of C2V3 region fragments with complete gp120 sequences. Pairwise sequence distances were determined for a commonly sequenced fragment of *env*, 289 nucleotides of the C2V3 region, and gp120 sequences from which the C2V3 regions were excised, leaving 790 nucleotides (gp120-del-C2V3). Sequence distances were calculated as $d = 1 - s$, where s is the proportion of shared nucleotides between the two sequences, and then modified by the Jukes-Cantor formula. Columns in the alignment that contained an insertion incorporated to maintain the alignment were deleted. The linear correlation coefficient, r, was 0.983 (65), and N, the number of pairwise combinations, was 1,326. Because the N value given is the result of all pairwise combinations of sequences, not all points are independent and the use of this statistic is not completely valid. Therefore, we recalculated the r values for several sets of single sequences compared with all others, and using these unlinked data sets with much smaller values of N, we obtained similarly high r values (for example, using MN compared with all others, $r = 0.974$). The least divergent cluster of points represents intracladal comparisons; the second cluster of points represents intercladal distances. The most divergent cluster of points is made up of the highly diverged HIV-1 sequences HIVMVP5180 and HIVANT70 (15, 28, 55, 78) compared with all others, and the second most divergent cluster of points represents the distances between the chimpanzee viral sequence CPZGAB and HIV-1 sequences. As an added statistical test, we removed the two clusters of highly diverged sequence datum points, which have great impact on the calculation of the linear correlation coefficient, from the original data set. Even with these points removed, we calculated a high linear correlation coefficient ($r = 0.949$). The slope of a line drawn through these points is approximately 0.6, indicating that the C2V3 region fragment is evolving at a higher rate than the gp120-del-C2V3 fragment.

region correspond to distances of 5 to 10% in gp120 (Fig. 2). Such ranges indicate that on occasion there could be inappropriate phylogenetic subtype associations of certain sequences when C2V3 region fragments are used, particularly with methods that depend on simple distance measures.

Another way to inspect the phylogenetic information of particular regions is to ask to what degree the subtype-defining branch points of interest in *env* phylogenetic trees reflect the information in short C2V3 fragments when compared with the information available in intact *env* genes. Neighbor-joining (21), parsimony (74), and maximum likelihood methods (21) reconstructed the major subtype groupings by using either the short C2V3 fragments or the remainder of the genes from which the fragments had been deleted (gp120-del-C2V3) (data not shown), further suggesting that C2V3 region fragments were suitable for basic genetic subtyping. Despite the reproducibility of the reconstruction of the major HIV-1 subtypes, the clustering patterns within a subtype varied dramatically, depending on both the input region of the gene used for the analysis and the phylogenetic reconstruction method employed. Therefore, while short C2V3 fragments were useful for subtype classification, they did not clearly define sequence relationships within clades.

Phenetic analysis of V3 loop protein sequences interpreted in the context of phylogenetic subtyping. Phylogenetic characterization for the purpose of genetic subtyping of international C2V3 region sequences was initially reported in the 1992 and 1993 Human Retroviruses and AIDS database compendiums (see Materials and Methods and reference 51). If protein similarity-based cluster analysis is now performed with equally long stretches of protein sequence, (i.e., the intact C2V3 region), the phenetic clustering patterns parallel the phylogenetic subtype classifications, as would be expected (data not shown). The phenetic analysis presented here goes further in that it is based on short variable stretches of amino acid sequence of a biologically important domain. In this context, clustering based on protein sequence similarity may reflect associations dictated by evolutionary conservation of structure rather than of genetic lineage.

For the phenetic analysis, the protein sequences were compared on the basis of three amino acid similarity scoring schemes to evaluate the overall protein conservation of the V3 loop: the amino acid class-covering scheme used in PIMA, based on chemical properties of amino acids (69); the BLO SUM60 matrix, based on the frequency of particular amino acid substitutions found in alignments of conserved protein regions (29); and the STR matrix that was calculated similarly to a BLOSUM matrix but used an alignment based on structural similarities found in a variety of proteins (30). Once a set of protein similarity scores was obtained, the scores were converted to distances and different clustering algorithms (maximum linkage, UPGMA, and neighbor joining) were used to group the protein sequences. We performed phenetic clustering analysis on three stretches of the V3 protein: first, that bounded by the disulfide bridge at the base of the loop (the complete loop, 30 to 40 amino acids); second, the loop plus the glycosylation sites in both flanks (31 to 43 amino acids); and third, an internal fragment of the loop that contains the primary neutralizing epitopes and phenotypic determinants (15 amino acids). This last fragment aligns with the sequence SIHIGPGRAFYTTGE of the B subtype V3 consensus sequence. The use of different scoring systems, clustering methods, or protein sequence fragments gave rise to differences in the fine structure of the final clustering pattern. The clustering patterns and protein alignments shown in Fig. 3 and 4 were generated by using the PIMA scoring system with a gap penalty of $0.5X$, a gap penalty selected because it resulted in clustering patterns that most closely correlated with the genetic subtype designations. Clustering patterns of sequences with unusual insertions and deletions were particularly sensitive to the gap penalty used; $0.5X$ improved the pattern for particular pairs of sequences that were coupled on the basis of having insertions or deletions of several amino acids but that were clearly inappropriately coupled in terms of the remaining sequence.

We will restrict our comments to two general observations that were robust, that is to say independent of the scoring scheme, clustering method, and V3 loop protein fragment boundaries used. A striking feature of the cluster analysis based on protein similarity scores shown in Fig. 3 is that D subtype V3 loop sequences possess the most divergent forms of the V3 loop. This excess divergence is shown in the greater depth of the branches among D subtype sequences; the actual V3 loop sequences, aligned by PIMA to correspond to the phenogram of Fig. 3, are shown in Fig. 4. As discussed above, B and D subtype sequences are the most genetically similar of the subtypes in both *gag* and *env*, suggesting that they may have diverged most recently. Yet, the intersubtype relationships

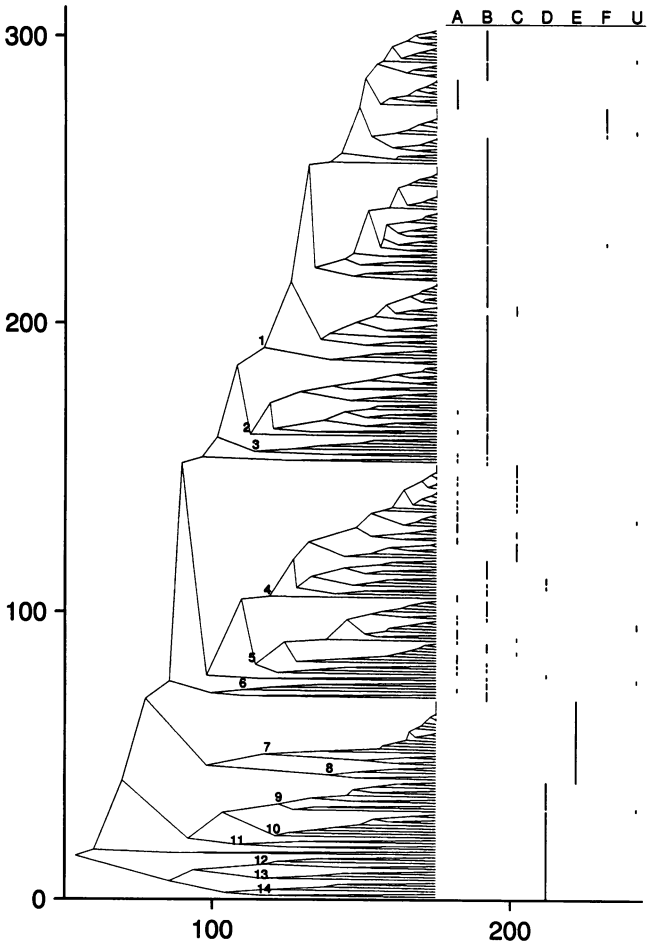


FIG. 3. HIV-1 V3 loop protein sequence relationships. The phenogram of HIV-1 V3 loop protein similarities is based on PIMA amino acid similarity scores obtained by using the maximum-linkage clustering algorithm (69, 70). The abscissa indicates the level of protein amino acid similarity; two identical V3 sequences 35 amino acids long would have a score of 175. Chemically conservative changes decrease the score less than nonconservative changes. The nodes in these phenograms do not reflect common ancestry but rather an amino acid class covering pattern (69, 70) that is general enough to accommodate all of the sequences which branch from that point. Each of the leaves on the phenogram indicates a particular sequence; the ordinate simply records the number of sequences analyzed. The genetic subtype association of each of the sequences is indicated by a mark under the columns labeled A through F for the different subtypes, and U for unclassified; the phylogenetic association could not be clearly determined for the unclassified sequences (51). The alignment of sequences shown in Fig. 4 corresponds to the organization of sequences shown here, with the top sequence in the phenogram being the top sequence in the alignment. The numbers on the branch points indicate clusters that were used to organize the sequences into phenetically associated groups for consensus sequence generation in Fig. 4.

characteristic of their V3 loop protein sequences are not similar (Fig. 3 and Table 4). And notwithstanding the higher level of protein variation in the D subtype V3 sequences, levels of intracladal variation between D subtype C2V3 region nucleotide sequences are not remarkable—a comparable range of intrasubtype nucleotide distances is manifest in both A and D subtype sequences (Table 4). The extent of the amino acid sequence variation between the sampled forms in the V3 loop in the D subtype amino acid sequences suggests that there may be different biological pressures shaping the evolution of the D clade V3 loops. Several features of this variation are summarized in Table 4. The median PIMA and STR protein similarity scores among pairwise comparisons of D subtype sequences are much lower than those for comparisons between any of the other sequence subtypes (Table 4). D subtype sequences also display a greater positive charge at the tips of their loops (Table 4), consistent with the observation that positively charged V3 loop sequences tend to show the greatest diversity among B subtype sequences (47). In Table 4, the greater positive charge is reflected in both the net charge and pI values given. The charge calculations were based on the 15-amino-acid internal fragment rather than the whole loop, because the internal fragment was better at distinguishing SI versus non-SI viruses in the set studied by Fouchier et al. (22) (data not shown). The greatest range of V3 loop length variation is seen in D subtype sequences (Table 4). Finally, the *ds/dn* ratios calculated for the D subtype sequences are far lower in the V3 loop than in the V3 loop flanking region, suggesting that the high rate of change within D subtype amino acid sequences is restricted to the V3 loop and not found beyond its cysteine borders (Table 4).

A second observation based on the phenetic organization of sequences is that some members of the A and C subtypes show remarkably similar V3 loop protein sequences despite having highly divergent nucleotide sequences. This conservation is reflected in Fig. 3 through the very close relationships of some of the A and C subtype sequences. The V3 loop amino acid sequences of the subset of the A and C subtype sequences are highly similar, as can be seen in Fig. 4. The conservation appears to be restricted to the V3 loop itself and does not extend to its flanking region, as is illustrated in Fig. 5. The ratios of the rates of synonymous over nonsynonymous substitutions (*ds/dn*) are 25- to 60-fold higher for pairwise comparisons of specific A and C subtype 35-amino-acid V3 loop sequences than for the 34 amino acids flanking either side of the V3 loop; these values are extraordinary relative to pairwise comparisons of all other sequences from all other clades included in this study (Fig. 5). Intersubtype silent mutation rates in the V3 loop encoding sequences are high and typical for intercladal comparisons. For example, sequences A_ZR.6657 and C_MW.6508 and sequences A_UG2.117 and C_MW.12229, which are identical in V3 loop amino acid sequences, have *Ps* values of 47 and 34%, respectively, in the V3 loop. Moreover, intracladal sequence comparisons exploring the relationship of *ds/dn* values of the V3 loop relative to its flanking regions show that the V3 loop is relatively well conserved among C subtype sequences, suggesting that there is preservation of a form of the V3 loop within the C clade (Table 4). To explore the hypothesis of the retention of an ancestral A and C clade form of the V3 loop, a hypothetical ancestral sequence at the branch point of A and C subtype sequences was reconstructed by parsimony (74) analysis of 24 A and C subtype V3 region sequences. This hypothetical ancestral sequence translates into an amino acid sequence that corresponds closely to the similar form found in A and C subtypes, differing from CONSENSUS 4 in the alignment shown in Fig. 4 by a single isoleucine-to-valine substitution in the V3 loop: the amino acid sequence IRIGPGQ in CONSENSUS 4 is VRIGPGQ in the A and C ancestral sequence.

DISCUSSION

The phenetic analysis of the V3 loop protein sequences resulted in two interesting observations of possible indications

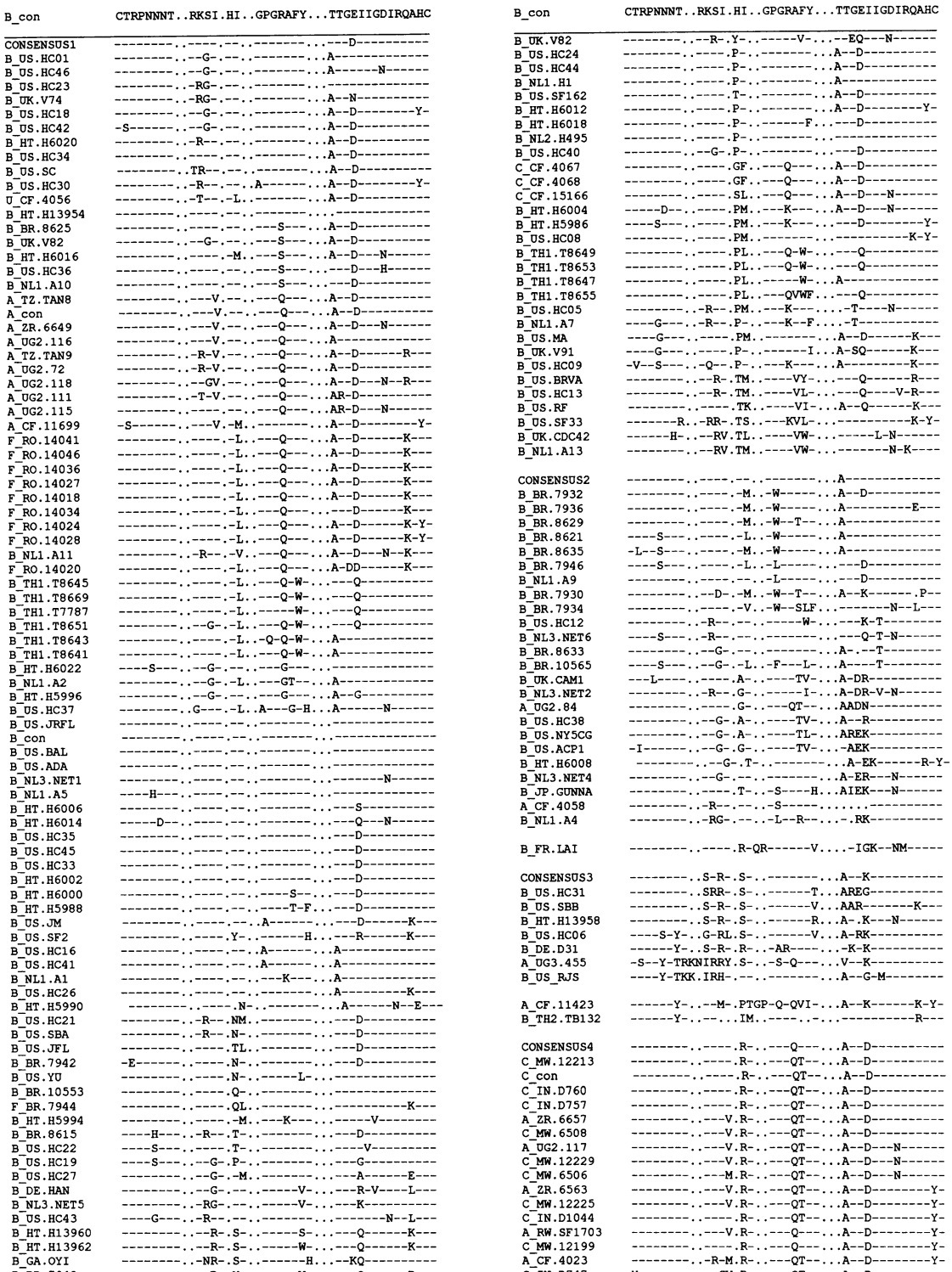


FIG. 4. Alignment of V3 loop protein sequences on the basis of the phenogram in Fig. 3. Sequences are labeled with the envelope genetic subtype association as the first letter of the name followed by an underscore, the two-letter country code designation, and the sequence identification. These sequence designations follow those from the V3 section of Human Retroviruses and AIDS database (51). Consensus sequences based on the most common amino acid in a given position were derived for each of the 14 clusters labeled in Fig. 3. Dashes represent identity with the top sequence in the alignment (B_con); dots represent insertions.

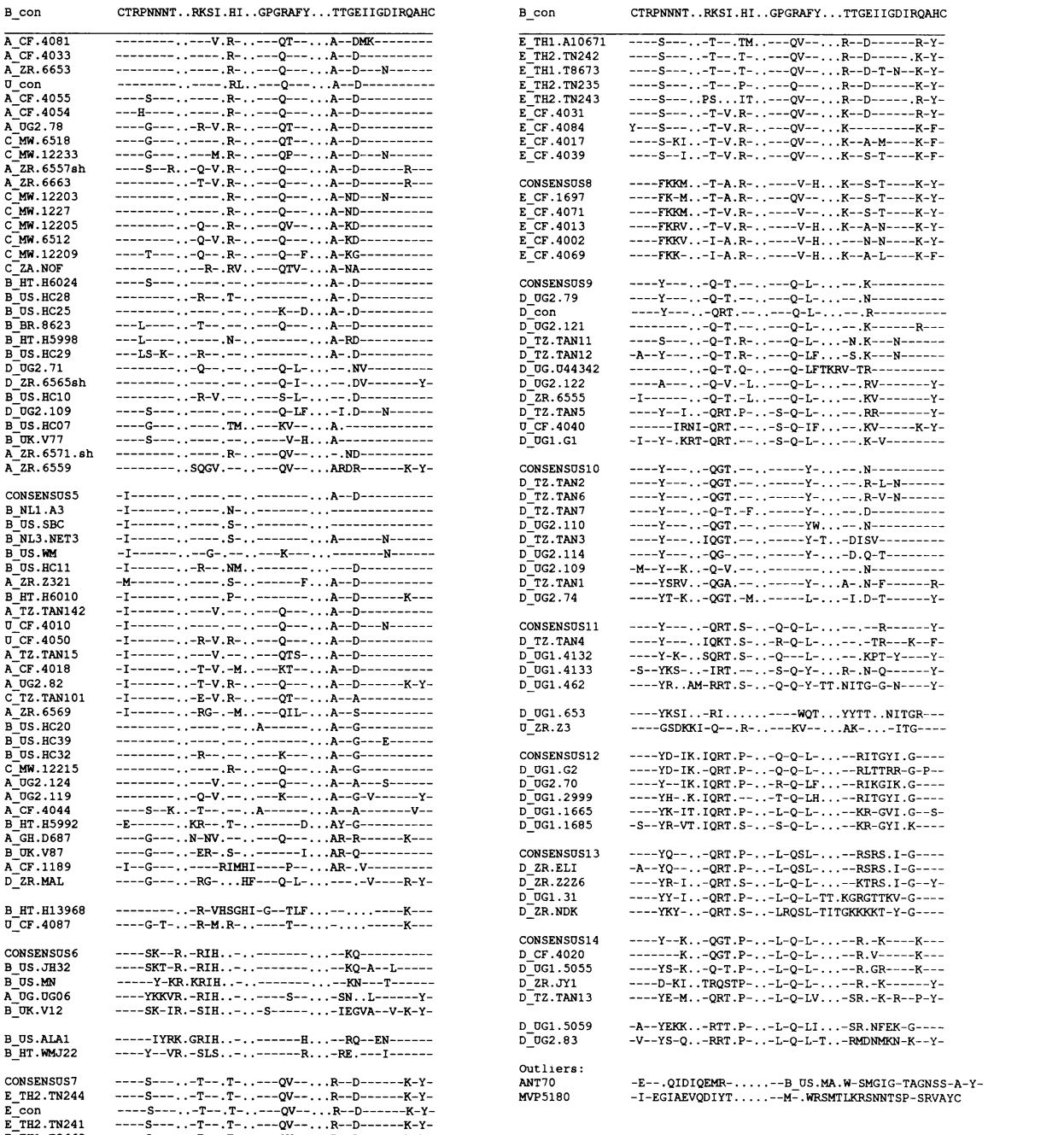


FIG. 4—*Continued.*

of distinct mutational patterns in the V3 loop in different HIV-1 lineages, with consequent changes in the biology of the virus. The first observation is that the interpatient set of D subtype V3 loop protein sequences are more divergent than other subtypes. The D subtype V3 loop protein sequences may actually be diverging more rapidly than the other subtypes; this suggestion is supported by the results of two different analytical methods. First, the intrasubtype nucleotide distances among D

TABLE 4. Intrasubtype comparisons of properties of V3 loop amino acid sequences and V3 region nucleotide similarities[a]

Subtype	N	Median (range) values for:						
		V3 region nuc *d* (IR)	V3 loop PIMA (IR)	V3 loop STR (IR)	V3 loop length (R)	pI (IR)	Charge (IR)	*ds/dn* V3 loop/flank ratio
A	41	15.1 (12.9–17.7)	148 (129–159)	173 (155–186)	35 (31–37)	6.0 (5.3–7.0)	0 (−0.5–0)	1.7/1.4 = 1.2
B	148	10.5 (8.1–12.2)	148 (139–157)	172 (160–184)	35 (31–36)	7.0 (6.1–9.1)	0 (0–1)	0.9/0.7 = 1.3
C	23	11.5 (8.5–17.5)	160 (154–165)	190 (183–197)	35 (35–35)	6.0 (6.0–8.1)	0 (0–0.5)	2.5/1.2 = 2.1
D	43	16.1 (13.3–18.8)	108 (93–125)	126 (110–146)	34 (30–38)	10.0 (8.8–11.1)	1.5 (1–2)	0.8/1.4 = 0.6
E	28	7.5 (2.0–12.5)	153 (131–168)	176 (154–199)	35 (32–35)	6.0 (6.0–9.7)	0 (0–2)	NA
F	10	2.4 (1.6–3.3)	172 (156–175)	202 (197–207)	35 (35–35)	5.0 (5.0–5.0)	−1 (−1–[−1])	NA

[a] Medians and interquartile ranges (IR) or ranges (R) are given. *N* refers to the number of sequences available from each subtype; all pairwise combinations of sequences were considered for nucleotide distances and PIMA and STR scores, and all sequences were considered for assessing the distribution of V3 loop lengths and pIs. Nucleotide distances (nuc *d*) are Hamming distances, or $d = (1 - s) \times 100\%$, where s is the fraction of shared bases in each position of an alignment. All positions containing gaps inserted to maintain the alignment were removed. This left approximately 200 positions in the V3 region of envelope for comparison. The E and F intrasubtype Hamming distances are small (and PIMA scores are high) because the E subtype sequences are predominately from Thailand and the F sequences are predominately from Romania, and both of these sets were taken early in the epidemic for the respective countries and show strong similarity, presumably due to a founder virus effect. The average distance between the Brazilian F subtype sequence and the Romanian F subtype sequences is 9.0%; Thai E subtype sequences and Central African Republic E subtype sequences differ by 8 to 12%. All pairwise V3 loop PIMA and STR scores were calculated for each subtype, and median and interquartile ranges are given. The higher the score, the greater is the amino acid conservation. The V3 loop length represents the number of amino acids from Cys to Cys. pIs were calculated by using MacVector (International Biotechnologies, Inc.), across the inner 15 amino acids of each loop, as discussed in Results. The net charge at pH 7 is also calculated for this same region. NA, not applicable. The *ds/dn* ratios are the median values for the V3 loop and for 34 amino acids flanking either side of the V3 loop (19 on the N-terminal side of the loop, and 15 on the C-terminal side). There are not enough changes in the E Thai and F Romanian sequences to include a *ds/dn* comparison, as many of the pairwise sequence comparisons do not show a single change over the short regions considered here.

subtype V3 region sequences are similar to those found among A subtype sequences (medians of 16% for D and 15% for A), while the protein similarity scores for the D subtype V3 loops are much lower (with medians of 108 for D and 148 for A) (Table 2). The second analysis is based on the ratio of synonymous divided by nonsynonymous substitutions (*ds/dn* ratios): the greater the *ds/dn* ratio, the greater is the relative conservation. In a comparison of the V3 loop to the flanking regions of the V3 loop, the D subtype viruses were found to have lower *ds/dn* ratios in the V3 loop. This suggests a greater rate of change at the protein level within the D subtype V3 loops than in the V3 loop flanking regions. All other subtypes

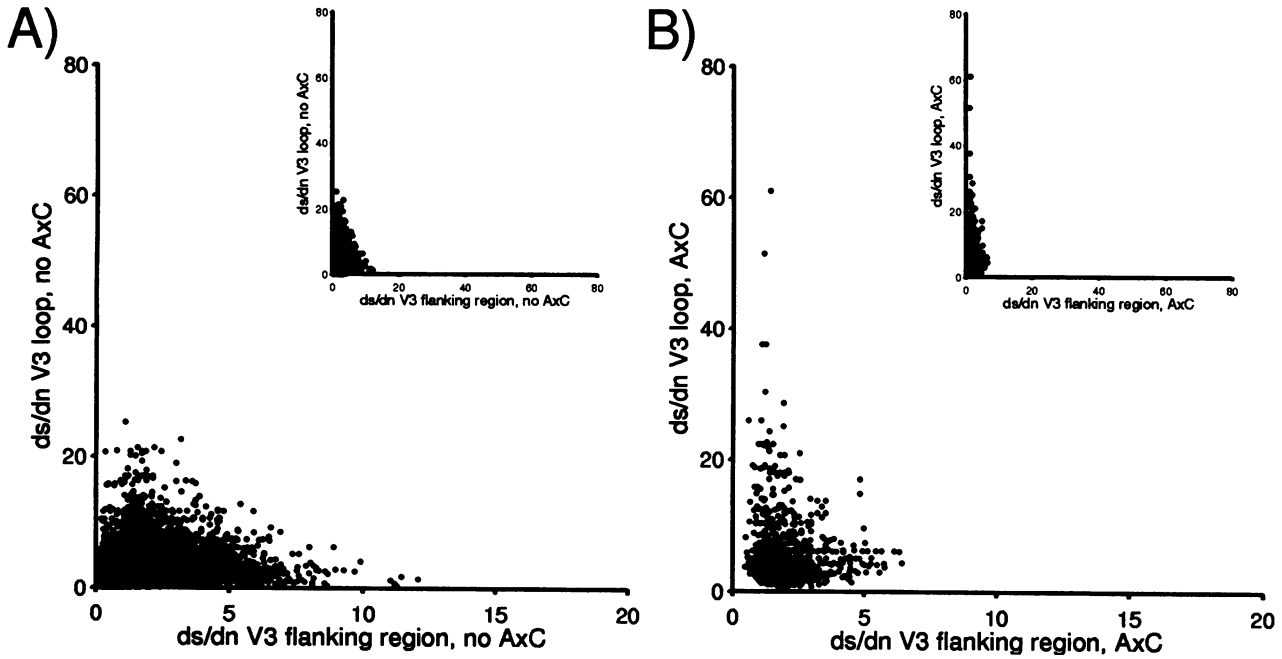


FIG. 5. Ratio of synonymous/nonsynonymous substitutions in the region encoding the 34 amino acids flanking the V3 loop contrasted with the region encoding the 35 amino acids of the V3 loop. Pairwise comparisons of all sequences included in this study were made, and *ds/dn* values were calculated for both regions. The sequence lengths under consideration here are short; hence, there is naturally some variation in the relative levels of conservation of the two regions. However, all of the points corresponding to sequences that are most extraordinarily conserved in the V3 region, without showing great conservation in the V3 flanking region, are clade A sequence comparisons contrasted with clade C sequence comparisons. Sequence comparisons for which there were no nonsynonymous substitutions are excluded to prevent denominators of zero. Part A is a plot of 28,516 pairwise comparisons of *ds/dn* values in the V3 loop (ordinate) and the V3 loop flanking regions (abscissa) of sequences, excluding clade A and C comparisons (no A×C). Part B is comparable to part A but represents only the clade A and C sequence comparisons (A×C). The inserts show the ordinate and abscissa with the same range (0 to 80) for direct comparison of the *ds/dn* values of the two regions.

had higher *ds/dn* ratios in the V3 loop, suggesting that for the other subtypes V3 loops were relatively more conserved than V3 loop flanking regions (Table 2). For example, A and D subtypes had comparable median *ds/dn* ratios in the V3 flanking regions (1.4), but within the V3 loop the D subtype had a much lower median value (0.8 for D versus 1.7 for A).

The D subtype viruses are also more positively charged. Given the association with positive charge in the V3 loop and SI phenotype in culture (22) and the correlation of the presence of SI viruses at seroconversion and rapid progression to disease (54), the possibility arises that the D subtype viral strains found in central Africa may represent a more pathogenic form of the virus. Another disturbing consequence of the greater variability in the V3 loop in the clade D viruses is that they may be engendering a more elusive vaccine target than are the viruses of other genetic lineages.

One hypothesis that might account for wider divergence among D subtype V3 loop sequences is that it is simply a sampling artifact and our observation is a consequence of differences in the stage of disease progression in infected people from which the samples were originally derived. For example, one might suspect that D subtype sequences were more likely to have come from people with full-blown AIDS and other subtype sequences to have come from healthy HIV-1-seropositive individuals. While this is a possible explanation, we consider it unlikely: many of the A and D samples came from the same studies conducted in central Africa, and therefore have been exposed to the same sampling biases, yet the V3 protein sequence characteristics of the two subtypes are very distinctive. Additionally, the World Health Organization (WHO) Global Programme on AIDS has collected samples from asymptomatic individuals from Rwanda, Uganda, Thailand, and Brazil; and the observations made here concerning charge and V3 variability have been substantiated in that data set, which includes D subtype sequences from 12 asymptomatic Ugandans, compared with the total set of 56 WHO HIV-1 isolates from asymptomatic people with virus associated with other subtypes by Myers and Korber and collaborators in the WHO technical working group on HIV isolation and characterization (50a).

The second observation based on the comparison of phenetic and phylogenetic clustering patterns pertains to a marked similarity of V3 loop sequences from a subset of A and C subtype sequences. There are three possible explanations of the similarity of V3 loop sequences in the A and C subtypes: a relatively recent recombination event between A and C subtype sequences, convergence, or a lack of divergence from a common precursor. Recombination is unlikely, because silent mutations permeate the A and C sequence comparisons over the V3 loop and are typical of what one would expect to find for intersubtype comparisons. Convergence at the molecular level is probably rare, except over short functional domains (71). It is extremely difficult to trace the convergent evolution of particular amino acid patterns (however, this has been claimed in comparisons of the lysozyme enzyme from the foregut of ruminants [72, 73]). There are several studies of HIV-1 sequences that propose convergence within the V3 loop. Holmes et al. have identified mutations within the V3 loop that they argue are a result of convergence due to mutational events in viruses that are associated with different phylogenetic branches found within a single patient (35). Albert et al. (1) found a higher degree of similarity between 1990 Ugandan HIV-1 V3 loops and North American and European isolates than in comparisons of the 1990 Ugandan sequences with earlier Ugandan isolates, which they considered to be possible evidence of convergence. The Ugandan sequences they were comparing were from A and D clades, and the North American and European sequences were from the B clade.

While the V3 loop is an example of a limited region where convergence may occur, we think that the third possible explanation listed above is a more likely interpretation of the similarity of the A and C subtype sequences: simply a lack of divergence of a favorable form which was a common precursor of both clades. Strong selection upon transmission for particular V3 regions has been proposed to be a contributing factor to the observed homogeneity of viral sequences obtained from recent seroconverters (83, 85). It has been further proposed that such selection may be dictated by conservation of V3 sequences which are associated with macrophage tropism (12, 83). These studies of macrophage tropism and transmission have concerned only viruses of the B clade. Similar selection for natural viral transmissibility may be the basis for the preservation of the form observed in many A and C subtype viruses. By close examination of clustered V3 loop sequences (Fig. 3 and 4), one can find a scattered set of sequences from a range of subtypes which are relatively close to the similar A and C forms; however, the most extraordinary V3 amino acid conservation in phylogenetically distant clades is observed between the A and C clades (Fig. 3 and 5). It is interesting that the V3 loop sequences of a subset of phylogenetically distinct HIV-2 sequences are relatively similar when compared by phenetic analysis (50) and that some simian immunodeficiency virus V3 loop sequences are highly similar (68). The similarity in these viral V3 loops may also reflect preservation of a biologically distinct form of the virus. It is also important to note that while the similarity of the A and C forms of the V3 loop (as well as the similarity of the two Gag prôtein sequences from HIVSF2 and HIVCAM1) is striking, it is also unusual and the general trend among the international set of HIV-1 protein sequences is divergence.

Phenetic cluster analysis may be particularly useful for designing an appropriate set of peptides for inclusion in a linear peptide vaccine cocktail. For example, clusters in Fig. 3 are numbered and the consensus sequences for each of the 14 clusters are included in the alignment in Fig. 4. These consensus sequences are derived from sequences associated on the basis of protein similarity. As such, these would better represent the range of possible forms of the V3 loop for inclusion in a vaccine than a set of peptides based on phylogenetically defined subtype consensus sequences. Phenetic associations could be used as a qualitative guide in conjunction with other available data, such as geographic distribution of variants (1, 45, 49, 51, 57, 59, 63, 64), antigenic cross-reactivity (6, 7, 10), structural information (27), and coordinate mutations found between sites in this region (36). The best utilization of phenetic analysis for vaccine design would be to use variants defined within the narrow boundaries of the highly immunogenic tip of the V3 loop, rather than the intact loop, as shown here. When peptides from other regions of HIV-1 proteins are employed for either vaccine design or peptide enzyme-linked immunosorbent assay, phenetic organization of sequences representing these protein fragments could provide additional insight into appropriate peptide cocktails.

Although phenetic organization of short peptide fragments is potentially useful for peptide-based vaccine design, most vaccine strategies currently being considered (9) employ intact proteins, either as subunit vaccines or through recombinant viral vectors or as live attenuated or whole killed virus. These strategies could elicit an immune response to discontinuous B-cell epitopes and a range of T-cell epitopes. For these strategies, representatives from the phylogenetically defined

HIV-1 clades that are present in target populations could serve as a guide for selecting a substrate for vaccine design, since the phenetic and phylogenetic clustering patterns are highly similar when longer stretches of sequence information are employed.

Several scenarios that account for the phylogenetic clustering of distinct clades could be imagined. The equidistance of the branch lengths between these major groups suggests that the HIV-1 clades may reflect a star phylogeny, where an instantaneous burst of new lineages arose from a single progenitor. The scenario that we think is the most plausible interpretation is that HIV-1 has been diverging within the human host and that genetic subtypes identified to date (A through F in *env*) reflect the spread of the virus into different human populations. The theory that a common ancestral sequence of the distinct genetic lineages of HIV-1 A through F may have existed contemporaneously with the beginning of the pandemic is consistent with available information concerning the macroscopic rates of evolution of lentiviruses (50). The equidistance of the branch lengths between clades is suggestive evidence that cladal founder viruses may have spread into different human subpopulations at a specific historical moment. This moment may have been dictated by changes in the behavior of the human host which allowed the virus to spread, such as increased density of urban populations or increased international travel. In this context, it is interesting that internal genetic distances between the available HIV-1 outlier sequences (28, 78) are comparable to those observed among sequences in the clades A through F. An alternative scenario that we regard as less likely because of the conservation of the genetic distances between the clades is that distinct zoonotic transmission events from simian precursers led to the introduction of each of the distinct HIV-1 clades in humans.

The scenario described above, that of the HIV-1 clades A through F diverging in the human host, can be envisioned hypothetically through the example of the E subtype viruses found in Thailand and the Central African Republic (45, 49, 59). The extreme genetic similarity of E subtype sequences found in Thailand testifies to the recent introduction of an E subtype virus that has spread rapidly through the Thai population. We think it is likely that the level of HIV-1 variation in Thailand will increase with the duration of the epidemic until, within a decade or two, it achieves the level of variation found in recently isolated U.S. or European B subtype viruses (37, 58) and comparable to that observed among E subtype viruses found in the Central African Republic (49). If a different E subtype strain had served as a founder virus in a distinct population in the late 1980s, coincident with the beginning of the epidemic in Thailand, two independent E sublineages with founder viruses approximately 12% distant in *env* would have emerged. This eventually could have resulted in intercladal distances between the two hypothetical E sublineages that were comparable to those observed between subtypes A through F. The plausibility of such a scenario happening within decades is put into perspective through consideration of the extraordinary level of intrapatient variation that develops within the course of a single infection (3, 35, 80, 84) from the relatively homogeneous viral forms found within an individual upon primary infection (79, 83, 85).

The differences in the phenetic clustering patterns of the V3 loop are most interesting when considered in the context of the phylogenetic relationships of the sequences. The equidistance of the branch lengths of the major clades to an ancestral node suggests that the viruses are diverging from an ancestral source, quite possibly at the same rate when averaged over long sequences of specific genes. However, the phenetic relationships in the V3 loop identify different mutational patterns in the different clades within the narrow boundaries of this important domain. The combined phylogenetic and phenetic analyses in this paper suggest that within the different lineages, the biologically important V3 loop may have adopted distinctive mutational characteristics. This analysis is consistent with the hypothesis that biologically distinct variants are emerging.

ACKNOWLEDGMENTS

We thank James Theiler for writing the program that enabled us to draw the clustering patterns obtained from PIMA and Steven Henikoff for helpful suggestions concerning protein similarity matrices. We also thank Ethan Allen and Miranda McEvilly for assistance maintaining the C2V3 region sequence database.

This work was supported by NIH/NIAID-DOE interagency agreement 3-Y01-A1-70001-11 through the NIAID DAIDS variation program (B.T.M.K., K.M., and G.M.), the ARIEL project of the Pediatric AIDS Foundation (B.T.M.K.), the W. M. Keck Center for Computational Biology, NIH grant 1R01-HG00973-01 (R.F.S.), and Baylor College of Medicine Human Genome Center grant P30-HG00210 (R.F.S.).

REFERENCES

1. **Albert, J., L. Franzen, M. Jansson, G. Scarlatti, P. K. Kataaha, E. Katabira, F. Mubiro, M. Rydaker, P. Rossi, U. Pettersson, and H. Wigzell.** 1992. Ugandan HIV-1 V3 loop sequences closely related to the US/European consensus. Virology **190**:674–681.
2. **Andeweg, A. C., P. Leeflang, A. D. M. E. Osterhaus, and M. L. Bosch.** 1993. Both the V2 and V3 regions of the human immunodeficiency virus type 1 surface glycoprotein functionally interact with other envelope regions in syncytium formation. J. Virol. **67**:3232–3239.
3. **Balfe, P., P. Simmonds, C. A. Ludlam, J. O. Bishop, and A. J. Leigh Brown.** 1990. 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. J. Virol. **64**:6221–6233.
4. **Berger, E. A., J. R. Sisler, and P. L. Earl.** 1992. Human immunodeficiency virus type 1 envelope glycoprotein molecules containing membrane fusion-impairing mutations in the V3 region efficiently undergo soluable CD4-stimulated gp120 release. J. Virol. **66**: 6208–6212.
5. **Bergeron, L., N. Sullivan, and J. Sodroski.** 1992. Target cell-specific determinants of membrane fusion within the human immunodeficiency virus type 1 gp120 third variable region and gp41 amino terminus. J. Virol. **66**:2389–2397.
6. **Berman, P. W., T. J. Matthews, L. Riddle, M. Champe, M. R. Hobbs, G. R. Nakamura, J. Mercer, D. J. Eastman, C. Lucas, A. J. Langlois, F. M. Wurm, and T. J. Gregory.** 1992. Neutralization of multiple laboratory and clinical isolates of human immunodeficiency virus type 1 (HIV-1) by antisera raised against gp120 from the MN isolate of HIV-1. J. Virol. **66**:4464–4469.
7. **Blomberg, J., A. Lawoko, R. Pipkorn, S. Moyo, B. E. Malmvall, J. Shao, R. Dash, and S. Tswana.** 1993. A survey of HIV-1 peptides with natural and chimeric sequences for differential reactivity with Zimbabwean and Tanzanian and Swedish HIV-1-positive sera. AIDS **7**:759–767.
8. **Boyd, M. T., G. R. Simpson, A. J. Cann, M. A. Johnson, and R. A. Weiss.** 1993. A single amino acid substitution in the V1 loop of human immunodeficiency virus type 1 gp120 alters cellular tropism. J. Virol. **67**:3649–3652.
9. **Cease, K. B., and J. A. Berzofsky.** Towards a vaccine for AIDS: the emergence of immunobiology-based vaccine development. Annu. Rev. Immunol., in press.
10. **Cheingsong-Popov, R., D. Callow, S. Beddows, S. Shaunak, C. Wasi, P. Kaleebu, C. Gilks, I. V. Petrascu, M. M. Garaev, D. M. Watts, N. T. Constantine, and J. N. Weber.** 1992. Geographical diversity of human immunodeficiency virus type 1: serologic reactivity to env epitopes and to neutralization. J. Infect. Dis. **165**:256–261.
11. **Cheng-Mayer, C., D. Seto, M. Tateno, and J. A. Levy.** 1988.

Biologic features of HIV-1 that correlate with virulence in the host. Science **240**:80–82.
12. **Chesebro, B., K. Wehrly, J. Nishio, and S. Perryman.** 1992. Macrophage-tropic human immunodeficiency virus isolates from different patients exhibit unusual V3 envelope sequence homogeneity in comparison with T-cell-tropic isolates: definition of critical amino acids involved in cell tropism. J. Virol. **66**:6547–6554.
13. **de Jong, J.-J., A. de Ronde, W. Keulen, M. Tersmette, and J. Goudsmit.** 1992. Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution. J. Virol. **66**:6777–6780.
14. **de Jong, J.-J., J. Goudsmit, W. Keulen, B. Klaver, W. J. A. Krone, M. Tersmette, and A. de Ronde.** 1992. Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. J. Virol. **66**:757–765.
15. **De Leys, R., B. Vanderborght, M. Vanden Haesevelde, L. Heyndrickx, A. van Geel, C. Wauters, R. Bernaerts, E. Saman, P. Nijs, B. Willems, H. Taelman, G. van der Groen, P. Piot, T. Tersmette, J. G. Huisman, and H. Van Heuverswyn.** 1990. Isolation and partial characterization of an unusual human immunodeficiency retrovirus from two persons of west-central African origin. J. Virol. **64**:1207–1216.
16. **Eigen, M., and K. Neiselt-Struwe.** 1990. How old is the immunodeficiency virus? AIDS **4**(Suppl.):S85–S93.
17. **Ewald, P. W.** 1991. Transmission modes and the evolution of virulence. Hum. Nat. **2**:1–30.
18. **Ewald, P. W.** 1993. The evolution of virulence. Sci. Am. **1993**:56–62.
19. **Fauikner, D. V., and A. Jurka.** 1988. Multiple aligned sequence editor (MASE). Trends Biochem. Sci. **13**:321–322.
20. **Felsenstein, J.** 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.
21. **Felsenstein, J.** 1993. PHYLIP (phylogeny inference package) version 3.4. Department of Genetics, University of Washington, Seattle.
22. **Fouchier, R. A. M., M. Groenink, N. A. Kootstra, M. Tersmette, H. G. Huisman, F. Miedema, and H. Schuitemaker.** 1992. Phenotype-associated sequence variation in the third variable region of the human immunodeficiency virus type 1 gp120 molecule. J. Virol. **66**:3183–3187.
23. **Freed, E. O., D. J. Myers, and R. Risser.** 1991. Identification of the principal neutralizing determinant of human immunodeficiency virus type 1 as a fusion domain. J. Virol. **65**:190–194.
24. **Freed, E. O., and R. Risser.** 1991. Identification of conserved residues in the human immunodeficiency virus type 1 principal neutralizing determinant that are involved in fusion. AIDS Res. Hum. Retroviruses **7**:807–811.
25. **Goudsmit, J., C. Debouck, R. H. Meloen, L. Smit, M. Bakker, D. M. Asher, A. V. Wolff, C. J. Gibbs, Jr., and C. Gajdusek.** 1988. Human immunodeficiency virus type 1 neutralization epitope with conserved architecture elicits early type-specific antibodies in experimentally infected chimpanzees. Proc. Natl. Acad. Sci. USA **85**:4478–4482.
26. **Groenink, M., R. A. M. Fouchier, S. Broersen, C. H. Baker, M. Koot, A. B. van't Wout, H. G. Huisman, F. Miedema, M. Tersmette, and H. Schuitemaker.** 1993. Relation of phenotype evolution of HIV-1 to envelope V2 configuration. Science **260**:1513–1516.
27. **Gupta, G., G. M. Anantharamaiah, D. R. Scott, J. H. Eldridge, and G. Myers.** 1993. Solution structure of the V3 loop of a Thailand HIV isolate. J. Biomol. Struct. Dyn. **11**:345–366.
28. **Gürtler, L. G., P. H. Hauser, J. Eberle, A. von Brunn, S. Knapp, L. Zekeng, J. M. Tsague, and L. Kaptue.** 1994. A new subtype of human immunodeficiency virus type 1 (MVP-5180) from Cameroon. J. Virol. **68**:1581–1585.
29. **Henikoff, S., and J. G. Henikoff.** 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89**: 10915–10919.
30. **Henikoff, S., and J. G. Henikoff.** 1993. Performance evaluation of amino acid substitution matrices. Proteins Struct. Funct. Genet. **17**:49–61.
31. **Hillis, D. M., M. W. Allard, and M. M. Miyamoto.** 1993. Analysis of DNA sequence data: phylogenetic inference. Methods Enzymol. **224**:456–487.
32. **Hillis, D. M., and J. J. Bull.** 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. **42**:182–189.
33. **Hillis, D. M., and J. P. Huelsenbeck.** 1992. Signal, noise and reliability in molecular phylogenetic analysis. J. Hered. **83**:189–195.
34. **Hillis, D. M., J. P. Huelsenbeck, and C. W. Cunningham.** Application and accuracy of molecular phylogenies. Submitted for publication.
35. **Holmes, E. C., L. Q. Zang, P. Simmonds, C. A. Ludlam, and A. J. Leigh Brown.** 1992. Convergent and divergent sequence evolution in the surface envelope glycoprotein of the human immunodeficiency virus type 1 within a single patient. Proc. Natl. Acad. Sci. USA **89**:4835–4839.
36. **Korber, B. T. M., R. M. Farber, D. H. Wolpert, and A. S. Lapedes.** 1993. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc. Natl. Acad. Sci. USA **90**:7176–7180.
37. **Kuiken, C. L., G. Zwart, E. Baan, R. A. Coutinho, J. A. R. van den Hoek, and J. Goudsmit.** 1993. Increasing antigenic and genetic diversity of the HIV-1 V3 domain in the course of the AIDS epidemic. Proc. Natl. Acad. Sci. USA **90**:9061–9065.
38. **Learmont, J., B. Tindall, L. Evans, A. Cunningham, P. Cunningham, J. Wells, R. Penny, J. Kaldor, and D. A. Cooper.** 1992. Long-term symptomless HIV-1 infection in recipients of blood products from a single donor. Lancet **340**:863–867.
39. **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.
40. **Li, W.-H., and D. Graur.** 1991. Fundamentals of molecular evolution, p. 113–144. Sinauer Associates, Inc., Sunderland, Mass.
41. **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.
42. **Louwagie, J., F. E. McCutchan, M. Peeters, T. P. Brennan, E. Buell, G. A. Eddy, G. van der Groen, K. Fransen, G. Gershy-Damet, R. Deleys, and D. S. Burke.** 1993. Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. AIDS **7**:769–780.
43. **Maddison, W. P., and D. R. Maddison.** 1992. MacClade: analysis of phylogenetic and character evolution. Sinauer Associates, Inc., Sunderland, Mass.
44. **Mastro, T. D., G. A. Satten, T. Nopkesorn, S. Sangkharomya, and I. M. Longini, Jr.** 1994. Probability of female-to-male transmission of HIV-1 in Thailand. Lancet **343**:204–207.
45. **McCutchan, F. E., P. A. Hegerich, T. P. Brennan, P. Phanuphak, P. Singharaj, A. Jugsudee, P. W. Berman, A. M. Gary, A. K. Fowler, and D. S. Burke.** 1992. Genetic variants of HIV-1 in Thailand. AIDS Res. Hum. Retroviruses **8**:1887–1895.
46. **McKeating, J. A., J. Cordell, C. J. Dean, and P. Balfe.** 1992. Synergistic interactions between ligands binding to the CD4 binding site and V3 domain of human immunodeficiency virus type 1 gp120. Virology **191**:732–742.
47. **Milich, L., B. Margolin, and R. Swanstrom.** 1993. V3 loop of the human immunodeficiency virus type 1 Env protein: interpreting sequence variability. J. Virol. **67**:5623–5634.
48. **Moore, J. P., M. Thali, B. A. Jameson, F. Vignaux, G. K. Lewis, S. Poon, M. Charles, M. S. Fung, B. Sun, P. J. Durda, L. Akerblom, B. Wahren, D. D. Ho, Q. J. Sattentau, and J. Sodroski.** 1993. Immunochemical analysis of the gp120 surface glycoprotein of human immunodeficiency virus type 1: probing the structure of the C4 and V4 domains and the interaction of the C4 domain with the V3 loop. J. Virol. **67**:4785–4796.
49. **Murphy, E., B. Korber, M. Georges-Courbot, B. You, A. Pinter, D. Cook, M. Kieny, A. Georges, C. Mathiot, F. Barre-Sinoussi, and M. Girard.** 1993. Diversity of V3 region sequences of human immunodeficiency virus type 1 from the Central African Republic. AIDS Res. Hum. Retroviruses **9**:997–1006.
50. **Myers, G., and B. Korber.** 1994. The future of HIV, p. 211–232. *In* S. S. Morse (ed.), Evolutionary biology of viruses. Raven Press, New York.

50a. **Myers, G., and B. Korber, et al.** Unpublished data.
51. **Myers, G., B. Korber, S. Wain-Hobson, R. F. Smith, and G. N. Pavlakis.** 1993. Human retroviruses and AIDS 1993. Los Alamos National Laboratory: Theoretical Biology and Biophysics, Los Alamos, N.M.
52. **Myers, G., and G. N. Pavlakis.** 1992. Evolutionary potential of complex retroviruses, p. 51–104. *In* J. A. Levy (ed.), The retroviruses. Plenum Press, New York.
53. **Nei, M., and T. Gojobori.** 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.
54. **Nielsen, C., C. Pedersen, J. D. Lundgren, and J. Gerstoft.** 1993. Biological properties of HIV isolates in primary HIV infection: consequences for subsequent course of infection. AIDS **7**:1035–1040.
55. **Nkengasong, J. N., M. Peeters, M. Vanden Haesevelde, S. S. Musi, B. Willems, P. M. Ndumbe, E. Delaporte, J.-L. Perret, P. Piot, and G. van der Groen.** 1993. Antigenic evidence of the presence of the aberrant HIV-1 ANT70 virus in Cameroon and Gabon. AIDS **7**:1536–1537.
56. **Olson, G. J., H. Matsuda, R. Hagstrom, and R. Overbeek.** 1994. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. **10**:41–48.
57. **Oram, J. D., R. G. Downing, M. Roff, N. Serwankambo, J. C. S. Clegg, A. S. R. Featherstone, and J. C. Booth.** 1991. Sequence analysis of the V3 loop of the env genes of Ugandan human immunodeficiency proviruses. AIDS Res. Hum. Retroviruses **7**:605–614.
58. **Ou, C., C. A. Clesieisky, G. Myers, C. I. Bandes, C. C. Luo, B. T. M. Korber, J. I. Mullins, G. Schochetman, R. L. Berkeiman, A. N. Economou, J. J. Witte, L. J. Furman, G. A. Satten, K. A. MacInnes, J. W. Curran, H. W. Jaffe, Laboratory Investigation Group, and Epidemiological Investigation Group.** 1992. Molecular epidemiology of HIV transmission in a dental practice. Science **256**:1165–1171.
59. **Ou, C., Y. Takebe, C. Luo, M. Kalish, W. Auwanit, C. Bandea, N. de la Torre, J. L. Moore, G. Schochetman, S. Yamazaki, H. D. Gayle, N. L. Young, and B. G. Weniger.** 1992. Wide distribution of two subtypes of HIV-1 in Thailand. AIDS Res. Hum. Retroviruses **8**:1471–1472.
60. **Overington, J., D. Donnelly, M. S. Johnson, A. Sali, and T. L. Blundell.** 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci. **1**:216–226.
61. **Palker, T. J., M. E. Clark, A. J. Langlois, T. J. Matthews, K. J. Weinhold, R. R. Randall, D. P. Bolognesi, and B. F. Haynes.** 1988. Type-specific neutralization of the human immunodeficiency virus with antibodies to env-encoded synthetic peptides. Proc. Natl. Acad. Sci. USA **85**:1932–1936.
62. **Pinter, A., W. J. Honnen, and S. A. Tilley.** 1993. Conformational changes affecting the V3 and CD4-binding domains of human immunodeficiency virus type 1 gp120 associated with *env* processing and with binding of ligands to these sites. J. Virol. **67**:5692–5697.
63. **Potts, K. E., M. L. Kalish, C. I. Bandea, G. M. Orloff, M. St. Louis, C. Brown, C. Malanda, M. Kavuka, G. Schochetman, C. Ou, and W. L. Heyward.** 1993. Genetic diversity of human immunodeficiency virus type 1 strains in Kinshasa, Zaire. AIDS Res. Hum. Retroviruses **9**:613–618.
64. **Potts, K. E., M. L. Kalish, T. Lott, G. Orloff, C. C. Luo, M. Bernard, C. Alves, R. Badaro, J. Suleiman, J. Ferreira, G. Schochetman, W. D. Johnson, C. Ou, J. L. Ho, and the Brazilian collaborative AIDS research group.** 1993. Genetic heterogeneity of the HIV-1 envelope glycoprotein in Brazil. AIDS **7**:1191–1197.
65. **Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling.** 1986. Numerical recipes, the art of scientific computing. Cambridge University Press, Cambridge.
66. **Rusche, J. R., K. Javaherian, C. McDanal, J. Petro, D. L. Lynn, R. Grimaila, A. J. 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.
67. **Shioda, T., J. A. Levy, and C. Cheng-Mayer.** 1992. Small amino acid changes in the V3 hypervariable region of gp120 can affect the T-cell-line and macrophage tropism of human immunodeficiency virus type 1. Proc. Natl. Acad. Sci. USA **89**:9434–9438.
68. **Shpaer, E. G., and J. I. Mullins.** 1993. Rates of amino acid change in the envelope protein correlate with pathogenicity of primate lentiviruses. J. Mol. Evol. **37**:57–65.
69. **Smith, R. F., and T. F. Smith.** 1990. Automatic generation of primary sequence patterns from sets of related protein sequences. Proc. Natl. Acad. Sci. USA **87**:118–122.
70. **Smith, R. F., and T. F. Smith.** 1992. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modeling. Protein Eng. **5**:35–41.
71. **States, D. J., and M. S. Boguski.** 1992. Similarity and homology, p. 89–157. *In* M. Gribskov and J. Devereux (ed.), Sequence analysis primer. W. H. Freeman and Company, New York.
72. **Stewart, C., J. W. Schilling, and A. C. Wilson.** 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature (London) **330**:401–404.
73. **Swanson, K. W., D. M. Irwin, and A. C. Wilson.** 1991. Stomach lysozyme gene of the langur monkey: tests for convergence and positive selection. J. Mol. Evol. **33**:418–425.
74. **Swofford, D.** 1993. PAUP (phylogenetic analysis using parsimony) version 3.1. Center for Biodiversity, Illinois Natural History Survey, Champaign, Ill.
75. **Swofford, D. L., and G. J. Olson.** 1991. Phylogeny reconstruction, p. 411–566. *In* D. M. Hillis and C. Moritz (ed.), Molecular systematics. Sinauer Associates, Inc., Sunderland, Mass.
76. **Temin, H. W.** 1989. Is HIV unique or merely different? J. Acquired Immune Defic. Syndr. **2**:1–9.
77. **Tersmette, M., and H. Schuitemaker.** 1993. Virulent HIV strains? AIDS **7**:1123–1125.
78. **Vanden Haesevelde, M., J. Decourt, R. J. De Leys, B. Vanderborght, G. van der Groen, H. van Heuverswijn, and E. Saman.** 1994. Genomic cloning and complete sequence analysis of a highly divergent African human immunodeficiency virus isolate. J. Virol. **68**:1586–1596.
79. **Wolfs, T. F. W., G. Zwart, M. Bakker, and J. Goudsmit.** 1992. HIV-1 genomic RNA diversification following sexual and parenteral virus transmission. Virology **189**:103–110.
80. **Wolinsky, S. M., C. M. Wike, B. Korber, C. Hutto, W. P. Parks, L. L. Rosenblum, K. J. Kunstman, M. R. Furtado, and J. Munoz.** 1992. Selective transmission of human immunodeficiency virus type 1 variants from mothers to infants. Science **255**:1134–1137.
81. **Wyatt, R., M. Thali, S. Tilley, A. Pinter, M. Posner, D. Ho, J. Robinson, and J. Sodroski.** 1992. Relationship of the human immunodeficiency virus type 1 gp120 third variable loop to a component of the CD4 binding site in the fourth conserved region. J. Virol. **66**:6997–7004.
82. **Yokoyama, S., L. Chung, and T. Gojobori.** 1988. Molecular evolution of the human immunodeficiency and related viruses. Mol. Biol. Evol. **5**:237–251.
83. **Zhang, L. Q., P. MacKenzie, A. Cleland, E. C. Holmes, A. J. Leigh Brown, and P. Simmonds.** 1993. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J. Virol. **67**:3345–3356.
84. **Zhang, Y.-M., S. C. Dawson, D. Landsman, H. C. Lane, and N. P. Salzman.** 1994. Persistence of four related human immunodeficiency virus subtypes during the course of zidovudine therapy: relationship between virion RNA and proviral DNA. J. Virol. **68**:425–432.
85. **Zhu, T., H. Mo, N. Wang, D. S. Nam, Y. Cao, R. A. Koup, and D. D. Ho.** 1993. Genotypic and phenotypic characterization of HIV-1 in patients with primary infection. Science **261**:1179–1181.