Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 9282–9286, October 1991
Evolution

Importance of purine and pyrimidine content of local nucleotide sequences (six bases long) for evolution of the human immunodeficiency virus type 1

(nonrandom mutation/error spectra/difference in codon frame/positive- and negative-selection pressures/evolutionary strategy)

HIROFUMI DOI

Biological Informatics Section, International Institute for Advanced Study of Social Information Science, Fujitsu Laboratories Ltd., 17-25 Shinkamata 1-chome, Ohta-ku, Tokyo 144, Japan

Communicated by Susumo Ohno, July 9, 1991

ABSTRACT **Human immunodeficiency virus type 1 evolves rapidly, and random base change is thought to act as a major factor in this evolution. However, segments of the viral genome differ in their variability: there is the highly variable *env* gene, particularly hypervariable regions located within *env*, and, in contrast, the conservative *gag* and *pol* genes. Computer analysis of the nucleotide sequences of human immunodeficiency virus type 1 isolates reveals that base substitution in this virus is nonrandom and affected by local nucleotide sequences. Certain local sequences 6 base pairs long are excessively frequent in the hypervariable regions. These sequences exhibit base-substitution hotspots at specific positions in their 6 bases. The hotspots tend to be nonsilent letters of codons in the hypervariable regions—thus leading to marked amino acid substitutions there. Conversely, in the conservative *gag* and *pol* genes the hotspots tend to be silent letters because of a difference in codon frame from the hypervariable regions. Furthermore, base substitutions in the local sequences that frequently appear in the conservative genes occurred at a low level, even within the variable *env*. Thus, despite the high variability of this virus, the conservative genes and their products could be conserved. These may be some of the strategies evolved in human immunodeficiency virus type 1 to allow for positive-selection pressures, such as the host immune system, and negative-selection pressures on the conservative gene products.**

The human immunodeficiency virus type 1 (HIV-1) has been shown to be the causative agent in AIDS (1, 2). Extensive genotypic variation—in particular, the presence of hypervariable regions in the extracellular envelope glycoprotein gp120—has been firmly established as a prominent feature of this virus (3–8). Sequence changes in the serial virus isolates taken from a single patient (WMJ) were almost exclusively nucleotide base substitutions (7). Such substitution, caused by HIV-1 reverse transcriptase (RT), has hence been assumed to be the primary mechanism for genomic change in this virus (7), although the rate-limiting step of mutations in HIV-1 variation has not been established. Another finding was that the *env* gene of this virus, encoding gp120 and the transmembrane envelope glycoprotein gp41, was more highly variable than the *gag* and *pol* genes, which encode the core proteins and the RT and endonuclease (3–8). The difference in variability between the *env* gene and the conservative genes *gag* and *pol* is generally thought to result from natural selection after random base changes.

The hypervariable regions, however, aid HIV-1 in evading positive-selection pressures, such as the host immune system (3, 4), whereas the conservative genes act under negative-selection pressures because mutations in those genes put this virus under great reproductive disadvantages. The effect of these selection pressures, hence, would be bolstered were the initial mutations in these genes nonrandom. In fact, despite the general idea of random base changes, previous data suggest that base substitutions are not random (4, 7). Furthermore, nonrandom mutations affected by local nucleotide sequences when HIV-1 RT acts as DNA-dependent DNA polymerase *in vitro* have been seen by Roberts *et al.* (9) and Bebenek *et al.* (10), with an assay system for the fidelity of DNA synthesis *in vitro*. These results suggest that HIV-1 RT could have its own specific local sequence *in vivo* that leads it to frequently cause single-base substitutions. If that were the operable cause, the specific sequence could be expected to appear at a higher frequency in the hypervariable regions and, therefore, could characterize the hypervariable regions.

To test this hypothesis, local sequences in excess within the hypervariable regions and those in the other genes were identified by computer analysis of nucleotide sequences of the HIV-1 genome. To classify the excess sequences and characterize the genes, I used the cyclic set defined by circular permutation of local sequences. Each gene can be characterized by the cyclic set of sequences composed of purine/pyrimidines six nucleotides in length (6-mers). Furthermore, error spectra of the 6-mers within *env* were estimated, and codon frames of 6-mers were tested in each gene. The data suggest that local nucleotide sequences *in vivo* affect base misreading by HIV-1 RT and that this virus has strategies for positive- and negative-selection pressures.

MATERIALS AND METHODS

Strains. I selected, considering geographical variations, six sample strains from the AIDS data base (8) for analysis of local sequences that are in excess: BRU, MN, RF, SF2, ELI, and MAL. The hypervariable regions and the conservative—i.e., not hypervariable—regions in the coding region of gp120 are called *gp120*-h and *gp120*-c, respectively, in all. The five regions of the genome sequences of the strains were analyzed: *gp120*-h, *gp120*-c, *gp41* (the coding region of gp41), *gag*, and *pol*. The strains were used for testing codon frames of the 6-mers in the regions. I used the *env* gene variation among strains WMJ1, WMJ2, and WMJ3, serially taken from a single patient (7) for estimating error spectra of the 6-mers.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: HIV-1, human immunodeficiency virus type 1; RT, reverse transcriptase; *gp120*-h, hypervariable regions in the coding region of the extracellular envelope glycoprotein gp120; *gp120*-c, conservative regions in the coding region of gp120; *gp41*, coding region of the transmembrane envelope glycoprotein gp41; 4-sequences, local sequences composed of adenine, cytosine, guanine, and thymine; 2-sequences, local sequences composed of one purine and one pyrimidine; *n*-mer, 2-sequence *n* bases long.

Nomenclature and Lengths of Local Sequences. Local sequences generated by the concatenation of the four nucleotide residues adenine, cytosine, guanine, and thymine are called 4-sequences—e.g., the dinucleotide AG is a 4-sequence in a length of two. Those local sequences composed of two residues, purines and pyrimidines, are called 2-sequences, and 2-sequences n base pairs long are called n-mers.

The hypervariable regions *gp120*-h of each sample strain have only 280–300 bases in total, and the possible numbers of 4- and 2-sequences in lengths of four and eight, which appear, as expected, more than once in *gp120*-h, are both 256. Accordingly, 4-sequences in lengths of one to four and 2-sequences in lengths of one to eight were analyzed.

Estimating Frequency of a Local Sequence in Length n. In the whole sequence of a sample strain (L base pairs long), when a local sequence j in length n occurs p times, the frequency f_j of the sequence j was expressed by a percentage:

$$f_j = [p \times 100/(L - n + 1)]\%, \qquad [1]$$

where $(L - n + 1)$ is the total number of occurrences of the local sequences n base pairs long in the whole sequence.

The frequency F_j of sequence j in the whole sequence of HIV-1, as a viral species, was estimated in the form of the mean value of f_j among sample strains. The frequency R_j of the sequence j in region R of HIV-1 genome was also obtained in the same way.

Extracting Excess Local Sequences in a Region. I have subtracted the frequency F_j of sequence j in the whole sequence from R_j in a region R: $R_j - F_j$. An excess of local sequence in the region is defined as that which has a positive value after subtraction—that is, appears significantly at a higher frequency in the region than in the whole sequence ($P < 0.025$, by the paired-sample t test).

Cyclic Sets and Their Frequency Diversities. The cyclic set, an unusual concept in molecular biology, presented here characterizes the nucleotide sequence. For example, in the sequence composed of three types of nucleotide bases, $a_1a_2a_3a_1a_2a_3a_1a_2a_3a_1a_2a_3$, the local sequence $a_1a_2a_3$ occurs three times, and $a_2a_3a_1$ and $a_3a_1a_2$, defined by the circular permutation of $a_1a_2a_3$, $a_1a_2a_3 \rightarrow a_2a_3a_1 \rightarrow a_3a_1a_2$, occur two times. These local sequences construct the cyclic set $\{a_1a_2a_3, a_2a_3a_1, a_3a_1a_2\}$. In a length of six, $a_1a_2a_3a_1a_2a_3$ and all other members in the cyclic set defined by the circular permutation of $a_1a_2a_3a_1a_2a_3$ also occur in the sequence. In lengths of four and five, however, no cyclic set whose members all occur in the sequence is found (e.g., $a_2a_3a_1a_1$, $a_3a_1a_1a_2$, and $a_1a_1a_2a_3$, defined by the circular permutation of $a_1a_2a_3a_1$, do not occur). Thus, the sequence is characterized by the two cyclic sets in lengths of three and six, but it is not characterized by any cyclic set in lengths of four and five. As shown here, the nucleotide sequence is characterized by cyclic sets.

I express a homo-residue cyclic set composed of a single sequence by putting brackets around the sequence and a hetero-residue cyclic set by putting brackets around the one that is the alphabetically youngest in the set. For example, the homo-residue cyclic set {AAA} is expressed as [AAA]: the hetero-residue cyclic set {AAG, AGA, GAA} is expressed as [AAG]; and {uuy, uyu, yuu}* is expressed as [uuy].

The members of a hetero-residue cyclic set would appear at a different frequency in a region of the nucleotide sequence. The ratio between the largest and smallest in frequencies of members is called the frequency diversity of the cyclic set. When the above example sequence is longer, the frequency diversities of the characteristic cyclic sets in lengths of three and six nearly equal 1.0, while conversely, the frequency diversity of a noncharacteristic set, for example, $[a_1a_2a_3a_1]$, is very large. In each length, in the frequency diversities of the sets, the largest and the smallest are called the maximum and the minimum frequency diversities in length.

Excess Cyclic Sets for a Region. I have categorized a specific cyclic set all members of which are excess sequences in a region as an excess cyclic set for the region. When the excess set for a region is composed of hetero-residues, the more nearly equal to 1.0 its frequency diversity is, the more characteristic the set is for the region. Conversely, in proportion to its frequency diversity, the set becomes less characteristic for the region.

Estimating Error Spectrums of 6-mers. Focusing on base substitutions [because sequence changes in WMJ strain genome variations have been almost solely base substitutions (7)] I estimated error spectra of the 6-mers within *env* from the variations. It was assumed that the consensus sequence among the variations was their ancestor and that a 6-mer j occurred p_j times in it. When bases at position i of the 6-mer j in the consensus were substituted for other ones in the *env* sequence of a WMJ strain, the base substitutions were counted as ones occurring in the strain. When v_{ij} is the number of counted base substitutions at the position occurring in the strain, the mutation frequency m_{ij} at position i of 6-mer j in the strain was defined as follows:

$$m_{ij} = v_{ij}/p_j. \qquad [2]$$

After normalizing the maximum frequency of m_{ij} in the strain to 1.0, the mean value of m_{ij} among the variations was used for the mutation frequency at position i of the 6-mer j within *env* of HIV-1.

Error Spectrums of Cyclic Sets. A cyclic set of the 6-mers defines a single cyclic permutation. The mutation frequency at each position of the 6-mers in the cyclic set, hence, can be presented in the form of the mean value among them at the corresponding position of the cyclic permutation—that is, the error spectrum of the cyclic permutation. I have defined the error spectrum of a cyclic set by that of the cyclic permutation and defined the mutation frequency of the cyclic set by the total of mutation frequencies at the six positions of the cyclic permutation. The presentation is advantageous to discuss mutation frequencies of excess local sequences and gene variabilities, which is to follow.

RESULTS AND DISCUSSION

Frequency Distributions of 4- and 2-Sequences and Frequency Diversities of Hetero-residue Cyclic Sets. I first tested frequencies of 4-sequences in lengths of one to four throughout the entire HIV-1 sequence. BRU, a sample strain of HIV-1, contains 3289 adenines, 1656 cytosines, 2232 guanines, and 2052 thymines in the whole sequence (8). Accordingly, in a length of two, from the nucleotide ratio the dinucleotide CC would be predicted to appear in the lowest-level frequency in the whole sequence but, instead, the dinucleotide CG appeared at the lowest level, as reported by Ohno and Yomo (11) (data not shown). When 4-sequences were grouped by cyclic sets, their frequencies varied more widely in a hetero-residue cyclic set—in particular, those that have the 4-sequences containing CG. Moreover, in proportion to length, both minimum and maximum frequency diversities increased. In each length, frequency diversities were 1.024 of [AT] and 5.122 of [CG], 1.056 of [ACT] and 6.514 of [ACG], and 1.125 of [ACCT] and 42.024 of [ACGT].

Conversely, each member of the hetero-residue cyclic set of shorter 2-sequences showed almost the same frequency in the whole sequence, and even the sets in lengths of seven and eight showed narrower variation of frequencies than those of 4-sequences in lengths of two to four. In a length of two, the

*In this paper the nonstandard abbreviations u and y are used for purine and pyrimidine, respectively.

frequency diversity of the single hetero-residue cyclic set [uy] was 1.000. In lengths of three to eight, the minimum and the maximum frequency diversities were 1.099 of [uuy] and 1.150 of [uyy], 1.017 of [uuuy] and 1.127 of [uyyy], 1.219 of [uuuuy] and 1.466 of [uuuyuy], 1.013 of [uuyuuy] and 1.735 of [uyyyyy], 1.223 of [uuuuuuy] and 2.217 of [uyyyyyy], and 1.056 of [uuuyuuuy] and 3.117 of [uyyyyyyy]. As compared with 4-sequences, the maximum frequency diversity also increased with length; but the minimum one in even-base pair-length was <1.06, and the minimum frequency diversity in odd-base pair-length was >1.09 and increased with length. In particular, the smallest frequency diversity in lengths of three to eight was 1.013 of the 6-base-long set [uuyuuy].

Excess Cyclic Sets for the Hypervariable Regions. Fig. 1 summarizes that the cyclic sets [A], [AA], [AT], [AAC], [AAT], and [ACT] were in excess in the hypervariable regions *gp120*-h, and for the length of four no excess cyclic set was found for *gp120*-h. However, except for [AAT] and [ACT] these cyclic sets also were specific for other regions; [A] and [AA] were in excess in *gp120*-c, *gag*, and *pol*; [AT] was in excess in *gp120*-c and *pol*; and [AAC] was in excess in *gp120*-c. Therefore, the cyclic sets [AAT] and [ACT] were characteristic for *gp120*-h to distinguish it from other regions and suggest that a length of three is specific for HIV-1.

As for 2-sequences, the cyclic sets [u], [uu], [uy], [uuy], and [uuyuuy] were in excess in *gp120*-h; in other lengths no excess cyclic set was found for *gp120*-h (Fig. 1). However, [u] and [uu] were also in excess in *gag* and *pol*; [uy] was in excess in *gp120*-c; and [uuy] was in excess in *pol*. The set [uuyuuy], hence, was the only one excess cyclic set for *gp120*-h that distinguished it from other regions, and this result suggests that a length of six is also characteristic for this virus. In the length, in addition to the 6-mers of [uuyuuy], uuuuyu and four 6-mers of [uuyuyy] were significantly in excess in *gp120*-h.

	gp120-h	*gp120-c*	*gp41*	*gag*	*pol*
1	[A]	[A], [T]	[G], [T]	[A], [C]	[A]
2	[AA], [AT]	[AA], [AC] [AT], [GT] [TT]	[CG], [CT] [GG], [TT]	[AA], [CC] [GG]	[AA], [AT]
3	[AAC], [AAT] [ACT]	[AAA], [AAC] [ATT], [CCC] [TTT]	[CTG], [CTT] [GGG], [GGT]	[AAA], [AGC] [CCC], [GGG]	[AAA], [AGT] [ATG], [ATT]
4	No	[AAAA] [AATT] [ACAC] [ATAT] [CCCC] [GTGT] [TTTT]	[CTTG] [GGTT]	[AAAA] [AAGG] [AGCC] [CCCC] [TTTT]	[AAAA] [AAAC] [AACT]
1	[u]	[y]	[y]	[u]	[u]
2	[uu], [uy]	[uy]	[yy]	[uu]	[uu]
3	[uuy]	No	[yyy]	[uuu]	[uuu], [uuy]
4	No	[uyuy]	[uyyy] [yyyy]	[uuuu] [yyyy]	[uuuu] [uuuy]
5	No	No	[uyyyy]	[uuuuu] [yyyyy]	[uuuuu] [uuuuy]
6	[uuyuuy]	[uyuyuy]	[uyyuyy]	[uuuuuu] [yyyyyy]	[uuuuuu]
7	No	[uuyuyuy] [uyuyuyy]	No	[uuuuuuu] [yyyyyyy]	[uuuuuuu] [uuuuuuy] [uuuuuyy]
8	No	[uyuyuyuy]	No	[uuuuuuuu] [yyyyyyyy]	[uuuyuuuy]

FIG. 1. Excess cyclic sets for the five regions. Numbers in the left column are lengths of sequences. Hypervariable regions *gp120*-h have no excess cyclic set of 4-sequences in the length of four or of 2-sequences in the lengths of four, five, seven, and eight. In each length, *gag* and *pol* had excess cyclic sets, and the homo-residue cyclic sets of 2-sequences were characteristic for *gag*. For *gp120*-c, no excess cyclic set was found in 3-mers and 5-mers; and for *gp41*, no excess cyclic set was found in 7-mers and 8-mers.

Excess Cyclic Sets Reveal the Importance of a Length of Six of 2-Sequences. A length of six of 2-sequences is much more specific for this virus than a length of three of 4-sequences. Fig. 1 shows that each of the three regions *gp120*-h, *gp120*-c, and *gp41* in *env* displayed a distinguishable excess in cyclic sets of 2-sequences only in the length of six. The frequency diversities of the excess sets were <1.2; the largest was 1.193 of [uuyuuy] in *gp120*-h. In the length the *gag* and the *pol* regions also have one or two excess, but homo-residue, cyclic sets of 2-sequences. Hence, the HIV-1 genome displays significant differences in 2-sequences in a length of six among the five regions. By contrast, the regions have many excess cyclic sets of 4-sequences in a length of three, and the frequency diversities of the excess hetero-residue sets in each region were >1.2 (the smallest was 1.232 of [ACT] in *gp120*-h). These results suggest that the cyclic sets of 2-sequences 6 bases long are more characteristic for this virus than those of 4-sequences 3 bases long.

Furthermore, Fig. 1 shows that the hetero-residue cyclic sets containing one or two pyrimidines were characteristic for *pol*, but the region has no excess hetero-residue cyclic set in a length of six. The length of six is the largest one in which the *gp41* region has excess cyclic sets. As stated, in the entire genome sequence, the smallest frequency diversity in lengths of three to eight was of [uuyuuy], and it is very surprising that the set was the one in excess for *gp120*-h. Thus, the results presented here suggest that a length of six of 2-sequences is important for this virus. Moreover, tRNA molecules recognize wobble bases at the third letters of codons and distinguish codons by the presence not of the four bases but of purine/pyrimidine (12–14). For these reasons I suggest that 2-sequences, in particular 6 base pairs long (6-mers), are better than 4-sequences to analyze error spectra in the HIV-1 genome.

Error Spectra Suggest That HIV-1 RT Base Substitutions Are Affected by Local Sequences in the Length of Six. Fig. 2 shows error spectra of cyclic sets of 6-mers estimated from strain WMJ *env* variations (error spectra of 6-mers are omitted). In Fig. 2 the error spectra of the cyclic sets differ from each other. In other words, Fig. 2 suggests that even only a single-base alteration in 6-mers could cause a different error spectrum. For example, the two cyclic sets [uuyuyy] and [uyyuyy] differ in the second letter, and their error spectra differ. The purines of uyy in [uuyuyy] had almost no change, but the ones in [uyyuyy] showed transitions. The first pyrimidines of uyy in [uuyuyy] showed high-level transversions, but the ones in [uyyuyy] showed low-level transversions. Thus, the present results suggest that HIV-1 RT is error-prone to cause base substitutions at specific positions during *in vivo* replication of the viral genome interacting with 6-mers. Furthermore, they suggest a difference in error-proneness of HIV-1 RT between when it acts as a RNA-dependent DNA polymerase and when it acts as a DNA-dependent polymerase (e.g., compare the error spectrum of [uuuuuu] with that of [yyyyyy]).

Excess of Highly Variable 6-mers and Deficiency of Conservative 6-mers in *gp120*-h May Be One Reason Why *gp120*-h Is Hypervariable. Although the molecular basis of nonrandom hypervariability of the type presented here is unknown, the effects on the variability and evolution of HIV-1 might be interpreted as follows. Fig. 2 shows that the cyclic set [uuyuuy] excess for *gp120*-h had the highest mutation frequency as a cyclic set within the WMJ *env*. [uuyuyy], of which most 6-mers, as already mentioned, excessively ap-

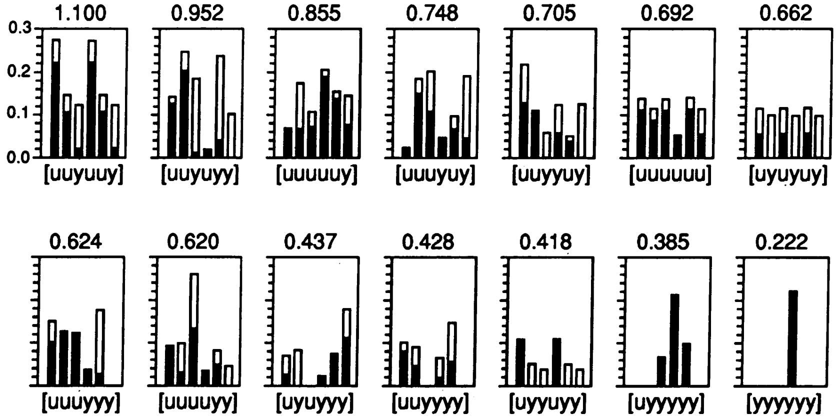

FIG. 2. Error spectra of the cyclic sets in lengths of six, as defined in text, within strain WMJ *env*, and mutation frequencies of cyclic sets (number on each box). Solid bar represents transition, and open bar represents transversion. The axis of ordinates refers to mutation frequency in each position.

peared in *gp120*-h, had the second highest frequency of any cyclic set within the *env*. In particular, at the first purines of subsequences uuy in [uuyuuy] and the second ones of uuy in [uuyuyy], transitions occurred at a much higher level; the purines were, hence, base-substitution hotspots. The first pyrimidines of uyy in [uuyuyy] also showed high-level transversions and were base-substitution hotspots. Conversely, most 6-mers in [uuuyyy], [uuyyyy], [uyyyyy], and [yyyyyy], in which mutations rarely occurred, seldom appeared in *gp120*-h (data not shown). Thus, excess of the 6-mers in the highly variable cyclic sets [uuyuuy] and [uuyuyy] and deficiency of the conservative 6-mers in *gp120*-h may be a reason why *gp120*-h is hypervariable.

In contrast, [yyyyyy], which appeared excessively in *gag* (Fig. 1), changed at the lowest frequency as a cyclic set—i.e., the most conservative, even within strain WMJ *env*. The set [uuuuuu], in excess for *gag* and *pol*, also changed at a lower level than [uuyuuy] and [uuyuyy], even within *env*. The excess 6-mers of [uuuuuu] and [yyyyyy] and their low-level mutations might explain a part of the gene conservation.

Codon Frames of Excess 6-mers for *gp120*-h Differ Between Hypervariable Regions and Conservative Genes. Besides excess, deficiency, and mutation frequency of the local sequences, structure–function relationships in the gene products also decide their variabilities. Because the hotspots in the 6-mers of the highly variable sets [uuyuuy] and [uuyuyy] would seriously affect the relationships according to their positions in codons, I tested three types of codon frames of the 6-mers: 123123, 231231, and 312312. There was a definite one-sided bias: 75–95% of the 6-mers in *gp120*-h made uuy and uyy codons (Fig. 3). Besides the 6-mers of the sets, *gp120*-h had many 6-mers in which the codon tended to be one-sided to make uuy and uyy codons (data of the 6-mers of [uyyuyy] are shown in Fig. 3, and data of the 6-mers of the other cyclic sets are not shown). In particular, 95–100% of the 6-mers of [uyyuyy] corresponded to codon frames that make uyy codons (Fig. 3), whereas in *gag* and *pol* most 6-mers of [uuyuuy], [uuyuyy], and [uyyuyy] (50–70% of the 6-mers) made uyu and yyu codons (Fig. 3). That is, most of the first, second, and third letters of uuy and uyy codons in *gp120*-h corresponded to the third, first, and second letters of uyu and yyu codons in *gag* and *pol*, respectively. The hotspots, the purines of uuy and the first pyrimidines of uyy in the 6-mers of [uuyuuy] and [uuyuyy], therefore, would alter the amino acids in different fashions in the hypervariable regions of the envelope glycoprotein gp120 and in the conservative *gag* and *pol* products. In the next section, amino acid changes caused by the hotspots are discussed.

Amino Acid Changes Caused by the Hotspots are Nonsilent in the Hypervariable Regions and Are Relatively Silent in *gag* and *pol* Products. Amino acid changes caused by hotspots in the hypervariable regions must be discussed relative to function of the regions. The potential N-linked glycosylation sites (Asn-Xaa-Ser/Thr), which occur more frequently in the regions are important for this virus because the oligosaccharides of gp120 have been candidates for viral attachment and addressing factors in the host (15–17), and change of the location or number of the sites could adjust the virus to various host-specific or tissue-specific properties of the cell-membrane receptor. Of the amino acids encoded by uuy and uyy codons, asparagine, serine, and threonine, therefore, would be important in the hypervariable regions. In fact, most amino acids encoded by uuy and uyy in the regions were these amino acids; for example, of the amino acids encoded by uuy of the 6-mer uuyuyy, 87% were asparagine or serine, and 70% of those encoded by uyy were threonine. Furthermore, the hypervariable regions have been understood to aid HIV-1 in evading the host immune system (3, 4).

In the hypervariable regions, amino acid changes caused by mutations in the two purines of frequently occurring codons uuy, which are hotspots, are nonsilent. Fig. 4*A* shows that by the frequent mutations, amino acids encoded by uuy codons are changed into amino acids that could alter the hydrophilicity, surface charge, and secondary structures of the regions or into amino acids that could break or newly form potential glycosylation sites. Amino acid changes by mutations at the second bases of uyy codons, the other hotspots in the excess 6-mers of [uuyuyy] in *gp120*-h, are also nonsilent to transversions, which are almost mutations at the bases, except for mutation C → G of AC(T/C), Thr → Ser,

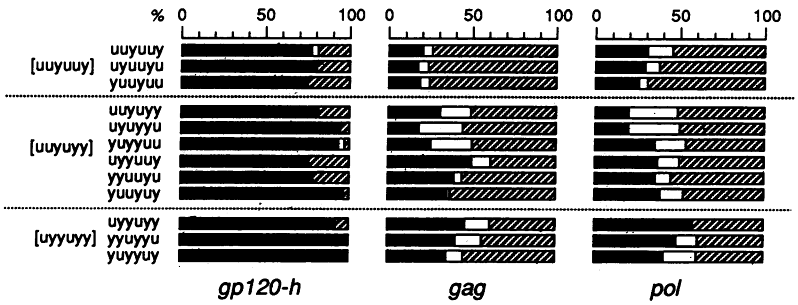

FIG. 3. Codon frames of the 6-mers of cyclic sets [uuyuuy], [uuyuyy], and [uyyuyy] in *gp120*-h, *gag*, and *pol*. Black bar, 6-mers whose uuy and uyy were codons; white bar, 6-mers whose yuu and yuy were codons; and hatched bar, 6-mers whose uyu and yyu were codons. For example, in uuyuuy, black shows codon frame 123123; white shows codon frame 231231; and hatched shows codon frame 312312. In uyuuyu, black presents codon frame 231231; white presents codon frame 312312; and hatched presents codon frame 123123. Percentages show how many 6-mers correspond to each type of codon frame in each region. (SDs are omitted for clarity.)

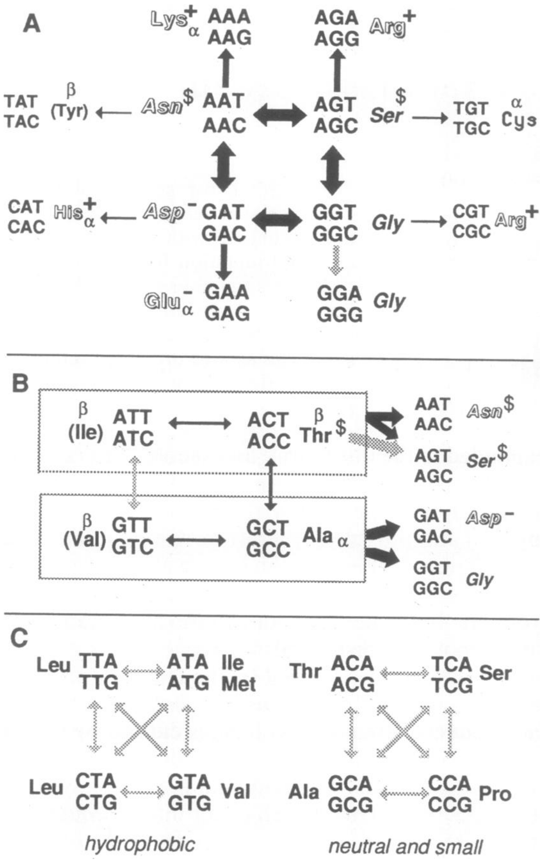

FIG. 4. Transition maps of amino acids caused by mutational hotspots in [uuyuuy] and [uuyuyy]. (*A* and *B*) Transition maps of amino acids encoded by uuy and uyy codons, produced by error spectra of the 6-mers of [uuyuuy] and [uuyuyy], respectively. (*C*) Transition map of amino acids encoded by uyu and yyu codons produced by mutations at the first letters. Solid arrows show nonsilent mutations or mutations that break or newly form potential glycosylation sites in the hypervariable regions; stippled arrow shows a silent mutation. The width of an arrow roughly indicates mutation frequencies. Amino acids in *A* and *B* are classified by the form of characters or by marks: amino acids shown in italics (Asn, Asp, Gly, and Ser) tend to form β-turns; amino acids marked with α (Ala, Glu, His, Lys, and Cys) tend to form α-helixes; amino acids marked with β (Ile, Thr, Tyr, and Val) tend to form β-sheets (18, 19). The amino acids shown by open characters (Arg, Asn, Asp, Glu, His, and Lys) are hydrophilic; amino acids in parentheses are hydrophobic; the others, except for cysteine, are neutral (20). $, amino acids that form potential glycosylation site; + and −, charge on amino acid.

in regard to forming potential glycosylation sites (Fig. 4*B*). Thus, the marked amino acid changes in the hypervariable regions of glycoprotein gp120 caused by hotspots in excess cyclic sets for *gp120*-h have the potential to alter the number and location of potential glycosylation sites in the regions. Moreover, they would change antigenicity of the regions.

In contrast, in *gag* and *pol*, as stated above, the 6-mers of [uuyuuy] and [uuyuyy] were accounted for by uyu and yyu codons, which encode hydrophobic or neutral and small amino acids. Fig. 4*C* shows that of hotspots in the cyclic sets, mutations at the first bases of uyu and yyu codons are silent because amino acid changes caused by the mutations are solely between hydrophobic amino acids or between neutral and small ones. The amino acids tend to be replaced with similar ones, thus leading to conservation of the protein. The third letter of uyu, another hotspot, is clearly silent. Therefore, in spite of hotspots in the 6-mers of [uuyuuy] and [uuyuyy], the products of *gag* and *pol* would be conserved.

I thank Dr. C. C. Martin for critically reading the manuscript and Drs. M. Furusawa, T. A. Kunkel, and S. Ohno for discussions and suggestions.

1. Weiss, A., Hollander, H. & Stobo, J. (1985) *Annu. Rev. Med.* **36**, 545–562.
2. Gallo, R. C. & Montagnier, L. (1988) *Sci. Am.* **259** (4), 40–48.
3. Modrow, S., Hahn, B. H., Shaw, G. M., Gallo, R. C., Wong-Staal, F. & Wolf, H. (1987) *J. Virol.* **61**, 570–578.
4. Starcich, B. R., Hahn, B. H., Shaw, G. M., McNeely, P. D., Modrow, S., Wolf, H., Parks, E. S., Parks, W. P., Josephs, S. F., Gallo, R. C. & Wong-Staal, F. (1986) *Cell* **45**, 637–648.
5. Alizon, M., Wain-Hobson, S., Montagnier, L. & Sonigo, P. (1986) *Cell* **46**, 63–74.
6. Srinivasan, A., Anand, R., York, D., Ranganathan, P., Feorino, P., Schochetman, G., Curran, J., Kalyanaraman, V. S., Luciw, P. A. & Sanchez-Pescador, R. (1987) *Gene* **52**, 71–82.
7. Hahn, B. H., Shaw, G. M., Taylor, M. E., Redfield, R. R., Markham, P. D., Salahuddin, S. Z., Wong-Staal, F., Gallo, R. C., Parks, E. S. & Parks, W. P. (1986) *Science* **232**, 1548–1553.
8. Myers, G., Rabson, A. B., Josephs, S. F., Smith, T. F. & Wong-Staal, F. (1988) *Human Retroviruses and AIDS 1988* (Los Alamos Natl. Lab., Los Alamos, NM).
9. Roberts, J., Bebenek, K. & Kunkel, T. A. (1988) *Science* **242**, 1171–1173.
10. Bebenek, K., Abbotts, J., Roberts, J. D., Wilson, S. H. & Kunkel, T. A. (1989) *J. Biol. Chem.* **264**, 16948–16956.
11. Ohno, S. & Yomo, T. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 1218–1222.
12. Crick, F. H. C. (1966) *J. Mol. Biol.* **19**, 548–555.
13. Yokoyama, S., Watanabe, T., Murao, K., Ishikura, H., Yamaizumi, Z., Nishimura, S. & Miyazawa, T. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 4905–4909.
14. Osawa, S., Jukes, T. H., Muto, A., Yamao, F., Ohama, T. & Andachi, Y. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52**, 777–789.
15. Matthews, T. J., Weinhold, K. J., Lyerly, H. K., Langlois, A. J., Wigzell, H. & Bolognesi, D. P. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 5424–5428.
16. Gruters, R. A., Neefjes, J. J., Tersmette, M., de Goede, R. e. Y., Tulp, A., Huisman, H. G., Miedema, F. & Ploegh, H. L. (1987) *Nature (London)* **330**, 74–77.
17. Wiley, R. L., Smith, D. H., Lasky, L. A., Theodore, T. S., Earl, P. L., Moss, B., Capon, D. J. & Martin, M. A. (1988) *J. Virol.* **62**, 139–147.
18. Chou, P. Y. & Fasman, G. D. (1978) *Annu. Rev. Biochem.* **47**, 251–276.
19. Levitt, M. (1978) *Biochemistry* **17**, 4277–4285.
20. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105–132.