J Mol Evol (1991) 32:360–363

Journal of Molecular Evolution

Mutation Pattern of Human Immunodeficiency Virus Genes

Etsuko N. Moriyama,[1] Yasuo Ina,[1] Kazuho Ikeo,[1,2] Nobuaki Shimizu,[3] and Takashi Gojobori[1]

[1] National Institute of Genetics, Mishima 411, Japan
[2] The Graduate University for Advanced Studies, Mishima 411, Japan
[3] National Institute of Health, Tokyo 141, Japan

Summary. Human immunodeficiency viruses (HIVs) show extensive genetic variation. This feature is the fundamental cause of pathogenicity of HIVs and thwarts efforts to develop effective vaccines. To understand the mutation mechanism of these viruses, we analyzed nucleotide sequences of *env* and *gag* genes of the viruses by use of molecular evolutionary methods and estimated the direction and frequency of nucleotide substitutions. Results obtained showed that the frequency of changes between A and G was extremely high and the mutation pattern of HIVs was distinct from those of nuclear genes of their host cells. This distinction may be caused by the characteristics of the reverse transcription of HIVs. The mutation pattern obtained would be helpful to construct effective antiviral drugs.

Key words: Human immunodeficiency viruses—Molecular evolution—Genetic variation—Mutation pattern—Antiviral drugs

Introduction

Human immunodeficiency viruses (HIVs) are the etiological agents of AIDS (acquired immune deficiency syndrome) in humans. HIVs as well as other retroviruses have RNA genomes and are transcribed by their own reverse transcriptase. It has been known that the RNA genomes of retroviruses evolve at a rate approximately a million times as great as that of eukaryotic nuclear genomes (Holland et al. 1982; Gojobori and Yokoyama 1985, 1987; Temin 1989). For HIVs, comparative studies of nucleotide sequences of their genomes showed that the rate of nucleotide substitutions of these viral genes are also approximately 10^{-3}/site/year (Hahn et al. 1986; Yokoyama and Gojobori 1987; Sharp and Li 1988). Experimental data suggest that the hypermutability of HIVs probably originates in the error-prone feature of the viral reverse transcriptase (Preston et al. 1988; Roberts et al. 1988; Ricchetti and Buc 1990).

The high rate of nucleotide substitutions of HIV genomes causes immunological hypervariability. This is an obstacle for making effective vaccines for the prevention of AIDS. We recently proposed an idea for the systematic development of an HIV vaccine by use of the pattern of amino acid substitutions in the envelope protein (Gojobori and Moriyama 1990). Here we describe detailed analyses for a pattern of nucleotide substitutions of HIV genes. These data prove helpful in understanding the mutation mechanism of HIVs and provide a basis for the systematic construction of antiviral drugs.

Methods

HIVs are classified into two major groups, HIV-1 and -2 (Gojobori et al. 1990). At present, nucleotide sequences of 15 isolates (including 6 from New York) of HIV-1s are comparable to each other for *env* and *gag* genes. Comparing the nucleotide sequences of these viral genes, we can calculate the frequencies of nucleotide substitutions for given nucleotide pairs. To estimate the direction of substitution, we used the phylogenetic trees for *env* and *gag* genes that were previously constructed (Gojobori et al. 1990). The directions were inferred from the tree topology and the alignment of nucleotide sequences (Fig. 1). First, we examined the nucleotide differences site by site in the aligned nucleotide sequences of these isolates. At sites where differences existed, the direction of the change was inferred from the tree topology and majority rule. If two different directions were equally likely, a

Offprint requests to: T. Gojobori

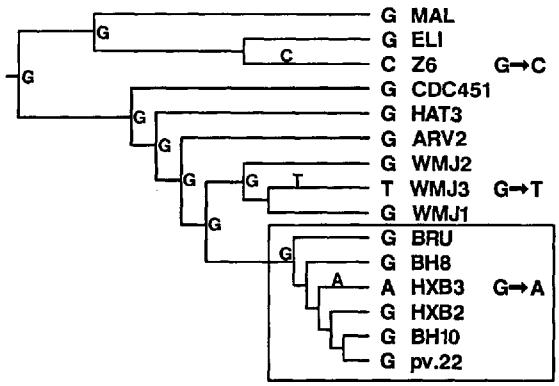

Fig. 1. Estimation of the direction of nucleotide substitution from a phylogenetic tree. To evaluate the pattern of mutation for the *env* gene of HIVs, we used the nucleotide sequences of six isolates of HIV-1s: BRU (Wain-Hobson et al. 1985; Alizon et al. 1986), BH8, BH10 (Ratner et al. 1985a), HXB3 (Crowl et al. 1985; Ratner et al. 1985b), HXB2 (Ratner et al. 1985a,b), and pv.22 (Muesing et al. 1985). These isolates were chosen because they were all derived from patients in New York and seemed to be evolutionarily related. In fact, the phylogenetic tree of HIVs suggests that they have diverged from a common ancestor within only the past decade. For this reason, almost all nucleotide changes between two isolates could be attributed to single substitutions of nucleotides. Other isolates used and their geographical locations: MAL, ELI (Alizon et al. 1986), Z6 (Srinivasan et al. 1987) (Zaire), CDC451 (Desai et al. 1986) (unknown), HAT3 (Ratner et al. 1985b; Starcich et al. 1986), WMJ1, WMJ2, WMJ3 (Hahn et al. 1986; Starcich et al. 1986) (Haiti), and ARV2 (Sanchez-Pescador et al. 1985) (San Francisco).

half-point (0.5) of a single substitution was given for each direction. When three or more directions can be inferred, the site was excluded from the analysis. For example, as shown in Fig. 1, if HXB3 has nucleotide A and all the others have G at a particular site, we assumed that a single nucleotide substitution occurred from G to A. We could identify directions for a total of 80 nucleotide changes in the 2,568 nucleotide sites compared.

Next, we computed the relative substitution frequency for each direction of nucleotide substitution by Gojobori et al.'s (1982) method. It is the frequency of changes from a particular nucleotide to another in a hypothetical sequence of 100 nucleotides with equal amounts of each nucleotide base (25A, 25T, 25C, and 25G). This value is more useful than the row frequency of nucleotide changes because it is unaffected by bias in the base composition in actual sequences.

Results and Discussion

Table 1 shows the relative substitution frequencies for the entire region of the *env* gene of HIV-1s. The frequency of transitional changes was very high. In particular, changes between A and G were much more frequent than any other change. We also counted the number of nucleotide changes at the first, second, and third codon positions separately throughout the *env* region. For all three codon positions in the *env* region, the rate of nucleotide substitution between A and G was always highest, followed by that of substitution between C and T and between C and A. For the third codon position, the rate of the substitution between A and G was more than twofold that of substitution between C and T. These features were true not only in the *env* region but also in the *gag* region (Table 1).

Table 1. Relative substitution frequencies of the *env* and *gag* genes of HIV-1s

	A[a]	T	C	G
env (all positions)				
A[b]	–	5.6	7.5	19.6
T	2.7	–	9.3	1.3
C	9.5	5.7	–	7.6
G	24.4[d]	0.0	6.8	–
[59.0][c]				
gag (all positions)				
A	–	0.0	0.0	16.0
T	0.0	–	31.2	10.4
C	0.0	10.0	–	0.0
G	32.4	0.0	0.0	–
[89.6]				
env (third position)				
A	–	0.0	5.8	17.3
T	2.7	–	11.0	2.7
C	4.1	4.1	–	4.1
G	38.6	0.0	9.6	–
[71.0]				
gag (third position)				
A	–	0.0	0.0	16.4
T	0.0	–	12.3	12.3
C	0.0	15.1	–	0.0
G	43.8	0.0	0.0	–
[87.6]				

[a] Mutated nucleotide
[b] Original nucleotide
[c] Values in brackets represent the frequency of transitional changes
[d] Underlined numbers represent the frequency of changes between A and G

As most of the substitutions at the third codon position do not change the amino acids, the nucleotide changes at that position are mostly free from functional constraints at the protein level. For this reason, the pattern of nucleotide changes at the third codon position can reflect the mutation pattern of the HIV genome. The frequency of changes between A and G was high, as mentioned above, particularly at the third codon position. It is consistent with the pattern of nucleotide substitutions in other retroviral oncogenes, but is different from those of eukaryotic genes, particularly nuclear pseudogenes (Gojobori et al. 1982; Li et al. 1984; Gojobori and Yokoyama 1985; see Table 2). The substitution pattern in pseudogenes can be thought of as the mutation pattern of the nucleotides, because the pseudogenes do not code for any functional proteins and thus can accumulate any kind of mutational change.

Table 2. Relative substitution frequencies of the viral oncogenes (Gojobori and Yokoyama 1985) and nuclear pseudogenes (Gojobori et al. 1982)

	A[a]	T	C	G
Viral oncogenes (all positions)				
A[b]	–	3.5	9.5	14.2
T	4.7	–	15.6	1.6
C	1.0	10.9	–	2.0
G	33.6	3.4	0.0	–
[74.3][c]				
Nuclear pseudogenes				
A	–	4.7	5.2	11.4
T	4.5	–	6.2	4.6
C	8.3	22.0	–	4.7
G	16.0	7.0	5.5	–
[55.6]				

[a] Mutated nucleotide
[b] Original nucleotide
[c] Values in brackets represent the frequency of transitional changes

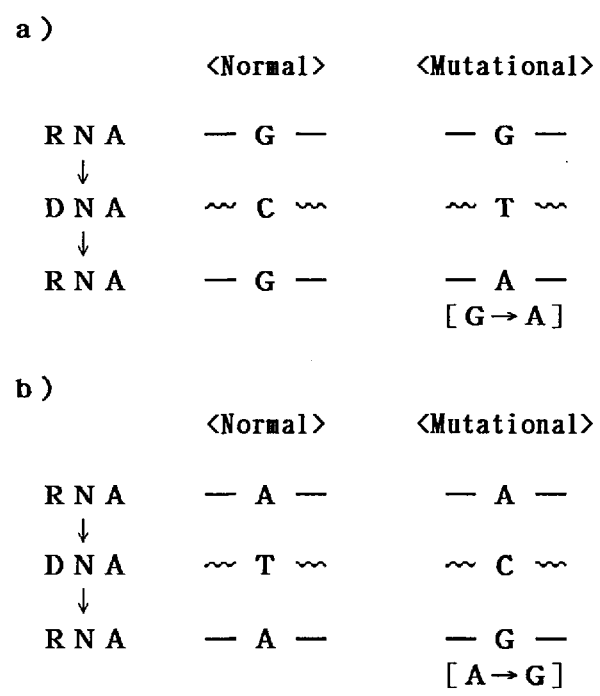

Fig. 2. Replication mechanism of viral reverse transcriptase. The normal and mutational processes of replication are shown. In the case of **a**, G is used as a template, and in the case of **b**, A is used. These mutational processes seem to take place much more frequently in the HIV genomes than in eukaryotic nuclear genomes as shown in Table 1. Analogues of thymidine such as AZT (3'-azido-2',3'-dideoxythymidine or azidothymidine) could inhibit the mutational process of **a**, and those of cytidine such as ddC (2',3'-dideoxycytidine) could inhibit that of **b**. Because mutations from G to A occur at a higher rate than those from A to G (Table 1), AZT is expected to be an antiviral drug more effective than ddC.

Therefore, the high rate of nucleotide substitution between A and G appears to be specific for retroviruses, including HIVs. It implies that the mutation mechanism for retroviruses is different from that for eukaryotic genes.

The replication mechanism unique to the retroviral life cycle is reverse transcription. This process seems to be error-prone (Preston et al. 1988; Roberts et al. 1988; Ricchetti and Buc 1990). Therefore, the high rate of mutation between A and G, particularly G to A, may reflect the characteristics of viral reverse transcriptase. This conclusion has been supported partially by an experimental analysis (Preston et al. 1988). The viral reverse transcriptase may recognize pyrimidines poorly during replication of the HIV genome. Suppose that nucleotide G is at a given site of a viral RNA genome (Fig. 2a). By reverse transcriptase activity, C should be incorporated at a corresponding site for the DNA complementary sequence. However, it seems that T instead of C is misincorporated there. Then, owing to the RNA polymerase of the host, the corresponding site at the HIV RNA genome of the next generation will be occupied by A. This nucleotide change is observed as a substitution from G to A (Fig. 2a). Thus, viral reverse transcriptase seems to recognize pyrimidines poorly when the template is a purine. It explains why pyrimidine analogues, particularly azidothymidine (AZT), are more effective than purine analogues as inhibitors of HIV replication (Fig. 2a and b).

Data obtained in this study are very helpful toward understanding the mutation mechanism of HIVs, which is distinct from that of DNA genomes. The mutation pattern of HIV genomes seems to result from the characteristics of viral reverse transcription and causes immunological hypervariability of HIV proteins, particularly the *env* protein (Shimizu et al. 1989). On the other hand, knowledge about the mutation pattern is also useful for developing effective antiviral drugs and vaccines. These analyses may guide the design of antiviral nucleoside analogues that are preferentially incorporated into viral genomes by reverse transcriptase of HIVs.

Acknowledgment. This study was supported by a grant-in-aid from the Ministry of Education, Science, and Culture, Japan.

References

Alizon M, Wain-Hobson S, Montagnier L, Sonigo P (1986) Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients. Cell 46:63–74

Crowl R, Ganguly K, Gordon M, Conroy R, Schaber M, Kramer R, Shaw G, Wong-Staal F, Reddy EP (1985) HTLV-III *env* gene products synthesized in *E. coli* are recognized by antibodies present in the sera of AIDS patients. Cell 41:979–986

Desai SM, Kalyanaraman VS, Casey JM, Srinivasan A, Andersen PR, Devare SG (1986) Molecular cloning and primary nucleotide sequence analysis of a distinct human immunodefi-

ciency virus isolate reveal significant divergence in its genomic sequences. Proc Natl Acad Sci USA 83:8380–8384

Gojobori T, Moriyama EN (1990) Molecular phylogeny of AIDS viruses and its application to vaccine development. In: Takahata N, Crow JF (eds) Population biology of genes and molecules. Baifukan, Japan, pp 323–340

Gojobori T, Yokoyama S (1985) Rates of evolution of the retroviral oncogene of Moloney murine sarcoma virus and of its cellular homologues. Proc Natl Acad Sci USA 82:4198–4201

Gojobori T, Yokoyama S (1987) Molecular evolutionary rates of oncogenes. J Mol Evol 26:148–156

Gojobori T, Li WH, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369

Gojobori T, Moriyama EN, Ina Y, Ikeo K, Miura T, Tsujimoto H, Hayami M, Yokoyama S (1990) Evolutionary origin of human and simian immunodeficiency viruses. Proc Natl Acad Sci USA 87:4108–4111

Hahn BH, Shaw GM, Taylor ME, Redfield RP, Markham PD, Salahuddin SZ, Wong-Staal F, Gallo RC, Parks ES, Parks WP (1986) Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. Science 232:1548–1553

Holland J, Spindler K, Horodyski F, Grabau E, Nichol S, VandePol S (1982) Rapid evolution of RNA genomes. Science 215: 1577–1585

Li WH, Wu CI, Luo CC (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71

Muesing MA, Smith DH, Cabradilla CD, Benton CV, Lasky LA, Capon DJ (1985) Nucleic acid structure and expression of the human AIDS/lymphadenopathy retrovirus. Nature 313: 450–458

Preston BD, Poiesz BJ, Loeb LA (1988) Fidelity of HIV-1 reverse transcriptase. Science 242:1168–1171

Ratner L, Haseltine W, Patarca R, Livak KJ, Starcich B, Josephs SF, Doran ER, Rafalski JA, Whitehorn EA, Baumeister K, Ivanoff L, Petteway SR Jr, Pearson ML, Lautenberger JA, Papas TS, Ghrayeb J, Chang NT, Gallo RC, Wong-Staal F (1985a) Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature 313:277–284

Ratner L, Starcich B, Josephs SF, Hahn BH, Reddy EP, Livak KJ, Petteway SR Jr, Pearson ML, Haseltine WA, Arya SK, Wong-Staal F (1985b) Polymorphism of the 3′ open reading frame of the virus associated with the acquired immune deficiency syndrome, human T-lymphotropic virus type III. Nucleic Acids Res 13:8219–8229

Ricchetti M, Buc H (1990) Reverse transcriptases and genomic variability: the accuracy of DNA replication is enzyme specific and sequence dependent. EMBO J 9:1583–1593

Roberts J, Bebenek K, Kunkel TA (1988) The accuracy of reverse transcriptase from HIV-1. Science 242:1171–1173

Sanchez-Pescador R, Powe MD, Barr PJ, Steimer KS, Stempien MM, Brown-Shimer SL, Gee WW, Renard A, Randolph A, Levy JA, Dina D, Luciw PA (1985) Nucleotide sequence and expression of an AIDS-associated retrovirus (ARV-2). Science 227:484–492

Sharp PM, Li WH (1988) Understanding the origin of AIDS viruses. Nature 336:315

Shimizu N, Okamoto T, Moriyama EN, Takeuchi Y, Gojobori T, Hoshino H (1989) Patterns of nucleotide substitutions and implications for the immunological diversity of human immunodeficiency virus. FEBS Lett 250:591–595

Srinivasan A, Anand R, York D, Ranganathan P, Feorino P, Schochetman G, Curran J, Kalyanaraman VS, Luciw PA, Sanchez-Pescador R (1987) Molecular characterization of human immunodeficiency virus from Zaire: nucleotide sequence analysis identifies conserved and variable domains in the envelope gene. Gene 52:71–82

Starcich BR, Hahn BH, Shaw GM, McNeely PD, Modrow S, Wolf H, Parks ES, Parks WP, Josephs SF, Gallo RC, Wong-Staal F (1986) Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. Cell 45:637–648

Temin H (1989) Retrovirus variation and evolution. Genome 31:17–22

Wain-Hobson S, Sonigo P, Danos O, Cole S, Alizon M (1985) Nucleotide sequence of the AIDS virus, LAV. Cell 40:9–17

Yokoyama S, Gojobori T (1987) Molecular evolution and phylogeny of the human AIDS viruses LAV, HTLV-III, and ARV. J Mol Evol 24:330–336

Received November 5, 1990/Revised December 25, 1990