Protease gene structure and *env* gene variability of the AIDS virus

Teruo Yasunaga*, Noriyuki Sagata and Yoji Ikawa

*Computation Center and Laboratory of Molecular Oncology, The Institute of Physical and Chemical Research, Wako, Saitama 351-01, Japan

Received 19 February 1986

The protease gene structure and the *env* gene variability have been precisely compared between the AIDS virus and members of the HTLV/BLV family. The conserved amino acid sequence (LVDT) which is repeated in the proteases of the HTLV/BLV family is not repeated in AIDS virus. Comparative analysis of the *env* gene sequences reveals the striking fact that the *env* gene of AIDS virus is 8–12-times more variable than those of the HTLV/BLV family. Within the AIDS virus *env* gene, the surface glycoprotein region is more liable to vary than is the transmembrane region; unexpectedly, however, this liability is not a characteristic feature of the AIDS virus because it is more prominent in other retroviruses including members of the HTLV/BLV family.

AIDS virus *BLV* *HTLV* *Retrovirus evolution* *Protease* *Gene structure* env *gene variability*

1. INTRODUCTION

AIDS viruses were isolated from patients with acquired immune deficiency syndrome (AIDS) and their complete nucleotide sequences have recently been determined for several AIDS virus isolates, human T-cell leukemia/lymphotropic virus type III (HTLV-III) [1], lymphadenopathy-associated virus (LAV) [2] and AIDS-associated retrovirus (ARV-2) [3]. Earlier, the AIDS virus was considered to be a member of the HTLV/BLV (human T-cell leukemia virus type I and bovine leukemia virus) family because it shares some biological properties with members of the HTLV/BLV family [4]. However, it now appears that the overall genomic structure of the AIDS virus is substantially different from that of the HTLV/BLV family [1–3]. In this communication, we report some other genomic features of the AIDS virus, especially focusing on the *env* gene variability, which distinguishes this virus from members of the HTLV/BLV family and from many other retroviruses as well.

2. METHODS

Amino acid sequence difference was calculated as the number of different sites/number of sites compared. Nucleotide sequence difference at synonymous sites was calculated as described [5]. A phylogenetic tree was constructed by the unweighted pair-group clustering method [6].

3. RESULTS AND DISCUSSION

3.1. *Phylogenetic relationships of the AIDS virus to other retroviruses*

We have previously constructed a phylogenetic tree of retroviruses based on the amino acid sequence difference data of the endonuclease domain of their *pol* genes, showing that BLV [7] and HTLV-I [8] are evolutionarily closely related to each other and constitute a distinct group (designated type E) of Oncovirinae (a subfamily of Retroviridae) [7]. Fig.1B shows a newly constructed phylogenetic tree which includes the recently sequenced AIDS virus (HTLV-III isolate)

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/86/$3.50

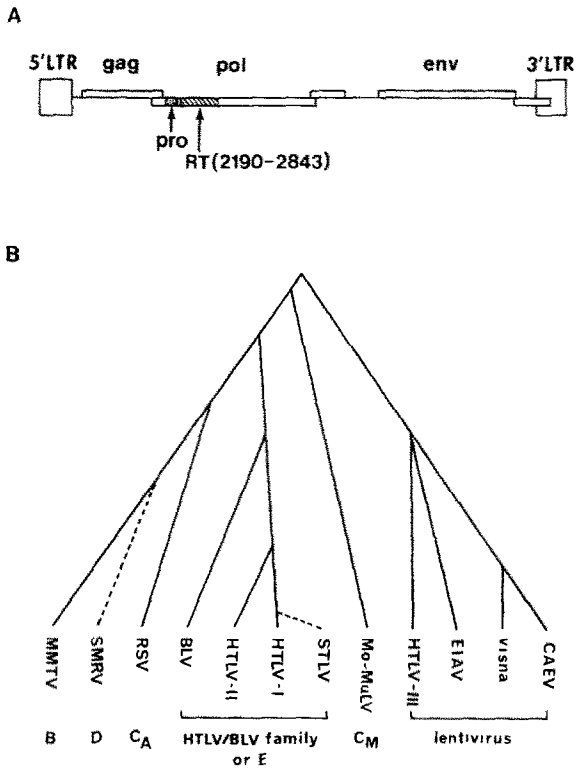

Fig.1. Organization of the AIDS virus genome (A) and phylogenetic tree of retroviruses based on the reverse transcriptase domain of their *pol* genes (B). (A) The genomic structure of the AIDS virus (HTLV-III isolate [1]) is represented schematically. Position of the reverse transcriptase (RT) domain of the *pol* gene used for construction of a phylogenetic tree (B) is indicated. The *gag* precursor-cleaving protease-coding region (*pro*) is also indicated for convenience (see fig.2). (B) The amino acid sequence of the reverse transcriptase domain of the *pol* gene of HTLV-III (amino acid positions 188–405 in [1]) and its corresponding regions of BLV [7], HTLV-I [8], HTLV-II [9], RSV (type C_A virus) [25], MMTV (type B) [13], Mo-MuLV (type C_M) [26], CAEV [13], EIAV [13] and visna [12] are aligned and compared pairwisely to calculate the amino acid differences: type E virus equates the HTLV/BLV family [7]. The tree is constructed by the unweighted pair-group clustering method [6] using the corrected values of the differences for multiple substitutions [23]. The evolutionary relationships of SMRV (type D) [27] and STLV [10] to the other retroviruses (indicated as dashed lines) are inferred from the sequence comparison of their *env* gene (for STLV) or endonuclease region of the *pol* gene (for SMRV) with those of the other retroviruses.

and some lentiviruses (Lentivirinae, a subfamily of Retroviridae): the present tree has been constructed based on the reverse transcriptase domain of the *pol* gene (fig.1A). The tree clearly shows that whereas HTLV-I, HTLV-II (human T-cell leukemia virus type II) [9], STLV (simian T-cell leukemia virus) [10] and BLV constitute a distinct group of retroviruses (HTLV/BLV family or type E), the AIDS virus (HTLV-III) is not included in this family. Rather, the AIDS virus should be classified as a member of lentiviruses which include EIAV (equine infectious anaemia virus) [13], CAEV (caprine arthritis-encephalitis virus) [13] and the visna virus [12] (fig.1B); this conclusion agrees with recent reports [11–13]. It may be interesting to note that the evolutionary relationships of these lentiviruses are topologically similar to those of their hosts; CAEV (goat) and the visna virus (sheep) are more closely related to each other than to any other lentiviruses.

3.2. *Difference in the protease gene structure between the AIDS virus and members of the HTLV/BLV family*

We have already reported that the *gag* precursor-cleaving proteases of both BLV and HTLV-I are encoded between (but out of frame with) the *gag* and *pol* genes, although the proteases of RSV (Rous sarcoma virus) and Mo-MuLV (Moloney murine leukemia virus) are encoded at the 3′-end of the *gag* gene and at the 5′-end of the *pol* gene, respectively [14]. Recently, it has been shown that the proteases of both the AIDS virus and the visna virus, like that of Mo-MuLV, are encoded at the 5′-end of the *pol* gene [1,2,12] (see fig.1A).

We compared precisely the amino acid sequence of the protease among various retroviruses. Fig.2 shows conserved sequences in the proteases of BLV, HTLV-I, HTLV-II, Mo-MuLV, RSV, AIDS virus and visna virus: there are two sequences (LVDTGA or its derivatives, and ILGRD or its derivatives) commonly conserved in all retroviruses compared and one additional short sequence (LVDT or its derivative) conserved only in BLV, HTLV-I and HTLV-II (i.e., HTLV/BLV family or type E virus). Interestingly, the short sequence LVDT conserved only in the HTLV/BLV family is a subsequence of one (LVDTGA) of the two commonly conserved sequences. Possibly, this

```
BLV     : -----LVDTGA-----54-----LVDT--5--ILGRD-----
HTLV-I  : -----LLDTGA----(55)----LVDT-(5)-IIGRD-----
HTLV-II : -----LLDTGA-----55-----LLDT--5--IIGRD-----

MuLV    : -----LVDTGA-------56-----------LLGRD-----
RSV     : -----LLDSGA-------67-----------ILGRD-----
visna   : -----LVDTGA-------57-----------VLGRD-----
HTLV-III: -----LLDTGA-------55-----------IIGRN-----
            91                            156
```

Fig.2. Comparison of the conserved sequences in the proteases of the AIDS virus and the other retroviruses. Highly conserved sequences among the proteases of BLV [7], HTLV-I [8], HTLV-II [9], Mo-MuLV [26], RSV [25], HTLV-III [1] and visna [12] are aligned. The figures indicate the numbers of the amino acid residues between the conserved sequences. For HTLV-I, they are in the parentheses because the potential coding frame for the protease is interrupted by stop codons and a deletion [14]. Amino acid positions are indicated for the HTLV-III protease [1].

short sequence plays an important role in the protease function in the HTLV/BLV family because the longer commonly conserved sequence which contains it is homologous to the active site sequence of acid proteases such as pepsinogen [15]. In contrast, proteases of the AIDS and visna viruses do not have such a short conserved sequence, like those of Mo-MuLV and RSV (fig.2); this may imply that protease function is somewhat different between these retroviruses and members of the HTLV/BLV family. At any rate, the differences in both the coding region of the protease gene and its internal sequence organization distinguish the AIDS virus from members of the HTLV/BLV family.

3.3. *Comparison of the* env *gene variability between the AIDS virus and members of the HTLV/BLV family*

From both comparison of the amino acid sequences and heteroduplex analysis between AIDS virus isolates, it has been pointed out that the *env* gene of the AIDS virus is more highly variable than the other viral genes such as *gag* and *pol* genes [16–18]. However, to date it was not known to what extent the *env* gene of the AIDS virus is variable as compared with those of the other retroviruses. To clarify this point, we have performed a detailed comparison of the *env* gene variability between known AIDS virus isolates, and between members of the HTLV/BLV family as well, using both nucleotide and amino acid sequence difference data.

Table 1 shows that the nucleotide sequence differences at synonymous sites for the *gag*, *pol* and *env* genes are in most cases nearly identical to each other when compared in a given pair of retroviruses; this is consistent with the notion that the synonymous substitutions usually occur approximately at uniform rate among different genes, although the rate of amino acid substitutions for a gene usually differs from gene to gene [5,22]. Among AIDS virus isolates, ARV-2 is significantly different from the other closely related isolates (BH10, H9pv.22, and LAV, table 1), although the almost identical synonymous substitutions (8.7–10%, table 1) between ARV-2 and the other isolate indicate that ARV-2 has also diverged from the common progenitor of the other isolates. The average value of the nucleotide differences at synonymous sites of the three genes (*gag*, *pol* and *env*) between ARV-2 and the other isolates is 9.4% and is comparable to that (11.1%) between BLV-J (Japanese isolate) and BLV-B (Belgian isolate); this means that the divergence time between ARV-2 and the other isolates is roughly equal to that between BLV-J and BLV-B. In contrast, the amino acid sequence differences of the *gag*, *pol* and *env* genes (4.0, 3.7 and 14.7% on average, respectively) between ARV-2 and the other three AIDS virus isolates are all greater than those (2.8, 2.2 and 2.5%) between BLV-J and BLV-B. These results clearly indicate that the amino acid sequences of all three AIDS virus genes are more liable to vary than those of BLV.

By simple calculation, it is now possible to estimate how many times more variable the amino acid sequences of AIDS virus genes are than those of BLV. For this, we first assume that the divergence time between ARV-2 and the other AIDS virus isolates is equal to that between BLV-J and BLV-B. In this situation, the ratio of the corrected amino acid difference of an AIDS virus gene to that of BLV represents the degree of variability of the AIDS virus gene relative to that of BLV (here, the amino acid difference (d) is corrected for multiple substitutions by the equation $-\ln(1-d)$ [23]). Such ratios of the *gag*, *pol* and *env* genes become 1.5, 1.7 and 6.4, respectively. Actually, however, the divergence time (9.4 in terms of the synonymous substitutions) between ARV-2 and the other isolates was slightly smaller than that (11.1) between BLV-J and BLV-B. Therefore, we

Table 1

Comparisons of nucleotide sequence differences at synonymous sites (%) and amino acid sequence differences (%) of the three viral genes among pairs of the HTLV/BLV family and of AIDS virus isolates

Pairs of retroviruses	Nucleotide sequence differences at synonymous sites (%)					Amino acid sequence differences (%)					
	gag	*pol*	*env*	(suf.	tra.)	*gag*	*pol*	*env*	(suf.	tra.	*R*)
HTLV/BLV family											
BLV-J/BLV-B	12.1[a]	11.1[b]	10.2	(10.5	9.9)	2.8[a]	2.2[b]	2.5	(3.3	1.4	2.4)
HTLV-I/HTLV-II	75.3	76.9[c]	82.0	(78.7	87.5)	24.3	36.9[c]	30.9	(36.7	20.9	2.0)
HTLV-I/STLV	–	–	36.3	(36.3	36.4)	–	–	6.6	(8.0	4.0	2.0)
AIDS virus											
BH10/H9pv.22[d]	1.8	1.5	1.1	(1.3	0.9)	0.0	0.3	0.9	(0.6	1.5	0.4)
BH10/LAV	2.6	2.2	1.7	(1.9	1.3)	1.0	1.2	2.1	(2.5	1.5	1.7)
LAV/H9pv.22[d]	0.7	2.2	1.3	(0.6	2.2)	1.0	1.3	1.9	(2.0	1.7	1.2)
ARV-2/BH10	10.0	9.7	9.6	(9.9	9.2)	4.2	3.9	14.6	(16.5	11.9	1.4)
ARV-2/H9pv.22[d]	8.9	9.5	9.3	(8.7	10.1)	4.2	4.0	14.9	(16.5	12.4	1.4)
ARV-2/LAV	8.7	9.0	9.5	(9.4	9.7)	3.6	3.2	14.7	(16.9	11.6	1.5)

[a] Three frame-shifted regions (nucleotide positions 1210–1239, 1315–1344 and 1444–1578 of BLV-J and their corresponding regions of BLV-B) were excluded from calculation

[b] A frame-shifted region (nucleotide position 3112–3120 of BLV-J and its corresponding region of BLV-J) was excluded from calculation

[c] Non-homologous region at the 5′-terminus was excluded from calculation; the region used for calculation is nucleotide position 2590–5184 for HTLV-I and the corresponding portion for HTLV-II

[d] H9pv.22 has a single base deletion in the overlapping region of the *gag* and *pol* genes, which causes frame shifts in both *gag* and *pol* genes of this isolate. For convenience, we used the same reading frames as those of the other AIDS virus isolates

The sequence data are taken from BLV-J [7], BLV-B [19,20], HTLV-I [8], HTLV-II [9], STLV [10], BH10 [1], H9pv.22 [21], LAV [2] and ARV-2 [3]. Amino acid sequence difference (%) was calculated as the number of different sites/number of sites compared × 100. Nucleotide sequence difference at synonymous sites (%) was calculated as described [5]. For the *env* gene, separate calculations for the surface glycoprotein region (suf.) and transmembrane protein region (tra.) were also done and shown in parentheses. *R* in parentheses indicates the ratio of the corrected amino acid sequence difference of the surface glycoprotein region to that of the transmembrane protein region. For both the *gag* and *pol* genes of AIDS virus isolates, only non-overlapping regions of the *gag* (nucleotide positions 347–1629 in [1]) and the *pol* (nucleotide positions 1869–4673 in [1]) genes were used for calculation

can re-estimate that the *gag*, *pol* and *env* genes of the AIDS virus are respectively 1.8-, 2.1- and 7.6-times more variable than those of BLV, by multiplying the above ratios by 1.2; the correction factor 1.2 is calculated as

$$[-\tfrac{3}{4}\ln(1-\tfrac{4}{3}\times 0.111)]/[-\tfrac{3}{4}\ln(1-\tfrac{4}{3}\times 0.094)],$$

where

$$-\tfrac{3}{4}\ln(1-\tfrac{4}{3}d)$$

gives the corrected value of nucleotide sequence difference d for multiple substitutions [24]. Similarly, we can also show that the *env* gene of the AIDS virus is as many as 11.6-times more variable than that of HTLV-I or STLV. Thus, the present data clearly show that among the three viral genes the *env* gene of AIDS virus is especially highly variable as compared with those of the HTLV/BLV family.

3.4. *Sequence variability within the* env *gene of the AIDS virus*

It has been thought that within the AIDS virus *env* gene the surface glycoprotein (amino-terminal half of the *env* protein) region is extremely variable as compared with the transmembrane (carboxyl-

terminal half) region [16,18]. Is this feature more prominent in the AIDS virus than in other retroviruses?

To compare the variability of the surface glycoprotein region of the *env* protein among retroviruses, we calculated the ratio (*R*) of the amino acid sequence difference of the surface glycoprotein region to that of the transmembrane region in each pair of retroviruses (table 1); for calculation, each of the differences was corrected for multiple amino acid substitutions [23]. In pairs of AIDS virus isolates, this ratio ranges from 1.2 (LAV vs H9pv.22) to 1.7 (BH10 vs LAV) (table 1). In pairs of the other retroviruses, on the other hand, the ratios are 2.4 (BLV-J vs BLV-B), 2.0 (HTLV-I vs HTLV-II), 2.0 (STLV vs HTLV-I) and 2.6 (M-MuLV vs AKV, not shown), which are significantly higher than the values obtained for pairs of AIDS virus isolates. Thus, although the surface glycoprotein of the AIDS virus *env* protein is more liable to vary than its transmembrane protein as previously suggested [16–18], this variability is rather smaller than those in the other retroviruses and, therefore, is not a characteristic feature of the AIDS virus.

Our data quantitatively show the unusually high variability of the AIDS virus *env* gene as a whole. The biological significance and mechanism of this phenomenon are not known. However, our data would provide valuable information when one considers eradication of the AIDS disease by such a method as vaccination.

ACKNOWLEDGEMENTS

We are grateful to Drs Naofumi Ogita and Takashi Miyata for discussion and encouragement. This work was supported in part by a grant from the Ministry of Education, Science and Culture of Japan and in part by Research Grant of the Princess Takamatsu Cancer Research Fund.

REFERENCES

[1] Ratner, L., Haseltine, W., Patarca, R., Livak, K.J., Starcich, B., Josephs, S.F., Doran, E.R., Rafalski, J.A., Whitehorn, E.A., Baumeister, K., Ivanoff, L., Petteway, S.R. jr, Pearson, M.L., Lautenberger, J.A., Papas, T.S., Ghrayeb, J., Chang, N.T., Gallo, R.C. and Wong-Staal, F. (1985) Nature 313, 277–284.

[2] Wain-Hobson, S., Sonigo, P., Danos, O., Cole, S. and Alizon, M. (1985) Cell 40, 9–17.

[3] Sanchez-Pescador, R., Power, M.D., Barr, P.J., Steimer, K.S., Stempien, M.M., Brown-Shimer, S.L., Gee, W.W., Renard, A., Randolph, A., Levy, J.A., Dina, D. and Luciw, P.A. (1985) Science 227, 484–492.

[4] Arya, S.K., Gallo, R.C., Hahn, B.H., Shaw, G.M., Popovic, M., Salahuddin, S.Z. and Wong-Staal, F. (1984) Science 225, 927–930.

[5] Miyata, T. and Yasunaga, T. (1980) J. Mol. Evol. 16, 23–36.

[6] Sokal, R.R. and Sneath, P.H. (1963) Principles of Numerical Taxonomy, Freeman, San Francisco.

[7] Sagata, N., Yasunaga, T., Tsuzuku-Kawamura, J., Ohishi, K., Ogawa, Y. and Ikawa, Y. (1985) Proc. Natl. Acad. Sci. USA 82, 677–681.

[8] Seiki, M., Hattori, S., Hirayama, Y. and Yoshida, M. (1983) Proc. Natl. Acad. Sci. USA 80, 3618–3622.

[9] Shimotohno, K., Takahashi, Y., Shimizu, N., Gojobori, T., Golde, D.W., Chen, I.S.Y., Miwa, M. and Sugimura, T. (1985) Proc. Natl. Acad. Sci. USA 82, 3101–3105.

[10] Watanabe, T., Seiki, M., Tsujimoto, H., Miyoshi, I., Hayami, M. and Yoshida, M. (1985) Virology 144, 59–65.

[11] Wain-Hobson, S., Alizon, M. and Montagnier, L. (1985) Nature 313, 743.

[12] Sonigo, P., Alizon, M., Staskus, K., Klatzmann, D., Cole, S., Danos, O., Retzel, E., Tiollais, P., Haase, A. and Wain-Hobson, S. (1985) Cell 42, 369–382.

[13] Chiu, I.M., Yaniv, A., Dahlberg, J.E., Gazit, A., Skuntz, S.F., Tronick, S.R. and Aaronson, S.A. (1985) Nature 317, 366–368.

[14] Sagata, N., Yasunaga, T. and Ikawa, Y. (1984) FEBS Lett. 178, 79–82.

[15] Toh, H., Kikuno, R., Hayashida, H., Miyata, T., Kugimiya, W., Inouye, S., Yuki, S. and Saigo, K. (1985) EMBO J. 4, 1267–1272.

[16] Ratner, L., Gallo, R.C. and Wong-Staal, F. (1985) Nature 313, 636–637.

[17] Rabson, A.B. and Martin, M.A. (1985) Cell 40, 477–480.

[18] Hahn, B.H., Gonda, M.A., Shaw, G.M., Popovic, M., Hoxie, J.A., Gallo, R.C. and Wong-Staal, F. (1985) Proc. Natl. Acad. Sci. USA 82, 4813–4817.

[19] Rice, N.R., Stephens, R.M., Couez, D., Deschamps, J., Kettmann, R., Burny, A. and Gilden, R.V. (1984) Virology 138, 82–93.

[20] Rice, N.R., Stephens, R.M., Burny, A. and Gilden, R.V. (1985) Virology 142, 357–377.

[21] Muesing, M.A., Smith, D.H., Cabradilla, C.D., Benton, C.V., Lasky, L.A. and Capon, D.J. (1985) Nature 313, 450–458.

[22] Miyata, T., Yasunaga, T. and Nishida, T. (1980) Proc. Natl. Acad. Sci. USA 77, 7328–7332.
[23] Zukerkandle, E. and Pauling, L. (1965) Evolving Genes and Proteins (Bryson, V. and Vogel, H.J. eds) pp.97–106, Academic Press, New York.
[24] Jukes, T.H. and Cantor, C.R. (1969) Mammalian Protein Metabolism, II (Munro, H.N. ed.) pp.21–123, Academic Press, New York.
[25] Schwartz, D.E., Tizard, R. and Gilbert, W. (1983) Cell 32, 853–869.
[26] Shinnick, T.M., Lerner, R.A. and Sutcliffe, J.G. (1981) Nature 293, 543–548.
[27] Chiu, I.M., Callahan, R., Tronick, S.R., Schlom, J. and Aaronson, S.A. (1984) Science 233, 364–370.