Identification of a potential protease-coding gene in the genomes of bovine leukemia and human T-cell leukemia viruses

Noriyuki Sagata, Teruo Yasunaga[+] and Yoji Ikawa

Laboratory of Molecular Oncology and [+]*Computation Center, The Institute of Physical and Chemical Research, Wako, Saitama 351, Japan*

Received 22 September 1984

The genomes of bovine leukemia and human T-cell leukemia viruses both contain an unidentified region between the *gag* and *pol* genes. These regions harbor an open reading frame that is in a different phase from the reading frames of the *gag* and *pol* genes. Based on the deduced amino acid sequences, we show here that they potentially encode a *gag* precursor-cleaving protease, which is known to be fused to the *gag* and *pol* products of avian and murine retroviruses, respectively. This finding raises the interesting question of the expression and evolution of retroviral genes.

Protease gene *Retrovirus genome* *Sequence comparison* *Evolution*

1. INTRODUCTION

Replication-competent retroviruses have only 3 structural genes: the *gag, pol* and *env* genes, in this order [1]. The *gag* gene encodes core proteins, which are initially produced as a *gag* precursor polyprotein [2]. A protease that cleaves the *gag* precursor is encoded at the 3′ end of the *gag* gene of the Rous avian sarcoma virus (RSV) [2,3], and probably at the 5′ end of the *pol* gene of the Moloney murine leukemia virus (M-MuLV) [4]. Human T-cell leukemia virus (HTLV), causing adult T-cell leukemia [5,6], and bovine leukemia virus (BLV), causing enzootic bovine leukosis [7], are closely related [8,9] and their entire provial genomes have now been sequenced [10,11]. Curiously, however, no possible protease-coding regions have yet been identified in their genomic structures. We show here that these retroviruses do encode a protease but in an independent open reading frame between the *gag* and *pol* genes.

2. MATERIALS AND METHODS

For sequence comparisons, we used the published nucleotide and deduced amino acid sequences of BLV [11], HTLV [10], M-MuLV [12] and RSV [3]. The homology matrix comparisons and sequence alignments were performed as described [13,14].

3. RESULTS AND DISCUSSION

We have recently determined the complete nucleotide sequence of the proviral genome of BLV [9,11]. The genomic structures of the M-MuLV [12], RSV [3], HTLV [10] and BLV [11] proviruses are compared in fig.1. The BLV and HTLV genomes share a common structure: 5′ LTR-*gag*-intercistronic region-*pol*-*env*-$pX_{(BL)}$-3′ LTR, where LTR represents a long terminal repeat, and $pX_{(BL)}$ an unidentified region. The presence of the $pX_{(BL)}$ sequences in these retroviruses distinguishes them from the other retroviruses [11]. Another remarkable feature of these retrovirus genomes is the existence of a very long intercistronic region (510 base pairs (bp) for BLV and 405 bp for HTLV) between the *gag* and *pol* genes; in M-MuLV and RSV, this intercistronic region is quite short (only a 3 bp amber stop codon

Published by Elsevier Science Publishers B.V.

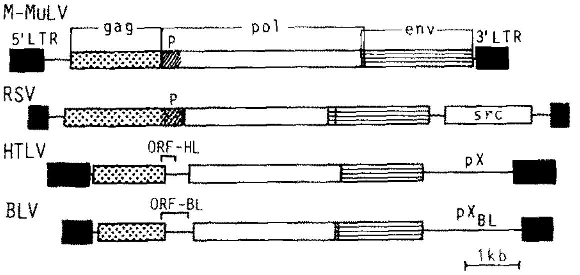

Fig.1. Comparison of the genomic structures of M-MuLV, RSV, HTLV and BLV. The genome of each virus is shown schematically in the proviral form (M-MuLV, RSV, HTLV and BLV from [12,3,10,11], respectively). The pX and pX$_{BL}$ in HTLV and BLV are unidentified regions. Domains of the putative M-MuLV protease (the N-terminal region of the *pol* gene) [4] and the RSV protease (the fifth *gag* protein) [2,3] are hatched and named P. See text and fig.2 for explanation of ORF-BL and ORF-HL. ▬ , LTR; ▨ , *gag* gene; □ , *pol* gene; ☰ , *env* gene.

for M-MuLV and 20 bp for RSV) (fig.1). Interestingly, the BLV and HTLV intercistronic regions correspond closely to a (*gag* precursor-cleaving) protease-coding region of the M-MuLV and RSV genomes (see fig.1 legend). These observations prompted us to examine whether the intercistronic regions represent as yet unidentified protease-coding genes of the BLV and HTLV genomes.

Inspection of the nucleotide sequence around the *gag-pol* junction of the BLV genome reveals an open reading frame (designated as ORF-BL), spanning nucleotides 1755–2258 and encoding 167 amino acid residues (fig.2a). HTLV also contains an open reading frame (designated as ORF-HL) in the corresponding position (nucleotides 2051–2318), but encoding a much smaller protein of 88 amino acids (fig.2b). Neither ORF-BL nor ORF-HL opens with an initiator ATG codon, and the 5′ ends of both overlap the 3′ ends of the *gag* genes. These open reading frames are both in a different phase from the reading frames of the *gag* and *pol* genes. To assess whether ORF-BL and ORF-HL encode a protease, we initially compared their deduced amino acid sequences with those of the M-MuLV and RSV proteases by a 2-dimensional homology matrix [13], finding partial but significant homologies between them (not shown). Based on this finding, we aligned the 4 sequences with minimal gaps to maximize the homology [14]. This gave the result in fig.3a, where the M-MuLV sequence is derived from the N-terminal 110-residue sequence (putative protease domain) of the *pol* product [4,12], while the RSV sequence includes, besides the fifth *gag* protein (p15) as a protease, a C-terminal 18-residue sequence of the fourth *gag* protein (p12) [2,3], because this has appreciable homology with the N-terminal regions of the other

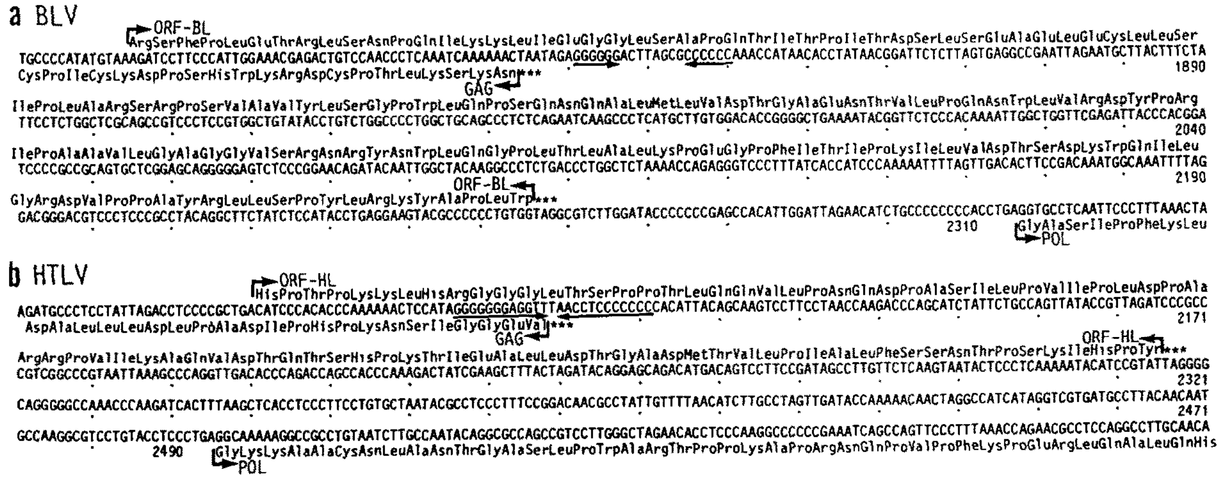

Fig.2. Nucleotide and deduced amino acid sequences of the open reading frames found arround the *gag-pol* junctions of the (a) BLV and (b) HTLV genomes. In both BLV [11] and HTLV [10], a 600 bp sequence spanning the *gag-pol* junction is presented. (a) ORF-BL and (b) ORF-HL represent open reading frames. Arrowed underlines denote palindromic sequences.

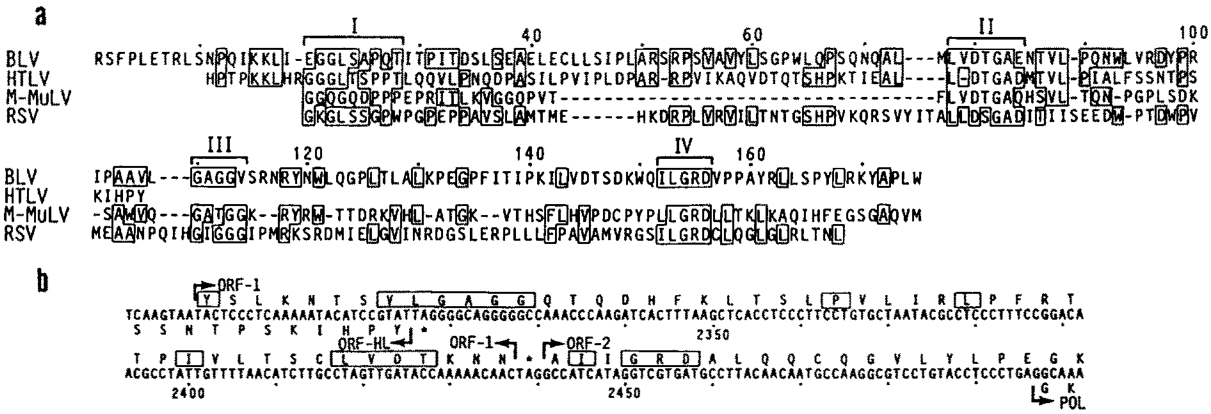

Fig.3. Amino acid sequence homologies between ORF-BL, ORF-HL and proteases of M-MuLV and RSV (a) and open reading frames encoding the possible C-terminal half of the putative HTLV protease (b). (a) Amino acid residues are expressed as one-letter codes. The BLV (ORF-BL) and HTLV (ORF-HL) sequences are from fig.2. The M-MuLV sequence is derived from an N-terminal 110-residue sequence of the *pol* gene [12], which probably represents a 13-kDa protease [4]. The RSV sequence includes a C-terminal 18-residue sequence of the fourth *gag* protein (p12) immediately preceding the protease (the fifth *gag* protein, 15 kDa) [2,3]. Conserved residues are boxed and highly conserved regions and marked I, II, III and IV. (b) Other open reading frames that could encode a possible C-terminal half of the putative HTLV protease. ORF-2 is separated from ORF-1 by a single amber stop codon. Assuming this stop codon is read through, the deduced amino acid sequence can be aligned with the C-terminal half of the ORF-BL sequence (positions 89–175 in (a)) without any gaps, and matched residues are boxed.

viral sequences. The overall homologies (counting gaps) between the sequence of ORF-BL and those of the M-MuLV and RSV proteases reach 25%. Moreover, there are 4 highly conserved regions in the 3 sequences (marked I, II, III and IV in fig.3a), each as long as 4–7 amino acid residues and located 30–50 residues from each other. Thus, the facts that ORF-BL is located at a position corresponding to the protease-coding regions of the M-MuLV and RSV genomes (fig.1) and that its deduced amino acid sequence bears appreciable homology with both protease sequences (fig.3a) strongly suggest that ORF-BL encodes a BLV protease. The amino acid sequence of ORF-HL is only about half as long as those of the ORF-BL and M-MuLV and RSV proteases and can be aligned with only their N-terminal 85-residue sequences (fig. 3a). It shows the highest homology (29%) with the putative BLV protease (ORF-BL), reflecting the closest relationship of HTLV with BLV [11], and also contains well conserved sequences (regions I and II). Thus it presumably represents an HTLV protease, in spite of its small size.

The apparent lack of a C-terminal 70-residue sequence in the putative HTLV protease is curious, because in the other viral proteases such sequences commonly contain well conserved sequences (regions III and IV, fig.3a). Further inspection of the HTLV sequence around and downstream of the ORF-HL revealed two more open reading frames, ORF-1 (nucleotides 2291–2437; fig.3b) and ORF-2 (nucleotides 2441–2752). The 5′ end of ORF-1 overlaps the 3′ end of ORF-HL, and ORF-2 is separated from ORF-1 by a single amber stop codon and ends with the 259th nucleotide of the *pol* gene. Surprisingly, these open reading frames contain two amino acid sequences (Gly-Ala-Gly-Gly in ORF-1 and Ile-Ile-Gly-Arg-Asp in ORF-2) that are homologous with the well conserved sequences (regions III and IV, respectively, in fig.3a) of the C-terminal halves of the other proteases. Remarkably, sequence alignment with the C-terminal half of the putative BLV protease can be performed without any gaps and reveals 3 highly homologous regions (two of these correspond to the above-mentioned conserved sequences), as shown by the boxed residues in fig.3b. Thus, we suggest that ORF-1 and ORF-2 together represent the C-terminal half of the putative HTLV protease, although they are separated by a single stop codon.

At present, we do not know the actual N- and C-termini of the putative BLV and HTLV proteases, nor do we know how their genes are transcribed and translated. However, the BLV protease is possibly produced as a *gag*-protease polyprotein because BLV mRNA produces a 70-kDa polyprotein that contains, besides a 45-kDa *gag* precursor, an additional but unknown polypeptide [7,15,16], which can be assumed to be a protease. If so, the mRNA for the polyprotein might be generated by such splicing as removes the stop codon between, and causes the in-frame fusion of, the *gag* and protease genes. Regarding this, we note that both BLV and HTLV harbor a palindromic structure around the *gag*-protease junction (see arrowed underlines in fig.2); such a palindromic structure may be involved in the possible splicing events [17]. The BLV protease may also be generated as a 145-kDa polyprotein (a putative *gag-pol* precursor [15]), since this molecule contains all the tryptic peptides of the 70-kDa polyprotein [7].

The present observations raise the interesting question of how organization of the protease gene and its surrounding genes has changed during retrovirus evolution. Roughly speaking, two models seems to be possible. One is that the primordial retrovirus had a single gene for *gag*-protease-*pol* and that during evolution this underwent mutation, resulting in a stop codon(s) at either or both ends of the protease-coding region, thereby producing a *gag*-fused (for RSV), *pol*-fused (for M-MuLV) or independent (for BLV and HTLV) protease gene. This model assumes the existence of a single gene of unusually enormous size and requires at least two steps to generate the independent protease gene of BLV or HTLV. The other simpler model that we favour is that the primordial retrovirus had 3 independent *gag*, protease and *pol* genes, like the present BLV (or HTLV). In this case, only a single step mutation, to abolish a stop codon and to cause a frameshift, between either the *gag* and protease genes or the protease and *pol* genes would generate a *gag*-fused (for RSV) or *pol*-fused (for M-MuLV) protease gene. This model is consistent with the protovirus hypothesis [18]. Further analysis of the genomic structures of other types of viruses would shed light on this problem.

ACKNOWLEDGEMENTS

We thank Mrs J. Tsuzuku-Kawamura and Miss K. Ohishi for preparation of the figures. This work was supported in part by a Grant-in-Aid for Cancer Research from the Ministry of Education, Science and Culture of Japan.

REFERENCES

[1] Coffin, J. (1982) in: Molecular Bilogy of Tumor Viruses, RNA Tumor Viruses (Weiss, R. et al. eds) pp. 261-368, Cold Spring Harbor Laboratory, NY.

[2] Dickson, C., Eisenman, R., Fan, H., Hunter, E. and Teich, N. (1982) in: Molecular Biology of Tumor Viruses, RNA Tumor Viruses (Weiss, R. et al. eds) pp. 513-648, Cold Spring Harbor Laboratory, NY.

[3] Schwartz, D.E., Tizard, R. and Gilbert, W. (1983) Cell 32, 853-869.

[4] Levin, J.G., Hu, S.C., Rein, A., Messer, L.I. and Gerwin, B.I. (1984) J. Virol. 51, 470-478.

[5] Poiesz, B.J., Ruscetti, F.W., Gazdar, A.F., Bunn, P.A., Minna, J.D. and Gallo, R.C. (1980) Proc. Natl. Acad. Sci. USA 77, 7415-7419.

[6] Yoshida, M., Miyoshi, I. and Hinuma, Y. (1982) Proc. Natl. Acad. Sci. USA 79, 2031-2035.

[7] Burny, A., Bruck, C., Chantrenne, H., Cleuter, Y., Dekegel, D., Ghysdael, J., Kettmann, R., Leclercq, M., Leunen, J., Mammerickx, M. and Portetelle, D. (1980) in: Viral Oncology (Klein, G. ed.) pp. 231-289, Raven, New York.

[8] Copeland, T.D., Oroszlan, S., Kalyanaraman, V.S., Sarngadharan, M.G. and Gallo, R.C. (1983) FEBS Lett. 162, 390-395.

[9] Sagata, N., Yasunaga, T., Ogawa, Y., Tsuzuku-Kawamura, J. and Ikawa, Y. (1984) Proc. Natl. Acad. Sci. USA 81, 4741-4745.

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

[11] Sagata, N., Yasunaga, T., Tsuzuku-Kawamara, J., Ohishi, K., Ogawa, Y. and Ikawa, Y. (1984) Proc. Natl. Acad. Sci. USA, in press.

[12] Shinnick, T.M., Lerner, R.A. and Sutcliffe, J.G. (1981) Nature 293, 543-548.

[13] Toh, H., Hayashida, H. and Miyata, T. (1983) Nature 305, 827-829.

[14] Needleman, S.B. and Wunsh, C.D. (1970) J. Mol. Biol. 48, 443-453.

[15] Ghysdael, J., Kettmann, R. and Burny, A.J. (1979) J. Virol. 29, 1087-1098.

[16] Mamoun, R.Z., Astier, T., Guillemain, B. and Duplan, J.F. (1983) J. Gen. Virol. 64, 1895-1905.

[17] Laprevotte, I., Hampe, A., Sherr, C.J. and Galibert, F. (1984) J. Virol. 50, 884-894.

[18] Temin, H.M. (1974) Annu. Rev. Genet. 8, 155-177.