The EMBO Journal vol.3 no.13 pp.3231 – 3237, 1984

Comparison of the entire genomes of bovine leukemia virus and human T-cell leukemia virus and characterization of their unidentified open reading frames

Noriyuki Sagata, Teruo Yasunaga[1], Kazue Ohishi, Junko Tsuzuku-Kawamura, Misao Onuma[2] and Yoji Ikawa

Laboratory of Molecular Oncology and [1]Computation Center, Institute of Physical and Chemical Research, Wako Saitama 351, and [2]Department of Veterinary Microbiology, College of Dairy Agriculture, Ebetsu 069-01, Hokkaido, Japan

Communicated by P. Kourilsky

We have compared the sequences of the entire genomes of bovine leukemia virus (BLV) and human T-cell leukemia virus type I (HTLV-I). Both the *gag* and *pol* genes show overall strong homologies indicating the close evolutionary relationship of the two retroviruses. However, a surface glycoprotein portion of the *env* gene shows no appreciable homology, which probably reflects a difference in their host ranges. The 3′ end portion of the BLV genome (designated as pX_{BL}) contains an unidentified long open reading frame that has a typical protein-coding property. The potential product of this open reading frame may be a glycoprotein of ~40 000 daltons. We note that its amino acid sequence shows low but appreciable homology, especially in its N-terminal quarter, to that of the HTLV-I counterpart (pX product), and we thus suggest that BLV pX_{BL} and HTLV-I pX have diverged from a common ancestral gene. It is tentatively concluded that both the putative pX_{BL} and pX products are respectively produced from a spliced mRNA.

Key words: homology matrix/pX and pX_{BL}/retrovirus evolution/splice acceptor/BLV/HTLV

Introduction

A number of replication-competent retroviruses are implicated in both naturally-occurring and experimentally-induced leukemias (Teich *et al.*, 1982). Their proviral genomes usually consist of only three structural genes flanked by long terminal repeats (LTRs) arranged in the order 5′ LTR-*gag-pol-env*-3′ LTR (Coffin, 1982).

Human T-cell leukemia virus type I (HTLV-I) is an exogenous, replication-competent retrovirus and is implicated in human adult T-cell leukemia (Poiesz *et al.*, 1980; Yoshida *et al.*, 1982; Weiss, 1982). The genomic structure of HTLV-I, deduced from its complete nucleotide sequence, is quite unique in that it harbors, downstream of the *env* gene, an unidentified region called pX (Seiki *et al.*, 1983). Although HTLV-I carries no typical oncogene of cellular origin (Seiki *et al.*, 1983), it can transform T-lymphocytes *in vitro* (Miyoshi *et al.*, 1982, 1983; Yamamoto *et al.*, 1982a). In addition, HTLV-I has no preferred chromosomal site for integration (Seiki *et al.*, 1984) and often lacks the 5′ half of its genome in the leukemic cells (Yoshida, 1983). It is thus conceivable that the HTLV-I genome (presumably its 3′ half) harbors a gene(s) that is somehow involved in the transforming events; in this connection, the pX region has been noted because it is located at the 3′ end of the viral genome and is capable of encoding an unidentified protein(s) (Seiki *et al.*, 1983; Haseltine *et al.*, 1984).

Bovine leukemia virus (BLV), a causative agent of enzootic bovine leukosis, is also an exogenous, replication-competent virus and contains no typical oncogene (Burny *et al.*, 1980; Deschamps *et al.* 1981). BLV shares several biological properties with HTLV-I (Burny *et al.*, 1980; Weiss, 1982) and, in fact, appears to be evolutionarily related judging from the appreciable homology between their core proteins (Oroszlan *et al.*, 1982; Copeland *et al.*, 1983b) and from the unique structures of their LTRs (Sagata *et al.*, 1984a). The ability of BLV to transform cells *in vitro* (Onuma *et al.*, 1981; Rhim *et al.*, 1983), the absence of preferred chromosomal sites for proviral integration (Kettmann *et al.*, 1983; our unpublished observation) and the presence of 5′ half-truncated proviruses in the leukemic cells (Kettmann *et al.*, 1982) also suggest that the 3′ half of the BLV genome possesses some gene implicated in cellular transformation. Collectively, it appears that the genomic structure of BLV has substantial similarity to that of HTLV-I, although this has not been elucidated so far at a molecular (nucleotide) level.

We have been interested in the evolutionary relationship of BLV to HTLV-I and the genomic structure of BLV (especially in its 3′ half) (Sagata *et al.*, 1983, 1984a), and we have recently determined the complete nucleotide sequence of the integrated BLV provirus (Sagata *et al.*, 1984b): the genomic structure of BLV proves to be very similar to that of HTLV-I, possessing at its 3′ end a region (designated as pX_{BL}) that corresponds to the pX region of HTLV-I. Based on this complete nucleotide sequence of BLV and the previously reported one of HTLV-I (Seiki *et al.*, 1983), we show here the detailed comparison of the entire genomes of these two retroviruses, including characterization of the pX_{BL}.

Results and Discussion

Comparison of the entire genomes of BLV and HTLV-I at a nucleotide level

The proviral genome of BLV consists of 8714 bp (Sagata *et al.*, 1984b), while that of HTLV-I consists of 9032 bp (Seiki *et al.*, 1983). They have a common genomic structure in the form of 5′ LTR-*gag-pol-env*-$pX_{(BL)}$ (pX for HTLV-I and pX_{BL} for BLV)-3′ LTR. Figure 1 shows a two-dimensional homology matrix comparison, under conditions detecting 20-bp unit sequences with >70% homology, of their entire genomes at the nucleotide level. On a nearly diagonal angle of the square frame of the matrix can be seen parts of a line, showing the existence of partial homology between the nucleotide sequences of the two retroviruses. They have the strongest homology in the *pol* gene (especially in its 5′-terminal region), intermediate homology in the *gag* gene and 3′ half of the *env* gene but no detectable homology in the LTR, *gag-pol* junction, 5′ half of the *env* gene or $pX_{(BL)}$ regions. Similar homology matrix comparisons between BLV and other (murine and avian) retroviruses showed virtually no

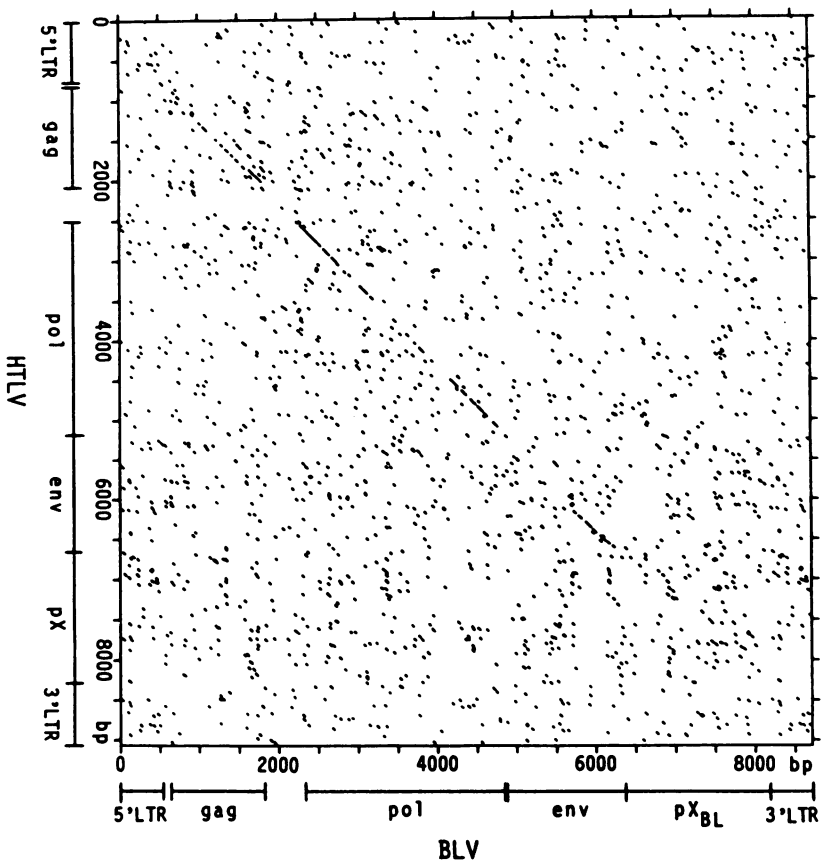


Fig. 1. A two-dimensional homology matrix comparison of the complete nucleotide sequences of BLV and HTLV-I. Complete nucleotide sequences of BLV and HTLV-I were respectively taken from Sagata *et al.* (1984b) and Seiki *et al.* (1983). Each short line (appearing as a 'dot' in the figure) represents a region where there is at least 70% homology in 20 contiguous nucleotides between the two viral genomes. Locations of the respective genes are indicated outside the square frame.

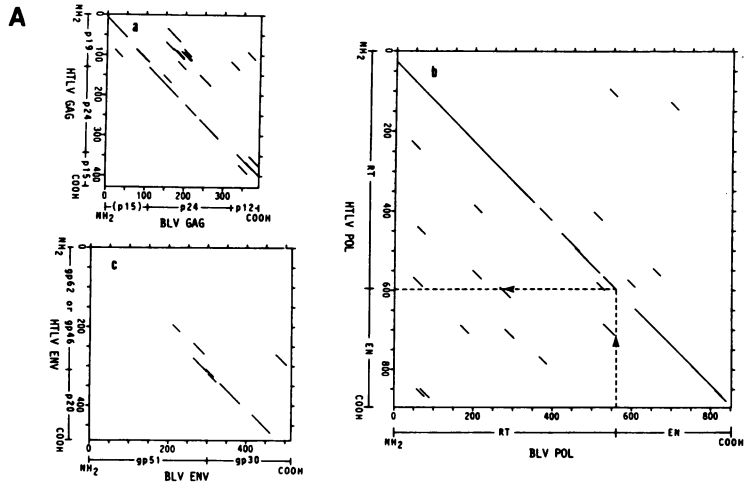
significant homology throughout their genomes, confirming at a complete nucleotide sequence level that BLV is most closely related to HTLV-I.

Comparison of the three structural genes of BLV and HTLV-I at an amino acid level

Figure 2A shows two-dimensional homology matrix comparisons of the three structural genes (*gag, pol* and *env*) of BLV and HTLV-I at an amino acid level, and Figure 2B shows their sequence alignments. They demonstrate very clearly the striking similarities in the gene structures of these two retroviruses.

gag gene. The BLV *gag* gene encodes a 392-amino acid residue precursor polyprotein and is predicted to have a sequence, NH_2-p15-p24-p12-COOH (Sagata *et al.*, 1984b), where p15, p24 and p12 are mature *gag* proteins representing phosphorylated, major internal and nucleic acid-binding proteins, respectively (Burny *et al.*, 1980; Copeland *et al.*, 1983a). The HTLV-I *gag* gene encodes a 429-residue precursor polyprotein with a sequence, NH_2-p19-p24-15-COOH (Seiki *et al.*, 1983; Yoshida, 1983), showing close structural similarity to the BLV *gag* precursor sequence. As represented by the diagonal lines in the homology matrix (Figure 2A-a), the BLV and HTLV-I *gag* precursor polyproteins show strong homology encompassing almost the entire region of the three mature protein portions. Sequence alignment shows that they have the strongest homology (62%, counting gaps) in the nucleic acid-binding protein (BLV p12 and HTLV-I p15) portion and less but still substantial homology (34%) in the other protein portions (Figure 2B-a). As previously mentioned (Copeland *et al.*, 1983b; Sagata *et al.*, 1984b), the extremely high homology between the nucleic acid-binding proteins is due to the existence of highly homologous sequences duplicated in both the BLV p12 and HTLV-I p15 (see the brackets in Figure 2B-a); interestingly, a similar sequence is also duplicated in the avian retrovirus homologue and is thought to have a nucleic acid-binding activity (Copeland *et al.*, 1983a). BLV p15 and HTLV-I p19, on the other hand, have strongly conserved sequences in the N-terminal region (amino acid position 20–30, Figure 2B-a), and the major internal core proteins (p24s) have these in both the N-terminal (positions 150–190) and middle (positions 240–300) regions. We presume that these highly homologous regions are functional domains of the respective proteins.

pol gene. The retroviral *pol* gene encodes reverse transcriptase and possibly endonuclease also (Dickson *et al.*, 1982). The predicted *pol* products of BLV and HTLV-I consist of 852 and 896 amino acid residues, respectivley (Seiki *et al.*, 1983; Sagata *et al.*, 1984b). We showed earlier that the C-terminal one-third (approximately at amino acid position 560 to the C terminus; see Figure 2A-b) of the BLV *pol* product represents an endonuclease domain, while the major N-terminal two-thirds represents a reverse transcriptase domain (Sagata *et al.*, 1984b). As shown in the homology matrix (Figure 2A-b), BLV and HTLV-I share substantial homology throughout their *pol* products. This in turn indicates that the HTLV-I *pol* gene also encodes, in addition to the reverse transcriptase, an



B

a GAG

```
                 ┌►(p15)                                                                                                  120
BLV  MGN------SPSYNPPAGISPSDWLNLLQSAQRLNPRPSPSDFTDLKNYIHW-FHKTQKKPWTFTSGGPTSCPPGRFGRVPLVLATL---NEVLSNEGGAPGASAPEEQPPPYDP-----
HTLV MGQIFSRSASPIPRPPRGLAAHHWLNFLQAAYRLEPGPSSYDFHQLKKFLKIALETPARICPINYSLLASLLPKGYPGRVNEILHILIQTQAQIPSRPAPPPPSSPTHDPPDSDPQIPPP
     └►p19 (p15)◄──►p24                                                                                                240
     ------PAILPIISE-GNRNRHRAWALRELQDIKKEIENKAPGSQVWIQTLRLAILQADPTPADLEQLCQYIASPVDQTAHMTSLTAAIAAAEAATPSRVLTPKTGTLTQQSAQPNAGDL
     YVEPTAPQVLPVMHPHGAPPNHRPWQMKDLQAIKQEVSQAAPGSPQFMQTIRLAVQQFDPTAKDLQDLLQYLCSSLVASLHHQQLDSLISEAETRGIT-GYNPLAGPLRVQANNPQQQGL
          p19◄──►p24                                                                                    p24◄──►p12  360
     RSQYQNLWLQAGKISLLVLQLQP-WSTIVQGPAESSVEFVNRLQISLADNLPDGVLRNPLLTPLVMQMLTESVSKFCRGEASGRGGAK-TAGLRTIG--PPRMKQPALLVHTPGPKMPGP
     RREYQQLWL-AAFAALPGSAKDPSWASILQGLEEPYHAFVERLNIALDNGLPEGTPKDPILRSLAYSNANKECQKL--LQARGHTNSPLGDMLRACQTWTPKDK-TKVLVVQ--------
                                                                                  448                  p24◄──►p15
     RQPAPKR[PPP-GPCYRCLKEGHWARDCPT]KATG[PPPGPCPICKDPSHWKRDCPTLKSKN
                                    p12◄┘
     ----PKK[PPPNQPCFRCGKAGHWSRDC-TQPR-[PPPGPCPLCQDPTHWKRDCPRLKPTIPEPEPEEDALLLDLPADIPHPKNSIGGEV
                                                                                              p15◄┘
```

b POL

```
BLV                  ┌►RT                                                                                                120
                     GASIPFKLERLQALQDLVHRSLEAGYISPWDGPGNNPVFPVRKPNGAWRFVHDLRATNALTKPIPALSPGPPDLTAIPTHPPHIICLDLKDAFF
HTLV GKKAACNLANTGASLPWARTPPKAPRNQPVPFKPERLQALQHLVRKALEAGHIEPYTGPGNNPVFPVKKANGTWRFIHDLRATNSLTIDLSSSSPGPPDLSSLPTTLAHLQTIDLRDAFF
     └►RT                                                                                                              240
     QIPVEDRFRSYLSFTLPSPGGLQPHRRFAWRVLPQGFINSPALFERALQEPLRQVSAAFSQSLLVSYMDDILYASPTEEQRSQCYQALAARLRDLGFQVASEKTSQTPSPVPFLGQMVHE
     QIPLPKQFQPYFAFTVPQQCNYGPGTRYAWKVLPQGFKNSPTLFEMQLAHILQPIRQAFPQCTILQYMDDILLASPSHEDLLLLSEATMASLISHGLPVSENKTQQTPGTIKFLGQIISP
                                                                                                                       360
     QIVTYQSLPTLQISSPISLHQLQAVLGDLQWVSRGTPTTRRPLQLLYSSLKRHHDPRAIIQLSPEQLQGIAELRQALSHNARSRYNEQEPLLAYVHLTRAGSTLVLFQKGAQFPLAYFQT
     NHLTYDAVPTVPIRSRWALPELQALLGEIQWVSKGTPTLRQPLHSLYCALQRHTDPRDQIYLNPSQVQSLVQLRQALSQNCRSRLVQTLPLLGAIMLTLTGTTTVVFQSKEQWPLVWLHA
                                                                                                                       480
     PLTDNQASPWGLLLLLGCQYLQTQALSSYAKPILKYYHNLPKTSLDNWIQSSEDPRVQELLQLWPQISSQGIQPPGPWKT-LITRAEV----FLTPQF--SPDPIPAALCLFSDGATGRG
     PLPHTSQCPWGQLLASAVLLLDKYTLQSYGLLCQTIHHNISTQTFNQFIQTSDHPSVPILLHHSHRFKNLGAQTGELWNTFLKTAAPLAPVKALMPVFTLSPVIIINTAPCLFSDGSTSRA
                                                                                                          RT◄──►EN600
     AYCLWKDHLL--DFQAVPAP-ESAQKGELAGLLAGLAAAPP-EPVNIWVDSKYLYSLLRTLVLGAWLQPDPVPSYALLYKSLLRHPAIVVGHVRSHSSASHPIASLNNYVDQLL----PL
     AYILWDKQILSQRSFPLPPPHKSAQRAELLGLLHGLSSARSWRCLNIFLDSKYLYHYLRTLALGTFQGRSSQAPFQALLPRLLSRKVVYLHHVRSHTNLPDPISRLNALTDALLITPVLQ
                                                                                                          RT◄──►EN720
     ETPEQWHKLTHCNSRALSRWPNPRISAWDPRSPATLCETCQKLNPTGGGKMRTIQRGWAPNHIWQADITHYKYKQFTYALHVFVDTYSGATHASAKRGLTTQTTIEGLLEAIVHLGRPKK
     LSPAELHSFTHCGQTALTL---QGATTTEASNILRSCHACRGGNPQHQMPRGHIRRGLLPNHIWQGDITHFKYKNTLYRLHVWVDTFSGAISATQKRKETSSEAISSLLQAIAHLGKPSY
                                                                                                                       840
     LNTDQGANYTSKTFVRFCQQFGVSLSHHVPYNPTSSGLDERTNGLLKLLLSKYHLDEPHLPMTQALSRALWTHNQINLL-PILKTRWELHHSPPLAVISEGGETPKGSDKLFLYLLPGQN
     INTDNGPAYISQDFLNMCTSLAIRHTTHVPYNPTSSGLVERFNGILKTLLYKYFTDKPDLPMDNALSIALWTINHLNVLTNCHKTRWQLHHSPRLQPIPETRSLSNKQTHWYYFKLPGLN
                                                                                                          EN◄┘
     NRRWLGPLPALVEASGGALL--ATDPPVWVPWRLLKAFKCLKN-DGPEDAHNRS--SDG
     SRQWKGPQEALQEAAGAALIPVSASSAQWIPWRLLKRAACPRPVGGPADPKEKDLQHHG
                                                                                                          EN◄┘
```

c ENV

```
                                                                                  gp51◄──►gp30                         120
BLV  PGLALPDAQIFWVNSSSFNTTQGWHHPSQRLLFNVSQGNALLLPPISLVNLSTASSAPPTRVRRSPVAA-LTLGLALSVGLTGINVAVSALSHQRLTSLIHVLEQDQQRLITAINQTHY
HTLV PSLALPAPHLTL----PFNWTHCF-DPQIQAIVSSPCHNSLILPPFSLSPVPTLG----SRSRRAVPVAYWLVSALAMGAGVAGGITGSMSLASGK--SLLHEVDKDISQLTQAIVKNHK
                                                                                  gp62 or gp46◄──►p20                  240
     NLLNVASVVAQNRRGLDWLYIRLGFQSLCPTINEPCCFLRIQNDSIIRLGDLQPLSQRVSTDWQWPWNWDLGLTAWVRETIHSVLSLFLLALFLLFLAPCLIKCLTSRLLKLLRQAPHFP
     NLLKIAQYAAQNRRGLDLLFWEQG--GLCKALQEQCRFPNITNSHVPILQERPPLENRVLTGW--GLNWDLGLSQWAREALQTGITLVALLLLVILAGPCI----------LRQLRHLP
                                                                                  gp30◄┘
     EISLTPKPDSDYQALLPSAPEIYSHLSPVKPDYINLRPCP
     S-----------------RVRYPHYSLIKPESSL
                      p20◄┘
```

Fig. 2. Two-dimensional homology matrix comparisons (**A**) and alignments (**B**) of the amino acid sequences of BLV and HTLV-I structural proteins. (**A**) Homology matrix comparisons of BLV and HTLV-I amino acid sequences according to the computer program detecting 20-residue unit sequences with an average 'score' value of at least 70 (Toh *et al.*, 1983). Location of the BLV *gag* p15 is tentatively identified (Sagata *et al.*, 1984b). The location of the HTLV-I reverse transcriptase (RT)-endonuclease (EN) junction is estimated (by the broken line with arrows) based on the previously determined one of BLV (Sagata *et al.*, 1984b). a, *gag*; b, *pol*; c, *env*. (**B**) Alignments of the BLV and HTLV-I amino acid sequences according to Needleman and Wunsh (1970). Amino acid (represented by one-letter code) sequences are deduced from nucleotide sequences (BLV from Sagata *et al.*, 1984b and HTLV-I from Seiki *et al.*, 1983). The HTLV-I *env* surface glycoprotein (gp62 or gp46)-transmembrane protein (p20) cleavage site is tentatively determined according to the consensus sequence recently proposed for the retroviral *env* product cleavage site (Sagata *et al.*, 1984b). See text for explanations of brackets, closed circles, underlines and boxes.

as yet unidentified endonuclease; the N terminus of this enzyme can be located approximately at amino acid position 600 of the *pol* product (estimation shown by the broken line with arrows in Figure 2A-b). It should be noted that the estimated mol. wt. (32 000 daltons) of the putative HTLV-I (and also BLV) endonuclease is identical to that of the avian retrovirus homologue (Grandgennet *et al.*, 1978, 1980).

Sequence alignment of the BLV and HTLV-I *pol* products can be performed without any extensive deletions or insertions (Figure 2B-b), and the overall homology is as high as 46%. We note the two highly conserved sequences: one in the N-terminal region of the reverse transcriptase (positions 31 – 85 in Figure 2B-b) having 75% homology, and the other in the C-terminal half of the putative endonuclease (positions 748 – 813) having 68% homology. Most probably, the former well conserved region represents an RNase H activity-carrying domain of the reverse transcriptase because it has strong homology with the N-terminal region of the avian retrovirus reverse transcriptase (Sagata *et al.*, 1984b), a region which is suggested to have an RNase H activity (Crouch and Dirksen, 1982). Both the BLV and HTLV-I reverse transcriptases possess most or all of the 10 amino acids (indicated by closed circles in Figures 2B-b) that are conserved at the corresponding positions of the other retrovirus reverse transcriptases (and the putative polymerases of hepatitis B and cauliflower mosaic viruses) (Toh *et al.*, 1983).

env gene. The primary product of the retroviral *env* gene undergoes processing to make a mature surface glycoprotein and a transmembrane protein (Dickson *et al.* 1982). The mature BLV *env* products are glycoproteins: gp51 (surface glycoprotein) and gp30 (transmembrane protein), with the translational order of NH_2-gp51-gp30-COOH (Burny *et al.*, 1980; Sagata *et al.*, 1984b). The HTLV-I *env* gene encodes a glycoprotein gp62 and a processed small protein p20 (a putative transmembrane protein) (Hattori *et al.*, 1983), but it is not known whether a mature HTLV-I surface glycoprotein is represented by the gp62 molecule or another glycoprotein found in HTLV-I-infected cells (i.e., gp46: Yamamoto *et al.*, 1982b; Lee *et al.*, 1984). Homology matrix comparison of the BLV and HTLV-I *env* products, each deduced from the nucleotide sequence (Seiki *et al.*, 1983; Sagata *et al.*, 1984b), reveals clearly that they share strong homology in the transmembrane protein portion but do not in almost the entire region of the surface glycoprotein portion (Figure 2A-c). The stronger conservation of the transmembrane protein has been noted between related retroviruses (Elder and Mullins, 1983; Sagata *et al.*, 1984b). The virtual absence of homology in the surface glycoprotein portion between BLV and HTLV-I probably reflects a difference in the host cell range of these retroviruses because the surface glycoprotein molecule is a host range determinant (Dickson *et al.*, 1982).

Sequence comparison of the BLV and HTLV-I *env* products around their transmembrane protein portions reveals striking structural similarities (Figure 2B-c). First, they both have three hydrophobic stretches at corresponding positions (at the C terminus of the surface glycoprotein and at the N terminus and near the C terminus of the transmembrane protein, each underlined in Figure 2B-c). Second, they share two common cystein residues spaced eight amino acids apart (at positions 150 and 157, indicated by closed circles in the figure) in the central domain of the transmembrane protein. And third, very near the conserved cystein residues, they both have a completely matched 8-residue sequence rich in hydrophilic residues (boxed in the figure). We note that all of these structural features are also at corresponding positions of the other (murine, feline and avian) retrovirus *env* products (Dickson *et al.*, 1982; Lenz *et al.*, 1982; Elder and Mullins, 1983; Sagata *et al.*, 1984), strongly suggesting their functional importance. In fact, the first two features may play a key role in formation of a surface glycoprotein-transmembrane protein complex and its membrane association (Dickson *et al.*, 1982; Lenz *et al.*, 1982).

The major difference in the BLV and HTLV-I transmembrane proteins occurs in the C-terminal region, where HTLV-I has large deletions totalling 29 amino acid residues (Figure 2B-c). The murine leukemia virus which shows stronger transmembrane protein homology to HTLV-I than to BLV, has similar large deletions in corresponding positions (our unpublished observation). We suggest therefore that the BLV transmembrane protein, which carries C-terminal extra sequences, interacts with the cell membrane in a somewhat different way than do those of other mammalian retroviruses.

Characterization of the pX_{BL} region

Between the *env* gene and the 3′ LTR, the BLV genome harbors a 1800-bp unidentified region (designated pX_{BL}), which corresponds to a pX region of the HTLV-I genome (Sagata *et al.*, 1984b). This pX_{BL} region contains six open reading frames that are capable of encoding at least 80 amino acid residues (Figure 3).

Protein-coding property of the pX_{BL} region. To assess whether the pX_{BL} region has any protein-coding property, we sequenced the pX_{BL} region of another BLV variant, a Belgian isolate (Deschamps *et al.*, 1981). Our rationale for the assessment was that if any open reading frame found in the pX_{BL} region of our original BLV isolate (Japanese isolate; Sagata *et al.*, 1983) actually encodes a protein, then the same open reading frame would also be found in the Belgian isolate, which (because of the synonymous changes of codons between the two isolates) would carry point mutations (base substitutions) preferentially at the third bases of its codons. Figure 3 shows a comparison of the pX_{BL} sequences of the two BLV isolates. In the 1800-bp sequence of the Belgian isolate 60-base substitutions and a single-base insertion can be seen, but none of these point mutations disrupt the open reading frames in our isolate. Although small open reading frames (pX_{BL}-II ~ pX_{BL}-VI) have too few mutations (at most 10) to assess their protein-coding properties, the longest open reading frame, pX_{BL}-I, has 34 point mutations and as many as 20 of these occur at the third bases, resulting in 16 synonymous changes out of the 29 affected codons. These results suggest that at least the pX_{BL}-I has a protein-coding property. This open reading frame could encode a 308-amino acid residue polypeptide with a mol. wt. of 34 200 daltons. It is interesting to note that the potential product has four periodically placed hydrophobic or apolar stretches of 15 – 18 residues (underlined in Figure 3) and two potential N-asparagine-linked glycosylation sites (indicated by CHOs in Figure 3), therefore it may be a (membrane) glycoprotein. If this polypeptide is actually glycosylated, its molecular size would amount to ~40 000 daltons.

Homology between the potential pX_{BL} and pX products

HTLV pX also contains several potential open reading frames (Seiki *et al.* 1983). The longest, pX-IV, together with the still open sequence upstream of the ATG triplet of the pX-IV frame, could encode a polypeptide 357 amino acids

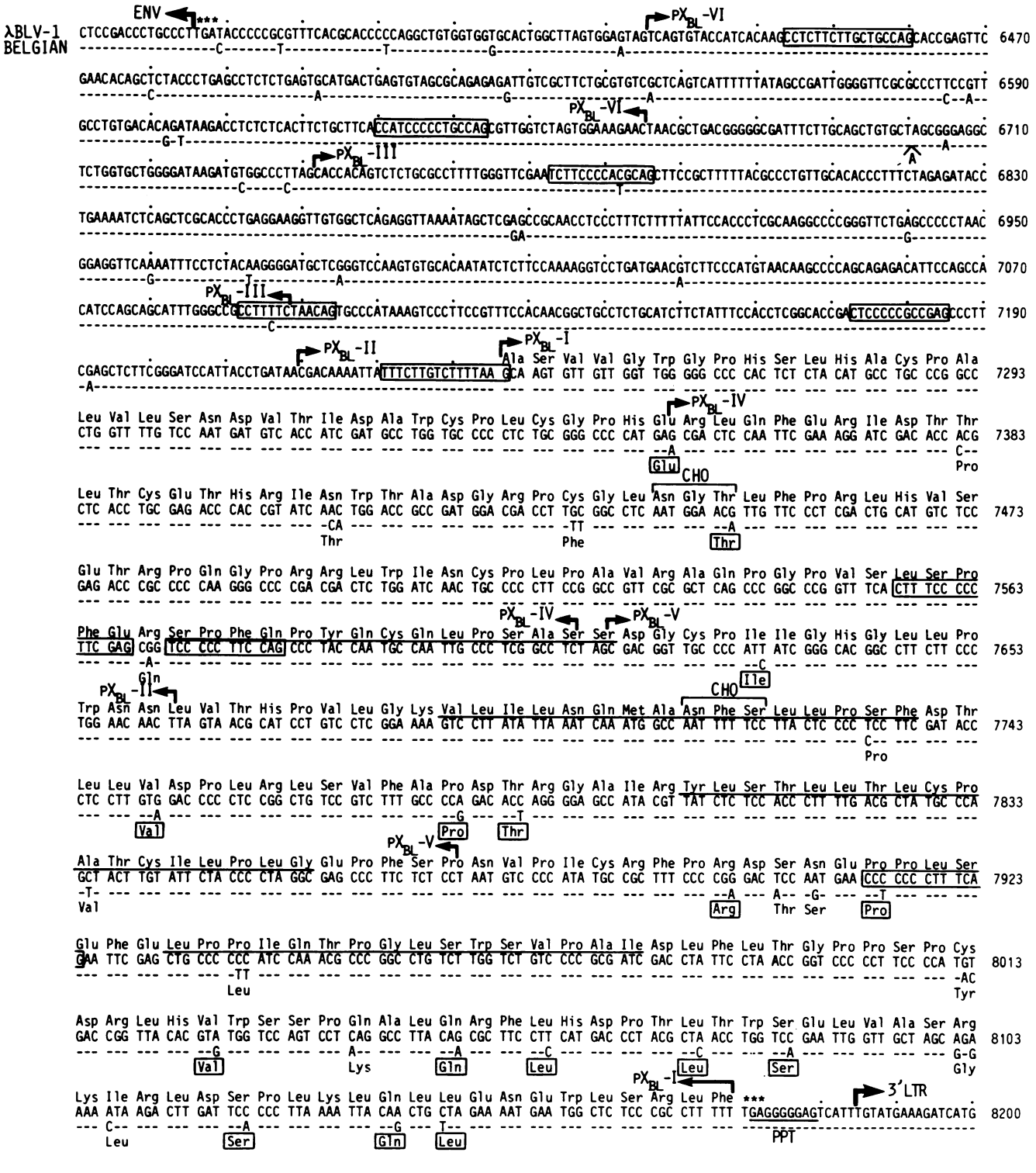


Fig. 3. Comparison of the pX_{BL} sequences between the two BLV isolates. The nucleotide sequences of the pX_{BL} regions of the Japanese BLV isolate (upper line; Sagata *et al.*, 1984b) and Belgian BLV isolate (lower line; determined in this paper) are compared; only the altered nucleotides are shown for the Belgian isolate (dashes, unaltered nucleotides). The Belgian isolate has a single base insertion between positions 6700 and 6701, the nucleotide numbering being according to the Japanese isolate (Sagata *et al.*, 1984b). pX_{BL}-I ~ pX_{BL}-VI are potential open reading frames, and only the pX_{BL}-I frame is translated into amino acids; boxed amino acid residues of the Belgian isolate represent synonymous changes. Nucleotide sequences representing potential splice acceptor sites are boxed. Apolar or mostly hydrophobic stretches of amino acids are underlined. CHO, potential glycosylation site; PPT, a polypurine tract as a primer for plus-strand DNA synthesis (Sagata *et al.*, 1984a).

long and corresponds, in both size and position, to the pX_{BL}-I of BLV. The two-dimensional homology matrix under a less stringent condition than that in Figure 2A, revealed that the potential pX_{BL}-I and pX-IV products have appreciable homology, especially in the N-terminal quarter (homology matrix omitted). Sequence alignment of their 80-residue N-terminal sequences reveals that they have as high as 35% homology and contain several highly conserved sequences (Figure 4A), indicating the significance of the homology observed. We also found low but still appreciable

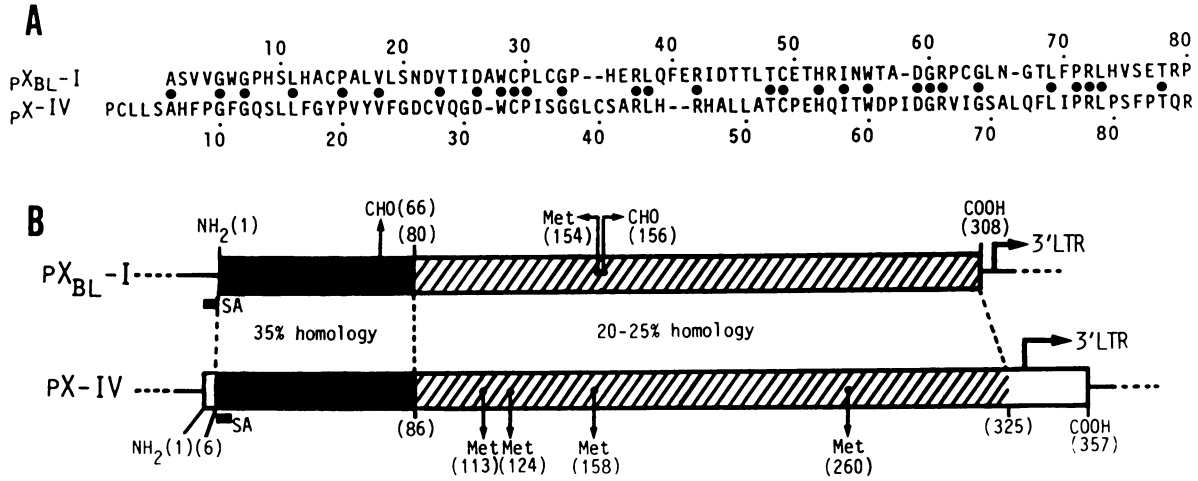


Fig. 4. Sequence alignment of the N-terminal quarter of the potential pX_{BL}-I and pX-IV products (**A**) and comparison of their entire structures (**B**). (**A**) The N-terminal sequence (80 residues) of BLV pX_{BL}-I is from Figure 3, while that of HTLV-I pX-IV (86 residues) is from the still open sequence upstream of the ATG triplet (methionine residue at position 113 in **B**) of the 'original' pX-IV frame (Seiki *et al.*, 1983). Closed circles indicate amino acid matches. (**B**) Closed boxes, showing 35% homology, correspond to the N-terminal sequences shown in **A**. Hatched boxes indicate less homologous regions (20–25%). Internal methionine (Met) residues are indicated. Numbers in parentheses are amino acid positions. CHO, potential glycosylation site; SA, potential splice acceptor site.

homology (20–25%) in the C-terminal three-fourths of the two products, although their overall sequence alignment could not be performed with great accuracy (and thus is omitted here). Based on these observations, a structural comparison of the potential pX_{BL}-I and pX-IV products is shown in Figure 4B. The C terminus of the pX_{BL}-I product is truncated compared with the pX-IV product, because of its premature termination immediately before the 3′ LTR. In contrast to the pX_{BL}-I product, the pX-IV product has no potential glycosylation site, which may suggest a different subcellular localization between the two products. Neither of these potential products has an initiator methionine, athough they contain an internal methionine residue(s); however, they probably begin with the 5′ ends of the respective open reading frames because they have stronger homology in the N-terminal quarter that precedes the internal methionine residue(s). In this respect, we suggest that each of these products is generated by a spliced mRNA because one of the several potential splice acceptor sites found in the pX_{BL} (see Figure 3) and pX regions occurs at the very 5′ ends of the respective open reading frames (Figure 4B). Supporting this suggestion, the pX region of HTLV type II, a retrovirus implicated in hairy cell leukemia and closely related to HTLV-I, contains a long open reading frame that is nearly identical to the HTLV-I pX-IV and has a potential splice acceptor site at its 5′ end (Haseltine *et al.*, 1984).

Origin of the $pX_{(BL)}$ sequence?

The low but significant homology observed between the putative pX_{BL} and pX products suggests that the pX_{BL} and pX sequences have diverged from a common ancestral gene. This is consistent with our recent proposal that the $pX_{(BL)}$ sequence was somehow acquired by the common progenitor of BLV and HTLV (Sagata *et al.*, 1984b). The pX sequence of HTLV has been reported to be of rodent origin (Fukui *et al.*, 1983). However, we think it difficult to experimentally determine where the $pX_{(BL)}$ sequence came from, because the sequences of pX_{BL} and pX have diverged greatly from each other (overall, <35% nucleotide homology), and also from their possible common ancestral (cellular?) sequence. In some support of this, we no longer found significant difference in the codon usage between the putative pX_{BL} product and the other structural proteins of BLV, nor did we find any significant homology between the pX_{BL} product and the >2500 known prokaryotic and eukaryotic gene products, including many oncogene products and growth factors (our unpublished observations).

Materials and methods

BLV DNA and sequence determination

The isolation and DNA sequencing of a molecular clone (λBLV-1; Japanese isolate) of BLV provirus was described (Sagata *et al.*, 1983, 1984a, 1984b). The subclone (clone 905-20) of the Belgian BLV isolate (Deschamps *et al.*, 1981) was the gift of Drs. R. Kettmann and A. Burny. The pX_{BL} region (3′ end portion) of this clone has a restriction map similar to our BLV isolate (λBLV-1) (Sagata *et al.*, 1983) and its DNA sequencing was done as previously (Sagata *et al.*, 1984b), according to the Maxam-Gilbert procedures (Maxam and Gilbert, 1980) using [γ-^{32}P]ATP (Amersham; 3000 Ci/mmol) and polynucleotide kinase (Takara-Shuzo, Kyoto, Japan).

Computer-assisted analyses of the nucleotide and amino acid sequences

Two-dimensional homology matrix comparisons of the amino acid sequences were performed by a computer program developed for distantly related proteins (Toh *et al.*, 1983). A DNA version of this method was adopted for the two-dimensional homology matrix comparison of the nucleotide sequences. Alignment of the amino acid sequences were performed according to Needleman and Wunsh (1970), checked by hand and corrected if necessary.

Acknowledgements

We thank Drs. R. Kettman and A. Burny, Brussels, Belgium, for the kind gift of the Belgian BLV clone. We also thank Dr. T. Miyata, Kyushu University, Japan, for help and critical discussions on the computer search for protein homologies and Mrs. J. Clarkin for reading and typing the manuscript. This work was supported in part by grants from the Ministry of Education, Science and Culture of Japan and the Tokushima Research Institute of Otuka Pharmaceutical Co., Ltd.

References

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 Klein,G. (ed.), *Viral Oncology*, Raven Press, NY, pp. 231-289.

Coffin,J. (1982) in Weiss,R., Teich,N., Varmus,H. and Coffin,J. (eds.), *RNA Tumor Viruses*, Cold Spring Harbor Laboratory Press, NY, pp. 261-368.

Copeland,T.D., Morgan,M.A. and Oroszlan,S. (1983a) *FEBS Lett.*, **156**, 37-40.

Copeland,T.D., Oroszlan,S., Kalyanaraman,V.S., Sarngadharan,M.G. and Gallo,R.C. (1983b) *FEBS Lett.*, **162**, 390-395.
Crouch,R.J. and Dirksen,M.L. (1982) in Linn,S. and Roberts,R. (eds.), *Nucleases*, Cold Spring Harbor Laboratory Press, NY, pp. 211-241.
Deschamps,J., Kettmann,R. and Burny,A. (1981) *J. Virol.*, **40**, 605-609.
Dickson,C., Eisenman,R., Fan,H., Hunter,E. and Teich,N. (1982) in Weiss, R., Teich,N., Varmus,H. and Coffin,J. (eds.), *RNA Tumor Viruses*, Cold Spring Harbor Laboratory Press, NY, pp. 513-648.
Elder,J.H. and Mullins,J.I. (1983) *J. Virol.*, **46**, 871-880.
Fukui,K., Noma,T., Takeuchi,K., Kobayashi,N., Hatanaka,M. and Honjo, T. (1983) *Mol. Biol. Med.*, **1**, 447-456.
Grandgenett,D.P., Vora,A.C. and Schiff,R.D. (1978) *Virology*, **89**, 119-132.
Grandgenett,D.P., Golomb,M. and Vora,A.C. (1980) *J. Virol.*, **3**, 264-271.
Haseltine,W.A., Sodroski,J., Patarca,R., Briggs,D., Perkins,D. and Wong-Staal,F. (1984) *Science (Wash.)*, **225**, 419-421.
Hattori,S., Imagawa,K., Shimizu,F., Hashimura,E., Seiki,M. and Yoshida, M. (1983) *Gann*, **74**, 790-793.
Kettmann,R., Deschamps,J., Cleuter,Y., Couez,D., Burny,A. and Marbaix, G. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 2465-2469.
Kettmann,R., Deschamps,J., Couez,D., Claustriaux,J.-J., Palm,R. and Burny,A. (1983) *J. Virol.*, **47**, 146-150.
Lee,T.H., Coligan,J.E., Homma,T., McLane,M.F., Tachibana,N. and Essex,M. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 3856-3860.
Lenz,J., Crowther,R., Straceski,A. and Haseltine,W. (1982) *J. Virol.*, **42**, 519-529.
Maxam,A. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499-560.
Miyoshi,I., Taguchi,H., Fujishita,M., Yoshimoto,S., Kubonishi,I., Shiraishi, Y. and Akagi,T. (1982) *Lancet*, **i**, 1016.
Miyoshi,I., Yoshimoto,S., Taguchi,H., Kubonishi,I., Fujishita,M., Ohtsuki, Y., Shiraishi,Y. and Akagi,T. (1983) *Gann*, **74**, 1-4.
Needleman,S.B. and Wunsh,C.D. (1970) *J. Mol. Biol.*, **48**, 443-453.
Onuma,M., Watarai,S., Suneya,M., Mikami,T. and Izawa,H. (1981) *Microbiol. Immunol.*, **25**, 445-454.
Oroszlan,S., Sarngadharan,M.G., Copeland,T.D., Kalyanaraman,V.S., Gilden,R.V. and Gallo,R.C. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 1291-1294.
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.
Rhim,J.S., Kraus,M. and Arnstein,P. (1983) *Int. J. Cancer*, **31**, 791-795.
Sagata,N., Ogawa,Y., Kawamura,J., Onuma,M., Izawa,H. and Ikawa,Y. (1983) *Gene*, **26**, 1-10.
Sagata,N., Yasunaga,T., Ogawa,Y., Tsuzuku-Kawamura,J. and Ikawa,Y. (1984a) *Proc. Natl. Acad. Sci. USA*, **81**, 4741-4745.
Sagata,N., Yasunaga,T., Tsuzuku-Kawamura,J., Ohishi,K., Ogawa,Y. and Ikawa,Y. (1984b) *Proc. Natl. Acad. Sci. USA*, in press.
Seiki,M., Hattori,S., Hirayama,Y. and Yoshida,M. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3618-3622.
Seiki,M., Eddy,R., Shows,T.B. and Yoshida,M. (1984) *Nature*, **309**, 640-642.
Teich,N, Wyke,J., Mak,T., Bernstein,A. and Hardy,W. (1982) in Weiss,R., Teich,N., Varmus,H. and Coffin,J. (eds.), *RNA Tumor Viruses*, Cold Spring Harbor Laboratory Press, NY, pp. 785-998.
Toh,H., Hayashida,H. and Miyata,T. (1983) *Nature*, **305**, 827-829.
Yamamoto,N., Okada,M., Koyanagi,Y., Kannagi,M. and Hinuma,Y. (1982a) *Science (Wash.)*, **217**, 737-739.
Yamamoto,N., Schneider,J., Hinuma,Y. and Hunsmann,G. (1982b) *Z. Naturforsch.*, **37c**, 731-732.
Yoshida,M. (1983) *Gann*, **74**, 777-789.
Yoshida,M., Miyoshi,I. and Hinuma,Y. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 2031-2035.
Weiss,R. (1982) in Weiss,R., Teich,N., Varmus,H. and Coffin,J. (eds.), *RNA Tumor Viruses*, Cold Spring Harbor Laboratory Press, NY, pp. 1205-1281.

Received on 12 September 1984

Note added in proof

The *gag-pol* junction which was commonly found in BLV and HTLV-1 genomes without apparent homology (Figure 1) was recently found to encode a protease for cleavage of a gag precursor polyprotein (N.Sagata *et al.*, *FEBS Lett.*, December 1984).