*Proc. Natl. Acad. Sci. USA*
Vol. 82, pp. 677–681, February 1985
Biochemistry

# Complete nucleotide sequence of the genome of bovine leukemia virus: Its evolutionary relationship to other retroviruses

**(splice donor site/endonuclease/unidentified reading frames/type "E")**

NORIYUKI SAGATA*, TERUO YASUNAGA†, JUNKO TSUZUKU-KAWAMURA*, KAZUE OHISHI*, YAŠUKI OGAWA‡, AND YOJI IKAWA*

*Laboratory of Molecular Oncology and †Computation Center, The Institute of Physical and Chemical Research, Wako, Saitama 351, Japan; and ‡Department of Epizootiology, Faculty of Veterinary Medicine, Hokkaido University, Sapporo, Hokkaido 060, Japan

*Communicated by Hitoshi Kihara, September 17, 1984*

ABSTRACT **We report the complete 8714-nucleotide sequence of the integrated bovine leukemia virus genome and deduce the following genomic organization: 5′ LTR-*gag-pol-env*-$pX_{BL}$-3′ LTR, where LTR represents a long terminal repeat and $pX_{BL}$ represents a region containing unidentified open reading frames. This genomic structure is similar to that of human T-cell leukemia virus. The LTR contains a putative splice donor site in the R region. The *gag* gene encodes a precursor protein with the form $NH_2$-p15-p24-p12-COOH. The $NH_2$- and COOH-terminal regions of the *pol* product show stronger homologies with those of avian, rather than murine, type C retrovirus, and its structure is identical to that of avian virus. The *env* gene encodes a surface glycoprotein (gp51) and a transmembrane protein (gp30). In contrast to the *pol* product, the gp30 shows stronger sequence homology with a murine, rather than avian homologue, indicating the chimeric nature of the bovine leukemia virus genome. Comparisons of the best conserved *pol* sequences and overall genomic organizations between several major oncoviruses allow us to propose that bovine leukemia and human T-cell leukemia viruses constitute a group, designated as type "E," of Oncovirinae.**

Bovine leukemia virus (BLV) is an exogenous replication-competent virus and is classified as a member of type C retrovirus (1). However, it differs from the major mammalian type C viruses in several aspects (1, 2). Recent biochemical (3) and molecular biological (4) studies suggest that BLV is most closely related to human T-cell leukemia virus (HTLV), which is also classified as a type C virus (5). In contrast to HTLV, little is known about the genomic structure of BLV. We report here the complete nucleotide sequence of BLV and propose that BLV and HTLV belong to another group of Oncovirinae, designated here as type "E."

## MATERIALS AND METHODS

**DNA Sequence Analysis.** The restriction map of a BLV clone (λBLV-1) was described (6). Each restriction site was labeled using [$\gamma$-$^{32}$P]ATP (Amersham; 3000 Ci/mmol; 1 Ci = 37 GBq) and polynucleotide kinase (Takara-Shuzo, Kyoto, Japan). According to the Maxam–Gilbert procedures (7), 82% of the BLV genome was sequenced in both DNA strands. All the restriction sites were read through.

**Computer-Assisted Analysis of the Deduced Amino Acid Sequences.** Sequence homology was examined by two-dimensional homology matrix (8). Sequence alignment was according to Needleman and Wunsh (9).

## RESULTS AND DISCUSSION

**Complete Nucleotide Sequence and Structure of BLV Genome.** Fig. 1 shows the complete 8714-nucleotide sequence of an integrated BLV genome. It can be anatomized in the form of 5′ LTR-*gag-pol-env*-$pX_{BL}$-3′ LTR, where $pX_{BL}$ represents an unidentified region and LTR represents a long terminal repeat.

**LTR and 5′ Leader Sequence.** The nucleotide sequence of the LTR [530 base pairs (bp)] was described (4). The 5′ leader sequence following the 5′ LTR and preceding the *gag* gene consists of 97 bp. Moloney murine leukemia virus (Mo-MuLV) contains a splice donor sequence for *env* mRNA in this region (10), while Rous avian sarcoma virus (RSV) contains this at the $NH_2$ terminus of the *gag* gene (11). BLV has no such sequence (consensus sequence, $^{C}_{A}$-A-G-G-T-$^{A}_{G}$-A-G-T; see ref. 12) in these regions. Inspection of the entire BLV sequence, however, reveals that the sequence C-A-G-G-T-A-A-G-G (8/9 match of the consensus) appears just once in only the long R region of the LTR (positions 303–311; Fig. 1). Inspection of the LTR sequences of two different HTLVs also reveals almost identical sequences in the R regions [T-A-G-G-T-A-A-G-T (8/9 match of the consensus) for HTLV-I (13) and A-A-G-G-T-A-A-G-T (9/9 match) for HTLV-II (14)]. The R regions of these retrovirus LTRs are much longer than those of other retroviruses and probably are implicated in transcription termination (4, 13). We suggest here that they are also involved in the splicing events.

***gag* Gene.** The first open reading frame, the *gag* gene, spans nucleotides 628–1806, beginning with the first ATG triplet appearing downstream of the 5′ LTR (Fig. 1). Its deduced amino acid sequence contains a reported 50-residue $NH_2$-terminal sequence of the major internal *gag* protein p24 (15) and a complete 69-residue sequence of the nucleic acid-binding protein p12 (16) at positions 955–1119 and 1597–1803, respectively. p24 presumably ends with the leucine residue (a COOH terminus of p24; see ref. 16) that immediately precedes p12, because the estimated molecular size (consisting of 214 residues) approximates 24 kDa. Upstream of the p24, there are 109 amino acid residues. This region probably corresponds to p15, which is a phosphorylated basic *gag* product (1). Thus, we propose that the BLV *gag* precursor protein (Pr45$^{gag}$; see ref. 1) has the sequence $NH_2$-p15-p24-p12-COOH. The nucleic acid-binding protein p12

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: BLV, bovine leukemia virus; bp, base pair(s); HTLV, human T-cell leukemia virus; LTR, long terminal repeat; Mo-MuLV, Moloney murine leukemia virus; RSV, Rous sarcoma virus.

```
►5'LTR,U3                                                Pvu II                        Pvu II
TGTATGAAAGATCATGCCGACCTAGGAGCCGCCACCGCCCCGTAAACCAGACAGAGACGTCAGCTGCCAGAAAAGCTGGTGACGGCAGCTGGTGGCTAGAATCCCCGTACCTCCCCAACTTCCCCTTTCCCGAAAATCCACACCCTGAG  150
                                                           U3◄►R                                               SacI CAT box
CTGCTGACCTCACCTGCTGATAAATTAATAAAATGCCGGCCCTGTCGAGTTAGCGGCACCAGAAGCGTTCTTCTCCTGAGACCCTCGTGCTCAGCTCTCGGTCCTGAGCTCTCTTGCTCCCGAGACCTTCTGGTCGGCTATCCGGCAGCG  300
                  TATA box Poly(A) signal          Sac I  Cap site                                                                                   R◄►U5
GTCAGGTAAGGCAAGCACGGTTTGGAGGGTGGTTCTCGGCTGAGACCACCGCGAGCTCTATCTCCGGTCCTCTGACCGTCTCCACGTGGACTCTCTCCTTTGCCTCCTGACCCCGCGCTCCAAGGGCGTCTGGCTTGCACCCGCGTTTGT  450
 Splice donor                                                           U5,5'LTR◄                                                          Poly(A) site
TTCCTGTCTTACTTTCTGTTTCTCGCGGCCCGCGCTCTCTCCTTCGGCGCCCTCTAGCGGCCAGGAGAGACCGGCAAACAATTGGGGGCTCGTCCGGGATTGATCACCCCGGAACCCTAACAACTCTCTGGACCCACCCCCTCGGCGGCA  600
                           ►GAG,(p15)                                                 Primer binding site
TTTTGGGTCTCTCCTTCAAATTATATCATGGGAAATTCCCCCTCCTATAACCCCCCCGCTGGTATCTCCCCCTCAGACTGGCTCAACCTTCTGCAAAGCGCGCAAAGGCTCAATCCGCGACCCTCTCCTAGCGATTTTACCGATTTAAAG  750
                           MetGlyAsnSerProSerThrAsnProProAlaGlyIleSerProSerAspTrpLeuAsnLeuLeuGlnSerAlaGlnArgLeuAsnProArgProSerProSerAspPheThrAspLeuLys

AATTACATCCATTGGTTTCATAAGACCCAGAAAAAACCATGGACTTTCACTTCTGGTGGCCCCACCTCATGTCCACCCGGGAGATTCGGCCGGGTTCCCCTCGTCTTGGCCACCCTAAACGAAGTACTCTCAAACGAAGGGGGCGCCCCG  900
AsnThrIleHisTrpPheHisLysThrGlnLysLysProTrpThrPheThrSerGlyGlyProThrSerCysProProGlyArgPheGlyArgValProLeuValLeuAlaThrLeuAsnGluValLeuSerAsnGluGlyGlyAlaPro
                                                    (p15)◄►p24
GGTGCATCGGCCCCAGAAGAACAACCCCCCCCTTATGACCCCCCCGCCATTTTGCCAATCATATCTGAAGGGAATCGCAACCGCCATCGTGCTTGGGCACTCCGAGAATTACAAGATATCAAAAAAGAAATTGAAAATAAGGCACCGGGT  1050
GlyAlaSerAlaProGluGluGlnProProProTyrAspProProAlaIleLeuProIleIleSerGluGlyAsnArgAsnArgHisArgAlaTrpAlaLeuArgGluLeuGlnAspIleLysLysGluIleGluAsnLysAlaProGly
                                            Pst I                                                 Sal I
TCGCAAGTATGGATACAAACACTACGACTTGCAATCCTGCAGGCCGACCCTACTCCGGCTGACCTAGAACAACTTTGCCAATATATTGCTTCCCCGGTCGACCAAACGGCCCATATGACCAGCCTAACGGCAGCAATAGCCGCCGCTGAA  1200
SerGlnValTrpIleGlnThrLeuArgLeuAlaIleLeuGlnAlaAspProThrProAlaAspLeuGluGlnLeuCysGlnTyrIleAlaSerProValAspGlnThrAlaHisMetThrSerLeuThrAlaAlaIleAlaAlaAlaGlu

GCGGCAACACCCTCCAGGGTTTTAACCCCCAAAACGGGTACCCTAACCCAACAATCAGCTCAGCCCAACGCCGGGGATCTTAGAAGTCAATATCAAAACCTCTGGCTTCAGGCCGGAAAAATCTCCCTACTCGTCCTTCAGCTACAACCT  1350
AlaAlaThrProSerArgValLeuThrProLysThrGlyThrLeuThrGlnGlnSerAlaGlnProAsnAlaGlyAspLeuArgSerGlnTyrGlnAsnLeuTrpLeuGlnAlaGlyLysIleSerLeuLeuValLeuGlnLeuGlnPro
                                              Hinc II
TGGTCCACCATCGTCCAAGGCCCCGCCGAAAGCTCTGTAGAGTTTGTCAACCGGTTACAAATTTCATTAGCTGACAACCTTCCCGACGGAGTCCTAAGGAACCCATTATTGACTCCCTTAGTTATGCAAATGCTAACAGAGAGTGTCAGC  1500
TrpSerThrIleValGlnGlyProAlaGluSerSerValGluPheValAsnArgLeuGlnIleSerLeuAlaAspAsnLeuProAspGlyValLeuArgAsnProLeuLeuThrProLeuValMetGlnMetLeuThrGluSerValSer
                                              Pst I                                      .p24◄►p12
AAATTTTGCAGGGGCGAGGCCAGTGGCCGCGGTGGGGCAAAAACTGCAGGCTTGCGCACAATTGGGCCCCCAAGAATGAAACAGCCTGCACTTCTCGTCCACACCCCAGGGCCCAAGATGCCCGGGCCTCGGCAACCGGCCCCCAAAAGG  1650
LysPheCysArgGlyGluAlaSerGlyArgGlyGlyAlaLysThrAlaGlyLeuArgThrIleGlyProProArgMetLysGlnProAlaLeuLeuValHisThrProGlyProLysMetProGlyProArgGlnProAlaProLysArg

CCTCCCCCAGGACCATGCTATCGATGCCTCAAAGAAGGCCATTGGGCCCGGGATTGTCCTACCAAGGCCACCGGCCCACCTCCGGGACCTTGCCCCATATGTAAAGATCCTTCCCATTGGAAACGAGACTGTCCAACCCTCAAATCAAAA  1800
ProProProGlyProCysTyrArgCysLeuLysGluGlyHisTrpAlaArgAspCysProThrLysAlaThrGlyProProProGlyProCysProIleCysLysAspProSerHisTrpLysArgAspCysProThrLeuLysSerLys
p12
GAG◄                                                                                                                                                Pst I
AACTAATAGAGGGGGGACTTAGCGCCCCCCAAACCATAACACCTATAACGGATTCTCTTAGTGAGGCCGAATTAGAATGCTTACTTTCTATTCCTCTGGCTCGCAGCCGTCCCTCCGTGGCTGTATACCTGTCTGGCCCCTGGCTGCAGC
Asn***
                                                                                      Bam HI
CCTCTCAGAATCAAGCCCTCATGCTTGTGGACACCGGGGCTGAAAATACGGTTCTCCCACAAAATTGGCTGGTTCGAGATTACCCACGGATCCCCGCCGCAGTGCTCGGAGCAGGGGGAGTCTCCCGGAACAGATACAATTGGCTACAAG  2100
                                                              Hinc II
GCCCTCTGACCCTGGCTCTAAAACCAGAGGGTCCCTTTATCACCATCCCAAAAATTTTAGTTGACACTTCCGACAAATGGCAAATTTTAGGACGGGACGTCCCTCCCGCCTACAGGCTTCTATCTCCATACCTGAGGAAGTACGCCCCCC  2250
                                                                   ►POL
TGTGGTAGGCGTCTTGGATACCCCCCCGAGCCACATTGGATTAGAACATCTGCCCCCCCCACCTGAGGTGCCTCAATTCCCTTTAAACTAGAACGCCTCCAGGCCCTTCAAGACCTGGTCCATCGCTCTCTGGAGGCAGGTTATATCTCC  2400
                                                                      GlyAlaSerIleProPheLysLeuGluArgLeuGlnAlaLeuGlnAspLeuValHisArgSerLeuGluAlaGlyTyrIleSer

CCCTGGGACGGGCCAGGCAATAATCCAGTCTTCCCGGTACGGAAACCAAATGGCGCCTGGAGGTTTGTGCATGACCTACGAGCTACAAATGCTCTTACAAAGCCCATTCCGGCACTCTCTCCCGGACCGCCAGACCTTACCGCTATCCCT  2550
ProTrpAspGlyProGlyAsnAsnProValPheProValArgLysProAsnGlyAlaTrpArgPheValHisAspLeuArgAlaThrAsnAlaLeuThrLysProIleProAlaLeuSerProGlyProProAspLeuThrAlaIlePro
                         Bgl II
ACGCACCCTCCACATATCATTTGCCTAGATCTCAAAGATGCCTTCTTCCAGATTCCAGTCGAAGACCGCTTCCGCTTCTACTTGTCTTTTACCCTCCCATCCCCCGGGGGACTCCAACCTCATAGACGCTTTGCCTGGCGGGTCCTACCT  2700
ThrHisProProHisIleIleCysLeuAspLeuLysAspAlaPhePheGlnIleProValGluAspArgPheArgSerTyrLeuSerPheThrLeuProSerProGlyGlyLeuGlnProHisArgArgPheAlaTrpArgValLeuPro

CAAGGCTTCATTAACAGCCCAGCTCTTTTCGAACGAGCACTACAGGAACCTCTTCGCCAAGTTTCCGCCGCCTTTTCCCAGTCTCTTCTGGTGTCCTATATGGACGATATCCTTTACGCTTCGCCTACAGAAGAACAGCGGTCACAATGT  2850
GlnGlyPheIleAsnSerProAlaLeuPheGluArgAlaLeuGlnGluProLeuArgGlnValSerAlaAlaPheSerGlnSerLeuLeuValSerTyrMetAspAspIleLeuTyrAlaSerProThrGluGluGlnArgSerGlnCys
                                                                                                                                                    Bgl II
TATCAAGCCCTGGCTGCCCGCCTCCGGGACCTAGGGTTTCAGGTGGCATCCGAAAAGACTAGCCAGACGCCTTCGCCCGTCCCCTTTTTGGGACAAATGGTCCATGAGCAGATTGTCACCTACCAGTCCCTACCTACCTTGCAGATCTCA  3000
TyrGlnAlaLeuAlaAlaArgLeuArgAspLeuGlyPheGlnValAlaSerGluLysThrSerGlnThrProSerProValProPheLeuGlyGlnMetValHisGluGlnIleValThrTyrGlnSerLeuProThrLeuGlnIleSer

TCCCCAATTTCTCTTCACCAATTACAGGCGGTCTTAGGAGACCTCCAATGGGTCTCTAGGGGCACACCCACTACCCGCCGGCCCCTGCAACTTCTCTACTCTTCCCTTAAAAGGCATCATGACCCTAGGGCCATCATCCAGCTTTCCCCG  3150
SerProIleSerLeuHisGlnLeuGlnAlaValLeuGlyAspLeuGlnTrpValSerArgGlyThrProThrThrArgArgProLeuGlnLeuLeuTyrSerSerLeuLysArgHisHisAspProArgAlaIleIleGlnLeuSerPro
  Pvu II                                      Bgl II
GAACAGCTGCAAGGCATTGCAGAGCTTCGACAAGCCCTGTCCCACAACGCAAGATCTAGATATAACGAGCAAGAACCCCTGCTAGCCTACGTACACCTAACCCGGGCGGGGTCCACCCTGGTACTCTTCCAAAAGGGCGCTCAATTTCCC  3300
GluGlnLeuGlnGlyIleAlaGluLeuArgGlnAlaLeuSerHisAsnAlaArgSerArgTyrAsnGluGlnGluProLeuLeuAlaTyrValHisLeuThrArgAlaGlySerThrLeuValLeuPheGlnLysGlyAlaGlnPhePro
                                                                                    Pst I
CTGGCCTACTTTCAGACCCCCTTGACTGACAACCAAGCCTCACCTTGGGGCCTCCTTCTCCTGCTGGGATGCCAATACCTGCAGACTCAGGCCTTAAGCTCGTATGCCAAGCCCATACTTAAATATTATCACAATCTTCCTAAAACCTCT  3450
LeuAlaTyrPheGlnThrProLeuThrAspAsnGlnAlaSerProTrpGlyLeuLeuLeuLeuLeuGlyCysGlnTyrLeuGlnThrGlnAlaLeuSerSerTyrAlaLysProIleLeuLysTyrTyrHisAsnLeuProLysThrSer
Xba I                           Xho I
CTAGACAATTGGATTCAATCATCTGAGGACCCTCGAGTCCAGGAGTTGCTGCAATTGTGGCCCCAGATTTCCTCTCAGGGAATACAGCCCCCGGGCCCTTGGAAGACCTTAATCACCAGGGCAGAGGTTTTTTTGACGCCCCAGTTCTCC  3600
LeuAspAsnTrpIleGlnSerSerGluASpProArgValGlnGluLeuLeuGlnLeuTrpProGlnIleSerSerGlnGlyIleGlnProProGlyProTrpLysThrLeuIleThrArgAlaGluValPheLeuThrProGlnPheSer

CCTGATCCGATTCCTGCGGCCCTTTGCCTCTTTAGTGACGGGGCTACAGGACGAGGAGCATATTGCTTGTGGAAAGACCACCTTTTAGACTTTCAGGCCGTTCCGGCCCCAGAATCCGCTCAAAAGGGAGAACTAGCAGGTCTCTTGGCG  3750
ProAspProIleProAlaAlaLeuCysLeuPheSerAspGlyAlaThrGlyArgGlyAlaTyrCysLeuTrpLysAspHisLeuLeuAspPheGlnAlaValProAlaProGluSerAlaGlnLysGlyGluLysAlaGlyLeuLeuAla

GGCTTAGCAGCCGCCCCGCCTGAACCTGTAAATATATGGGTAGATTCCAAATACCTGTACTCTTTGCTCAGAACCCTAGTTCTGGGAGCTTGGCTTCAACCTGACCCCGTACCCTCCTACGCCCTCCTATATAAAAGCCTCCTCCGACAT  3900
GlyLeuAlaAlaAlaProProGluProValAsnIleTrpValAspSerLysTyrLeuTyrSerLeuLeuArgThrLeuValLeuGlyAlaTrpLeuGlnProAspProValProSerTyrAlaLeuLeuTyrLysSerLeuLeuArgHis

CCAGCAATCGTTGTTGGTCATGTCCGGAGCCACTCTTCAGCATCCCACCCTATTGCTTCCCTGAACAATTATGTAGATCAACTGCTTCCCTTAGAAACTCCAGAGCAATGGCATAAGCTCACCCACTGCAACTCTCGGGCCTTGTCTCGA  4050
ProAlaIleValValGlyHisValArgSerHisSerSerAlaSerHisProIleAlaSerLeuAsnAsnTyrValAspGlnLeuLeuProLeuGluThrProGluGlnTrpHisLysLeuThrHisCysAsnSerArgAlaLeuSerArg
                                                                    Hind III
TGGCCGAACCCACGTATCTCTGCCTGGGACCCCCGTTCCCCCGCTACGCTGTGTGAAACCTGCCAAAAGCTTAATCCAACTGGAGGAGGAAAGATGCGAACTATTCAGAGAGGGTGGGCCCCGAATCATATTTGGCAGGCCGATATAACC  4200
TrpProAsnProArgIleSerAlaTrpAspProArgSerProAlaThrLeuCysGluThrCysGlnLysLeuAsnProThrGlyGlyGlyLysMetArgThrIleGlnArgGlyTrpAlaProAsnHisIleTrpGlnAlaAspIleThr

CATTATAAATACAAACAGTTCACCTACGCTCTGCATGTGTTTGTAGATACTTACTCTGGAGCTACTCATGCCTCGGCGAAGCGTGGGCTCACCACTCAAACGACCATTGAGGGCCTTCTTGAGGCCATAGTGCATCTGGGTCGCCCAAAA  4350
HisTyrLysTyrLysGlnPheThrTyrAlaLeuHisValPheValAspThrTyrSerGlyAlaThrHisAlaSerAlaLysArgGlyLeuThrThrGlnThrThrIleGluGlyLeuLeuGluAlaIleValHisLeuGlyArgProLys
                                        Splice acceptor
AAGCTAAACACTGACCAAGGTGCAAACTACACCTCCAAAACCTTTGTCAGGTTTTGCCAGCAGTTCGGAGTTTCCCTTTCTCATCATGTTCCCTACAACCCCACAAGTTCGGGGTTAGATGAACGGACAAATGGACTGCTCAAACTTCTT  4500
LysLeuAsnThrAspGlnGlyAlaAsnTyrThrSerLysThrPheValArgPheCysGlnGlnPheGlyValSerLeuSerHisHisValProTyrAsnProThrSerSerGlyLeuAspGluArgThrAsnGlyLeuLeuLysLeu
                                                     Xho I
CTATCTAAATATCACCTAGACGAACCCCACCTTCCCATGACTCAGGCCCTTTCTCGAGCCCTCTGGACTCACAATCAGATTAACCTCCTACCAATTCTAAAGACCAGATGGGAGCTACACCATTCACCCCCACTTGCTGTCATTTCAGAG  4650
LeuSerLysTyrHisLeuAspGluProHisLeuProMetThrGlnAlaLeuSerArgAlaLeuTrpThrHisAsnGlnIleAsnLeuLeuProIleLeuLysThrArgTrpGluLeuHisHisSerProProLeuAlaValIleSerGlu
```

(*Fig. 1 continues on following page.*)

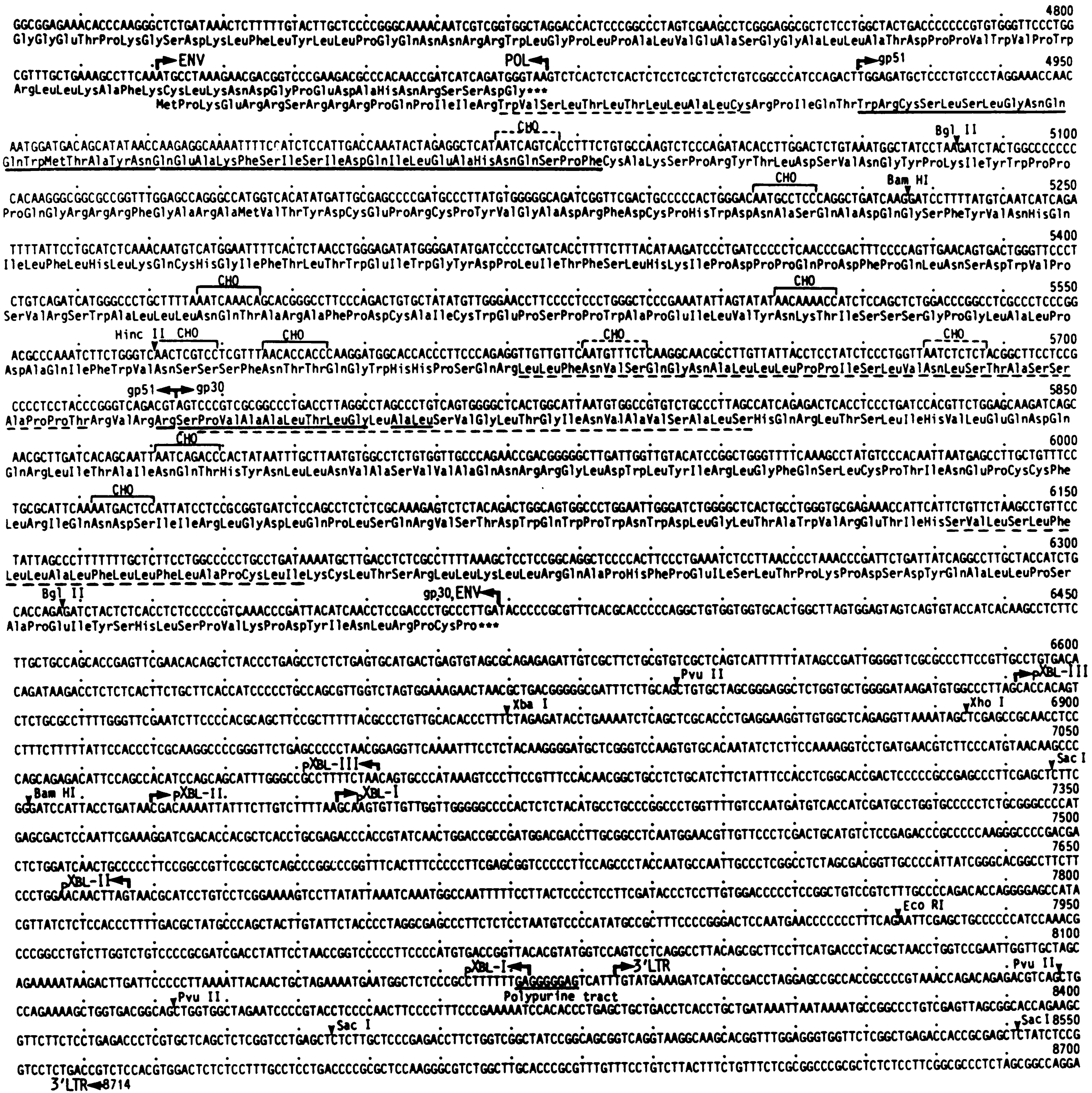

FIG. 1. Complete nucleotide sequence of proviral BLV genome. Deduced amino acid sequence is given below the nucleotide sequence. Amino acid residues that match experimentally determined ones (see text) are underlined with a solid line. Stretches of hydrophobic amino acid residues are underlined with a broken line. Dots mark every 10th nucleotide. See ref. 4 for LTRs. Major structural features are indicated. CHO, potential glycosylation site (three CHOs with broken lines may not be glycosylated, as described in the text); $pX_{BL}$-I, -II, and -III are unidentified open reading frames.

contains internally duplicated sequences (brackets in Fig. 1) that have periodically placed cysteine residues. A similar sequence is also duplicated in the avian (but not murine) retrovirus homologue (16).

***pol* Gene.** About 500 bp downstream of the *gag* gene appears the second open reading frame encoding 852 amino acid residues (positions 2317–4875). This is the largest reading frame, does not open with an ATG codon, and is located in the middle of the viral genome, all of which are properties associated with the retrovirus *pol* gene (10, 11). Fig. 2 shows two-dimensional homology matrices between the putative BLV *pol* product and those of Mo-MuLV and RSV, where we find two major homologous regions termed NHR (for the $NH_2$-terminal homologous region) and CHR (for the COOH-terminal homologous region). Between BLV and Mo-MuLV (Fig. 2*A*), NHR is shifted upward from the extension of CHR, indicating that Mo-MuLV *pol* harbors an insertion(s) between NHR and CHR. Furthermore, Mo-MuLV *pol* has an additional 180-residue $NH_2$-terminal sequence that is completely lacking in the BLV *pol*. Between BLV and RSV (Fig. 2*B*), on the other hand, NHR and CHR are both on a nearly diagonal line of the square frame. In addition, both NHR and CHR sequence homologies observed between BLV and RSV are significantly higher than those between

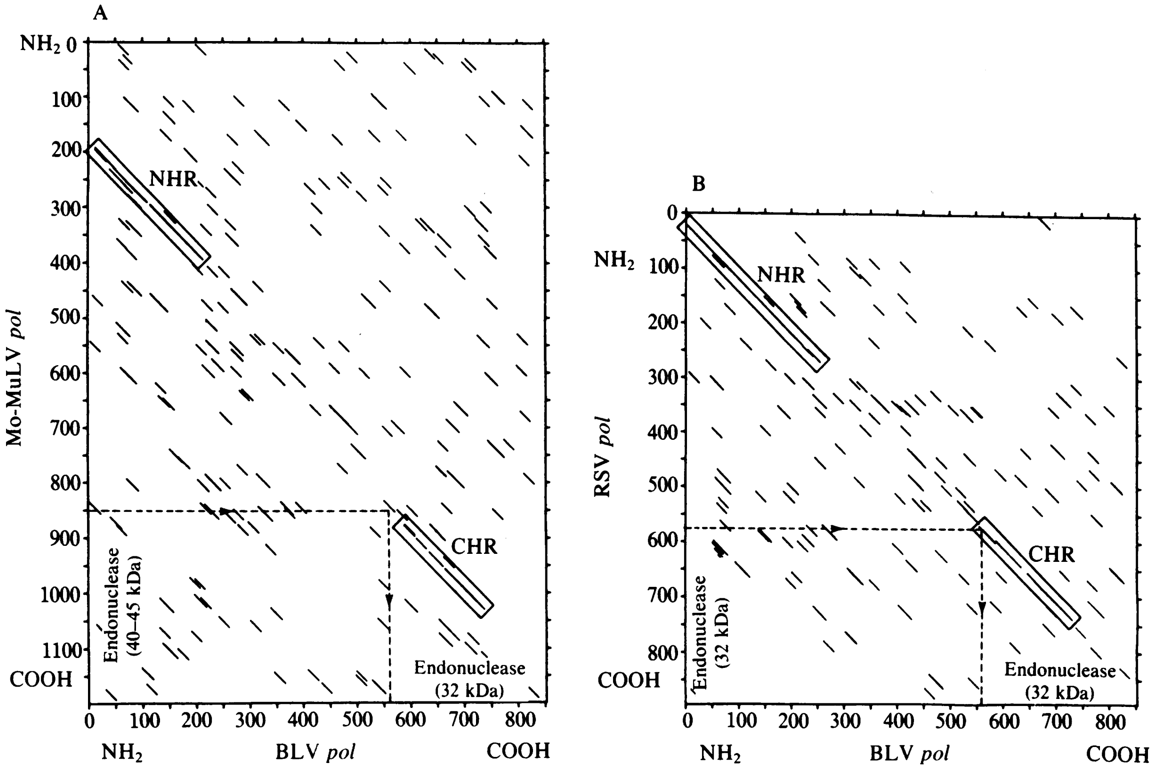

FIG. 2. Two-dimensional homology matrix comparisons of amino acid sequences encoded by the BLV *pol* gene and those of Mo-MuLV (*A*) and RSV (*B*). Each diagonal line represents a region in which there is an average score value of at least 65 (8) in 20 contiguous amino acids between two viruses. NHR and CHR denote the major $NH_2$- and COOH-terminal homologous regions, respectively. Putative endonuclease domain of the BLV *pol* gene is estimated by broken lines with arrows (see text for details).

BLV and Mo-MuLV (see Fig. 4*A*). Thus, very interestingly, the BLV *pol* more closely resembles the RSV *pol* in both sequence organization and homology.

The BLV *pol* product (≈95 kDa) could be somewhat larger than the *pol*-encoded BLV reverse transcriptase (70 kDa; see ref. 1). The RSV *pol* and presumably the Mo-MuLV *pol* also encode an endonuclease at their COOH termini, with molecular sizes of 32 and 40–45 kDa, respectively (17, 18, 25). These endonuclease domains closely correspond to CHRs (as shown by broken lines in Fig. 2), which in turn indicates that the BLV *pol* also encodes an endonuclease represented by the CHR. This putative BLV endonuclease (≈32 kDa; see Fig. 2) would explain the discrepancy of the sizes between the *pol* product and the reverse transcriptase described above.

***env* Gene.** BLV *env* products are gp51 (surface glycoprotein) and gp30 (transmembrane protein) (1). The third open reading frame spans nucleotides 4821–6368 and encodes 515 amino acid residues with an initiator methionine (Fig. 1). The 5′ end of this open reading frame overlaps the 3′ end of the *pol* gene and its predicted amino acid sequence contains 10 *N*-asparagine-linked glycosylation sites (Fig. 1), showing features associated with the *env* gene (10, 11). During preparation of this paper, we learned of a 12-residue $NH_2$-terminal sequence of gp30 and a 38-residue $NH_2$-terminal sequence and a COOH-terminal residue (arginine) of gp51 (19), almost all residues of which were found in our deduced sequence (underlining in Fig. 1). Thus, gp51 spans positions 4920–5723 and gp30 spans positions 5724–6365 (Fig. 1). gp51 contains 8 potential glycosylation sites, but 3 of these may not be glycosylated because they are either embedded in a long hydrophobic sequence or immediately followed by a proline residue (Fig. 1); the hydrophobic sequence may interact with the $NH_2$-terminal hydrophobic sequence of gp30 (20). gp30 contains two glycosylation sites and has two hydrophobic sequences, the COOH-terminal one of which may anchor the membrane (20). The $NH_2$-terminal 33-residue sequence preceding gp51 could be a signal sequence for membrane insertion of the primary product because it has hydrophilic $NH_2$-terminal regions followed by a hydrophobic core (Fig. 1) (21). The splice acceptor site for the *env* mRNA is most likely the sequence C-C-T-T-T-G-T-C-A-G-G (positions 4391–4401), which fits well the consensus sequence (12).

The sequence of *env* transmembrane protein is more strongly conserved between related retroviruses than is the sequence of surface glycoprotein, which is a host determinant (22). Fig. 3 shows the sequence alignment of the first 120 amino acid residues of BLV gp30, Mo-MuLV p15E (10), and RSV gp37 (11). Overall homology between BLV and Mo-MuLV (33%) is significantly higher than that between BLV and RSV (20%). This is of particular interest because the BLV *pol* more closely resembled the RSV *pol*. We do not know how such a chimeric genome was generated. However, one possibility may be that the transmembrane protein sequence is somehow restricted by the host range (i.e., mammalian vs. avian). Alternatively, *env* gene recombination might have occurred between progenitors of BLV and MuLV; such an event appears to have recently occurred between progenitors of certain mammalian type C and type D

```
              .        20         .        40         .        60
MuLV EPVSLTL-ALLLGGLTMGGIAAGIGTGTT--ALMATQQFQQLQAAV---QDDLREVEKSI
      ●●   ●  ● ●  ●  ●    ●●       ●●   ●    ●       ●●  ●     ●
BLV  SPVA-AL-TLGL-ALSVG--LTGINVAVS--AL-SHQRLTSLIHVL--EQDQQRLIT-AI
     ● ●   ●                ●            ● ●            ●  ● ●
RSV  S-VS-HLDDTCSDEVQLWGPTARIFASILAPGVAAAQALREIERLACWSVKQANLTTSLL
              .        80         .       100         .       120
MuLV SNLEKSLTSLSEVVLQNRRGLDLLFLKEG--GLCAALKEECCF-Y-ADHTGLVRDSMAKL
           ●     ●● ●●●●●●● ●    ●   ●●    ● ●●●          ●     ●
BLV  NQTHYNLLNVASVVAQNRRGLDWLYIRLGFQSLCPTINEPCCF-LRIQNDSIIR--LGDL
                  ● ●●●   ● ●   ●     ●      ●●● ●     ●●
RSV  GDLLDDVTSIRHAVLQNRAAIDFL-L-LAHGHGCEDVAGMCCFNLSDHSESIQKK-FQLM
```

FIG. 3. Sequence alignment of transmembrane proteins of Mo-MuLV, BLV, and RSV. The $NH_2$-terminal 120 amino acids (expressed by one-letter code) of the *env* transmembrane proteins of Mo-MuLV p15E (10), BLV gp30 (this paper), and RSV gp37 (11) are aligned with gaps (-) (9). Residues of BLV that match those of Mo-MuLV and RSV are shown by closed circles in upper and lower lines, respectively.

retroviruses (23). Finally, the cysteine residues (positions 94, 101, and 102; Fig. 3) well conserved between the three transmembrane proteins probably make S—S bridges between the transmembrane protein and surface glycoprotein (20).

**$pX_{BL}$, Unidentified Open Reading Frames.** In addition to the *gag*, *pol*, and *env* genes, the BLV genome harbors a 1800-bp region at its 3′ end containing several unidentified open reading frames (Fig. 1). This region corresponds to a pX region of HTLV (13) and is designated as $pX_{BL}$ for BLV. As with pX (13), the $pX_{BL}$ showed no significant hybridization with the host (bovine) DNA, indicating that it is not a recently acquired cellular gene. The largest open reading frame $pX_{BL}$-I, opening with a GCA triplet and ending with a terminator TGA at the beginning of the polypurine tract (Fig. 1), could encode 308 amino acid residues.

**Evolutionary Relationship of BLV to Other Retroviruses.** The *pol* gene is the best conserved of the three retroviral genes (8, 23) and sequencing data of its well conserved region (corresponding to the CHR in Fig. 2) are available for all, except for type A, of the major oncovirus genera—i.e., types B, D, and mammalian and avian type C viruses (23). Fig. 4*A* shows the amino acid homology and divergence of the CHRs between BLV, HTLV, and the other oncoviruses. Clearly, BLV is most closely related to HTLV (45% homology) and is more closely related to types B, D, and avian type C viruses (29%) than to mammalian type C virus (25%). (A similar trend was obtained between NHRs of BLV, HTLV, and mammalian and avaian type C viruses, although the homologies were generally 6%–8% higher than those observed for CHRs.) Interestingly, HTLV has generally stronger sequence homology with other viruses than has BLV (Fig. 4*A*), implying that HTLV has evolved more slowly than BLV.

Fig. 4*B* shows the evolutionary tree of the oncovirus genera based on the CHR sequence divergences in Fig. 4*A*. Both BLV and HTLV (with a divergence of 0.8) are only distantly related to the other groups (with a divergence of 1.2–1.25), especially to the mammalian type C virus (with a divergence of 1.3–1.4). This allows us to propose that BLV and HTLV constitute a group of Oncovirinae designated here as type "E" (Fig. 4*B*). In support of this, these viruses appear to be morphologically and biochemically somewhat different from the major mammalian type C viruses (refs. 1 and 2; M. Nakai, personal communication) and have now turned out to be genetically very different from other type viruses in that they have pX sequences. [We presume that BLV and HTLV somehow acquired the pX sequence after their progenitor had branched away from the common progenitor of types B, D, and avian type C viruses (Fig. 4*B*).] Fig. 4 also shows that avian and mammalian type C viruses are evolutionarily very distant from each other (with a divergence of 1.44) and are designated here as types $C_A$ and $C_M$, respectively. It should be noted that the evolutionary tree proposed reflects well the divalent cation dependency of the reverse transcriptase of each type virus: $Mg^{2+}$ for types B, D, $C_A$, and E, and $Mn^{2+}$ for type $C_M$ viruses (23).

A

|  | BLV | HTLV | SMRV | MMTV | RSV | MuLV |
|---|---|---|---|---|---|---|
| BLV | – | 0.81 | 1.23 | 1.23 | 1.26 | 1.38 |
| HTLV | 44.6 | – | 1.09 | 1.21 | 1.21 | 1.29 |
| SMRV | 29.4 | 33.7 | – | 0.65 | 0.97 | 1.33 |
| MMTV | 29.4 | 29.9 | 52.2 | – | 0.96 | 1.33 |
| RSV | 28.3 | 29.9 | 37.8 | 38.3 | – | 1.44 |
| MuLV | 25.1 | 27.4 | 26.5 | 26.5 | 23.8 | – |

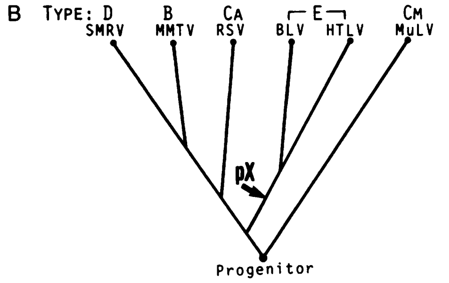

FIG. 4. Amino acid homology and divergence of COOH-terminal *pol* sequences between various oncoviruses (*A*) and proposed evolutionary tree of the oncovirus genera (*B*). (*A*) COOH-terminal *pol* sequence (equivalent to CHR in Fig. 2) of BLV (positions 3979–4530; Fig. 1) and its corresponding regions of HTLV (13), squirrel monkey retrovirus, type D (SMRV) (23), mouse mammary tumor virus, type B (MMTV) (23), RSV (avian type C; ref. 11) and Mo-MuLV (mammalian type C; ref. 10) were aligned (9), and the percentage homology and the divergence (corrected for multiple hits; ref. 22) are shown in lower left and upper right halves of the matrix, respectively. (*B*) The evolutionary tree was made from corrected divergences in *A* by the unweighted pair-group clustering method (24), in which the divergence between two viruses is expressed as height from their branching point. The avian and mammalian type C viruses are tentatively designated as type $C_A$ and $C_M$, respectively, because of their distant relationship in the tree. The pX sequence is assumed to have been acquired by the progenitor of BLV and HTLV around the position indicated.

We have determined the complete nucleotide sequence of the BLV genome, deduced its genomic structure, and proposed another classification of BLV and HTLV in Oncovirinae. A detailed comparison of the respective genes, including the pX sequence, of BLV and HTLV and identification of a potential protease-coding gene between the *gag* and *pol* genes of these retroviruses have been described elsewhere (26, 27).

This work was supported by a grant-in-aid for Cancer Research from the Ministry of Education, Science and Culture of Japan.

1. Burny, A., Bruck, C., Chantrenne, H., Cleuter, Y., Dekegel, D., Ghysdael, J., Kettmann, R., Leclercq, M., Leunen, J., Mammerickx, M. & Portetelle, D. (1980) in *Viral Oncology*, ed. Klein, G. (Raven, New York), pp. 231–289.
2. Weiland, F., Ueberschär, S., Straub, O. C., Kaaden, O. R. & Dietzschold, B. (1974) *Intervirology* **4**, 140–149.
3. Copeland, T. D., Oroszlan, S., Kalyanaraman, V. S., Sarngadharan, M. G. & Gallo, R. C. (1983) *FEBS Lett.* **162**, 390–395.
4. Sagata, N., Yasunaga, T., Ogawa, Y., Tsuzuku-Kawamura, J. & Ikawa, Y. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 4741–4745.
5. Poiesz, B. J., Ruscetti, F. W., Gazdar, A. F., Bunn, P.A., Minna, J. D. & Gallo, R. C. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 7415–7419.
6. Sagata, N., Ogawa, Y., Kawamura, J., Onuma, M., Izawa, H. & Ikawa, Y. (1983) *Gene* **26**, 1–10.
7. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
8. Toh, H., Hayashida, H. & Miyata, T. (1983) *Nature (London)* **305**, 827–829.
9. Needleman, S. B. & Wunsh, C. D. (1970) *J. Mol. Biol.* **48**, 443–453.
10. Shinnick, T. M., Lerner, R. A. & Sutcliffe, J. G. (1981) *Nature (London)* **293**, 543–548.
11. Schwartz, D. E., Tizard, R. & Gilbert, W. (1983) *Cell* **32**, 853–869.
12. Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459–472.
13. Seiki, M., Hattori, S., Hirayama, Y. & Yoshida, M. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3618–3622.
14. Shimotohno, K., Golde, D. W., Miwa, M., Sugimura, T. & Chen, I. S. Y. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 1079–1083.
15. Oroszlan, S., Copeland, T. D., Henderson, L. E., Stephenson, J. R. & Gilden, R. V. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 2996–3000.
16. Copeland, T. D., Morgan, M. A. & Oroszlan, S. (1983) *FEBS Lett.* **156**, 37–40.
17. Levin, J. G., Hu, S. C., Rein, A., Messer, L. I. & Gerwin, B. I. (1984) *J. Virol.* **51**, 470–478.
18. Schiff, R. D. & Grandgenett, D. P. (1978) *J. Viol.* **28**, 279–291.
19. Schultz, A. M., Copeland, T. D. & Oroszlan, S. (1984) *Virology* **135**, 417–427.
20. Lenz, J., Crowther, R., Straceski, A. & Haseltine, W. (1982) *J. Virol.* **42**, 519–529.
21. Emr, S. D. & Silhavy, T. J. (1982) *J. Cell Biol.* **95**, 689–696.
22. Zukerkandl, E. & Pauling, L. (1965) in *Evolving Genes and Proteins*, eds. Bryson, V. & Vogel, H. J. (Academic, New York), pp. 97–106.
23. Chiu, I.-M., Callahan, R., Tronick, S. R., Schlom, J. & Aaronson, S. A. (1984) *Science* **233**, 364–370.
24. Sokal, R. R. & Sneath, P. H. A. (1963) *Principles of Numerical Taxonomy* (Freeman, San Francisco).
25. Hippenmeyer, P. J. & Grandgenett, D. P. (1984) *Virology* **137**, 358–370.
26. Sagata, N., Yasunaga, T., Ohishi, K., Tsuzuku-Kawamura, J., Onuma, M. & Ikawa, Y. (1984) *EMBO J.*, in press.
27. Sagata, N., Yasunaga, T. & Ikawa, Y. (1984) *FEBS Lett.* **178**, 79–82.