JOURNAL OF VIROLOGY, Nov. 1986, p. 589–598
0022-538X/86/110589-10$02.00/0

Vol. 60, No. 2

Nucleotide Sequence of Human Endogenous Retrovirus Genome Related to the Mouse Mammary Tumor Virus Genome

MASAO ONO,[1]* TERUO YASUNAGA,[2] TAKASHI MIYATA,[3] AND HIROMI USHIKUBO[1]

Department of Molecular Biology, School of Medicine, Kitasato University, Sagamihara-shi, Kanagawa 228,[1] Computation Center, Institute of Physical and Chemical Research, Wako-shi, Saitama 351,[2] and Department of Biology, Faculty of Science, Kyushu University, Fukuoka-shi, Fukuoka 812,[3] Japan

Received 14 May 1986/Accepted 18 July 1986

We determined the complete nucleotide sequence of the human endogenous retrovirus genome HERV-K10 isolated as the sequence homologous to the Syrian hamster intracisternal A-particle (type A retrovirus) genome. HERV-K10 is 9,179 base pairs long with long terminal repeats of 968 base pairs at both ends; a sequence 290 base pairs long, however, was found to be deleted. It was concluded that a composite genome having the 290-base-pair fragment is the prototype HERV-K provirus *gag* (666 codons), protease (334 codons), *pol* (937 codons), and *env* (618 codons) genes. The size of the protease gene product of HERV-K is essentially the same as that of A- and D-type oncoviruses but nearly twice that of other retroviruses. A comparison of the deduced amino acid sequences encoded by the *pol* region showed HERV-K to be closely related to types A and D retroviruses and even more so to type B retrovirus. It was noted that the *env* gene product of HERV-K structurally resembles the mouse mammary tumor virus (type B retrovirus) *env* protein, and the possible expression of the HERV-K *env* gene in human breast cancer cells is discussed.

Recently, human endogenous retrovirus genomes, some homologous to mammalian type B (2, 3, 8, 42) or type C (1, 18, 21, 22, 28) retroviruses or to retrovirus proviruses of an unidentified type (16), were isolated and characterized. However, the types, structures, and functions of human endogenous retroviruses and their causal relationships to human cancer have yet to be fully clarified.

Intracisternal A-particle (IAP; type A retrovirus) genomes are moderately repetitive endogenous retrovirus genomes interspersedly present in rodents such as mice, rats, and Syrian hamsters (15). Our recent determination of the complete nucleotide sequence of a Syrian hamster IAP genome confirmed the presence of conserved *pol* regions among retrovirus genomes and revealed a close evolutionary relationship between the type A retrovirus and types B and D oncoviruses (25). Based on these findings, sequences hybridizable with a fragment primarily encoding the conserved *pol* region of the Syrian hamster IAP genome were isolated (24) from a human fetal liver gene library (14).

Typical human endogenous retrovirus genomes, termed HERV-K, were found to be 9.2 or 9.5 kilobases (kb) long, with long terminal repeats (LTRs) of ca. 970 base pairs (bp) (24). These HERV-K proviruses are present at ca. 50 copies per haploid human genome and are homologous to the mouse mammary tumor virus (MMTV; type B retrovirus), as well as to the type A retrovirus. Since the $tRNA_{1,2}^{Lys}$ was identified as a presumed primer for reverse transcription, the genes were tentatively classified as members of the HERV-K family (K is an abbreviated form of the Lys).

For greater clarification of the organization of HERV-K genomes, the complete nucleotide sequence of a typical HERV-K provirus, HERV-K10, was determined. HERV-K10 is 9,179 bp long with an LTR-*gag*-protease (*prt*)-*pol*-*env*-LTR structure, although a deletion 290 bp long was noted in the boundary region between the *pol* and *env* genes. A composite genome 9,469 bp long containing the 290-bp sequence was named HERV-K10(+) and was concluded to be a prototype HERV-K genome. A comparison of the deduced amino acid sequences encoded by the *pol* regions indicated that this provirus is closely related to IAP-H18 (type A retrovirus) (25), a squirrel monkey retrovirus (type D retrovirus) (4), and type D retrovirus (SRV-1) (26) linked to the simian acquired immune deficiency syndrome and even more so to MMTV (type B retrovirus) (4, 8).

* Corresponding author.

MATERIALS AND METHODS

Clones and DNA sequencing. HERV-K proviruses cloned from a human fetal liver gene library have been described (24). In the present study, DNA fragments were labeled either at the 5′ ends with [γ-^{32}P]ATP and T4 polynucleotide kinase or at the 3′ ends with [α-^{32}P]ddATP and terminal deoxynucleotidyl transferase. The nucleotide sequences of the fragments were determined by the method of Maxam and Gilbert (19). All sequences in both DNA strands were determined.

Computer-assisted analyses of nucleotide and amino acid sequences. A two-dimensional homology matrix comparison of the amino acid sequences was conducted with a computer program developed for distantly related proteins (40). A DNA version of this method developed by H. Hayashida and T. Miyata (unpublished) was used to make a two-dimensional homology matrix comparison of the nucleotide sequences (29). Alignment of the amino acid sequences was performed by the method of Needleman and Wunsch (20), checked by hand, and corrected as necessary.

RESULTS

Structural features of HERV-K10(+) gene. In a previous paper, four characterized HERV-K proviruses (HERV-K8, HERV-K10, HERV-K18, and HERV-K22) were shown to be divided into two groups (24). The length of the proviruses of one group (HERV-K10 and HERV-K18) averages 9.2 kb, whereas the proviruses in the other group (HERV-K8 and


5'LTR
TGTGGGGAAAAGCAAGAGAGATCAAATTGTTACTGTGTCTGTGTAGAAAGAAGTAGACATAGGAGACTCCATTTGTTATGTGCTAAG
TATGACCTTACCCCCAACCCCGTGCTCTCTGAAACGTGTGCTGTGTCAACTCAGGGTTGAATGGATTAAGGGCGGTGCAGGATGTGCTTTGTTAAACAGATGCTTGAAGGCAGCATGCTC
CTTAAGAGTCATCACCACTCCCTAATCTCAAGTACCCAGGGACACAAAAACTGCGGAAGGCCGCAGGGACCTCTGCCTAGGAAAGCCAGGTATTGTCCAAGGTTTCTCCCCATGTGATAG
TCTGAAATATGGCCTCGTGGGAAGGGAAAGACCTGACCGTCCCCCAGCCCGACACCTGTAAAGGGTCTGTGCTGAGGAGGATTAGTAAAAGAGGAAGGAATGCCTCTTGCAGTTGAGACA
AGAGGAAGGCATCTGTCTCCTGCCTGTCCCTGGGCAATGGAATGTCTCGGTATAAAACCCGATTGTATGCTCCATCTACTGAGATAGGGAAAAACCGCCTTAGGGCTGGAGGTGGGACCT
GCGGGCAGCAATACTGCTTTGTAAAGCATTGAGATGTTTATGTGTATGCATATCCAAAAGCACAGCACTTAATCCTTTACATTGTCTATGATGCCAAGACCTTTGTTCACGTGTTTGTCT
GCTGACCCTCTCCCCACAATTGTCTTGTGACCCTGACACATCCCCCTCTTTGAGAAACACCCACAGATGATCAATAAATACTAAGGGAACTCAGAGGCTGGCGGGATCCTCCATATGCTG
AACGCTGGTTCCCCGGGTCCCCTTATTTCTTTCTCTATACTTTGTCTCTGTGTCTTTTTCTTTTCCAAATCTCTCGTCCCACCTTACGAGAAACACCCACAGGTGTGTAGGGGCAACCCA
*ThrLeuValProArgValProLeuPheLeuSerLeuTyrPheValSerValSerPheSerPheProAsnLeuSerSerHisLeuThrArgAsnThrHisArgCysValGlyAlaThrHis
5'LTR PBS SD
CCCCTACATCTGGTGCCCAACGTGGAGGCTTTTCTCTAGGGTGAAGGTACGCTCGAGCGTAATCATTGAGGACAAGTCGACGAGAGATCCCGAGTACATCTACAGTCAGCCTTACGGTAA
ProTyrIleTrpCysProThrTrpArgLeuPheSerArgValLysValArgSerSerValIleIleGluAspLysSerThrArgAspProGluTyrIleTyrSerGlnProTyrGlyLys
GCTTGCGCGCTCGGAAGAAGCTAGGGTGATAATGGGGCAAACTAAAAGTAAAATTAAAAGTAAATATGCCTCTTATCTCAGCTTTATTAAAATTCTTTTAAAAAGAGGGGGAGTTAAAGT
LeuAlaArgSerGluGluAlaArgValIleMetGlyGlnThrLysSerLysIleLysSerLysTyrAlaSerTyrLeuSerPheIleLysIleLeuLeuLysArgGlyGlyValLysVal
ATCTACAAAAAATCTAATCAAGCTATTTCAAATAATAGAACAATTTTGCCCATGGTTTCCAGAACAAGGAACTTCAGATCTAAAAGATTGGAAAAGAATTGGTAAGGAACTAAAACAAGC
SerThrLysAsnLeuIleLysLeuPheGlnIleIleGluGlnPheCysProTrpPheProGluGlnGlyThrSerAspLeuLysAspTrpLysArgIleGlyLysGluLeuLysGlnAla
AGGTAGGAAGGGTAATATCATTCCACTTACAGTATGGAATGATTGGGCCATTATTAAAGCAGCTTTAGAACCATTTCAAACAGAAGAAGATAGCATTTCAGTTTCTGATGCCCCTGGAAG
GlyArgLysGlyAsnIleIleProLeuThrValTrpAsnAspTrpAlaIleIleLysAlaAlaLeuGluProPheGlnThrGluGluAspSerIleSerValSerAspAlaProGlySer
CTGTTTAATAGATTGTAATGAAAACACAAGGAAAAAATCCCAGAAAGAAACCGAAAGTTTACATTGCGAATATGTAGCAGAGCCGGTAATGGCTCAGTCAACGCAAAATGTTGACTATAA
CysLeuIleAspCysAsnGluAsnThrArgLysLysSerGlnLysGluThrGluSerLeuHisCysGluTyrValAlaGluProValMetAlaGlnSerThrGlnAsnValAspTyrAsn
TCAATTACAGGAGGTGATATATCCTGAAACGTTAAAATTAGAAGGAAAAGGTCCAGAATTAATGGGGCCATCAGAGTCTAAACCACGAGGCACAAGTCCTCTTCCAGCAGGTCAGGTGCT
GlnLeuGlnGluValIleTyrProGluThrLeuLysLeuGluGlyLysGlyProGluLeuMetGlyProSerGluSerLysProArgGlyThrSerProLeuProAlaGlyGlnValLeu
CGTAAGATTACAACCTCAAAAGCAGGTTAAAGAAAATAAGACCCAACCGCAAGTAGCCTATCAATACTGCCGCTGGCTGAACTTCAGTATCGGCCACCCCCAGAAAGTCAGTATGGATAT
ValArgLeuGlnProGlnLysGlnValLysGluAsnLysThrGlnProGlnValAlaTyrGlnTyrCysArgTrpLeuAsnPheSerIleGlyHisProGlnLysValSerMetAspIle
***AspProThrAlaSerSerLeuSerIleLeuProGluAlaGluLeuGlnTyrArgProProProGluSerGlnTyrGlyTyr
CCAGGAATGCCCCCAGCACCACAGGGCAGGGCGCCATACCATCAGCCGCCCACTAGGAGACTTAATCCTATGGCACCACCTAGTAGACAGGGTAGTGAATTACATGAAATTATTGATAAA
GlnGluCysProGlnHisHisArgAlaGlyArgHisThrIleSerArgProLeuGlyAspLeuIleLeuTrpHisHisLeuValAspArgValValAsnTyrMetLysLeuLeuIleAsn
ProGlyMetProProAlaProGlnGlyArgAlaProTyrHisGlnProProThrArgArgLeuAsnProMetAlaProProSerArgGlnGlySerGluLeuHisGluIleAspLys
TCAAGAAAGGAAGGAGATACTGAGGCATGGCAATTCCCAGTAACGTTAGAACCGATGCCACCTGGAGAAGGAGCCCAAGAGGGAGAGCCTCCCACAGTTGAGGCCAGATACAAGTCTTTT
GlnGluArgLysGluIleLeuArgHisGlyAsnSerGln***
SerArgLysGluGlyAspThrGluAlaTrpGlnPheProValThrLeuGluProMetProProGlyGluGlyAlaGlnGluGlyGluProProThrValGluAlaArgTyrLysSerPhe
TCGATAAAAATGCTAAAAGATATGAAAGAGGGAGTAAAACAGTATGGACCCAACTCCCCTTATATGAGGACATTATTAGATTCCATTGCTTATGGACATAGACTCATTCCTTATGATTGG
SerIleLysMetLeuLysAspMetLysGluGlyValLysGlnTyrGlyProAsnSerProTyrMetArgThrLeuLeuAspSerIleAlaTyrGlyHisArgLeuIleProTyrAspTrp
GAGATTCTGGCAAAATCGTCTCTCTCACCCTCTCAATTTTTACAATTTAAGACTTGGTGGATTGATGGGGTACAAGAACAGGTCCGAAGAAATAGGGCTGCCAATCCTCCAGTTAACATA
GluIleLeuAlaLysSerSerLeuSerProSerGlnPheLeuGlnPheLysThrTrpTrpIleAspGlyValGlnGluGlnValArgArgAsnArgAlaAlaAsnProProValAsnIle
GATGCAGATCAACTATTAGGAATAGGTCAAAATTGGAGTACTATTAGTCAACAAGCATTAATGCAAAATGAGGCCATTGAGCAAGTTAGAGCTATCTGCCTTAGAGCTTGGGAAAAAATC
AspAlaAspGlnLeuLeuGlyIleGlyGlnAsnTrpSerThrIleSerGlnGlnAlaLeuMetGlnAsnGluAlaIleGluGlnValArgAlaIleCysLeuArgAlaTrpGluLysIle
CAAGACCCAGGAAGTACCTGCCCCTCATTTAATACAGTAAGACAAGGTTCAAAAGAGCCCTACCCTGATTTTGTGGCAAGGCTCCAAGATGTTGCTCAAAAGTCAATTGCCGATGAAAAA
GlnAspProGlySerThrCysProSerPheAsnThrValArgGlnGlySerLysGluProTyrProAspPheValAlaArgLeuGlnAspValAlaGlnLysSerIleAlaAspGluLys
GCCGGTAAGGTCATAGTGGAGTTGATGGCATATGAAAACGCCAATCCTGAGTGTCAATCAGCCATTAAGCCATTAAAAGGAAAGGTTCCTGCAGGATCAGATGTAATCTCAGAATATGTA
AlaGlyLysValIleValGluLeuMetAlaTyrGluAsnAlaAsnProGluCysGlnSerAlaIleLysProLeuLysGlyLysValProAlaGlySerAspValIleSerGluTyrVal
AAAGCCTGTGATGGAATCGGAGGAGCTATGCATAAAGCTATGCTTATGGCTCAAGCAATAACAGGAGTTGTTTTAGGAGGACAAGTTAGAACATTTGGAGGAAAATGTTATAATTGTGGT
LysAlaCysAspGlyIleGlyGlyAlaMetHisLysAlaMetLeuMetAlaGlnAlaIleThrGlyValValLeuGlyGlyGlnValArgThrPheGlyGlyLysCysTyrAsnCysGly
CAAATTGGTCACTTAAAAAAGAATTGCCCAGTCTTAAACAAACAGAATATAACTATTCAAGCAACTACAACAGGTAGAGAGCCACCTGACTTATGTCCAAGATGTAAAAAAGGAAAACAT
GlnIleGlyHisLeuLysLysAsnCysProValLeuAsnLysGlnAsnIleThrIleGlnAlaThrThrThrGlyArgGluProProAspLeuCysProArgCysLysLysGlyLysHis
TGGGCTAGTCAATGTCGTTCTAAATTTGATAAAAATGGGCAACCATTGTCGGGAAACGAGCAAAGGGGCCAGCCTCAGGCCCCACAACAAACTGGGGCATTCCCAATTCAGCCATTTGTT
TrpAlaSerGlnCysArgSerLysPheAspLysAsnGlyGlnProLeuSerGlyAsnGluGlnArgGlyGlnProGlnAlaProGlnGlnThrGlyAlaPheProIleGlnProPheVal
***LysTrpAlaThrIleValGlyLysArgAlaLysGlyProAlaSerGlyProThrThrAsnTrpGlyIleProAsnSerAlaIleCysSer
CCTCAGGGTTTTCAGGGACAACAACCCCCACTGTCCCAAGTGTTTCAGGGAATAAGCCAGTTACCACAATACAACAATTGTCCCTCACCACAAGCGGCAGTGCAGCAGTAGATTTATGTA
ProGlnGlyPheGlnGlyGlnGlnProProLeuSerGlnValPheGlnGlyIleSerGlnLeuProGlnTyrAsnAsnCysProSerProGlnAlaAlaValGlnGln***
SerGlyPheSerGlyThrThrThrProThrValProSerValSerGlyAsnLysProValThrThrIleGlnGlnLeuSerLeuThrThrSerGlySerAlaAlaValAspLeuCysThr
CTATACAAGCAGTCTCTCTGCTTCCAGGGGAGCCCCCACAAAAAATCCCCTACAGGGGTATATGGCCCACTGCCTGAGGGGACTGTAGGACTAATCTTGGGAAGATCAAGTCTAAATCTAA
IleGlnAlaValSerLeuLeuProGlyGluProProGlnLysIleProThrGlyValTyrGlyProLeuProGluGlyThrValGlyLeuIleLeuGlyArgSerSerLeuAsnLeuLys
AAGGAGTTCAAATTCATACTAGTGTGGTTGATTCAGACTATAAAGGCGAAATTCAATTGGTTATTAGCTCTTCAATTCCTTGGAGTGCCAGTCCAAGAGACAGGATTGCTCAATTATTAC
GlyValGlnIleHisThrSerValValAspSerAspTyrLysGlyGluIleGlnLeuValIleSerSerSerIleProTrpSerAlaSerProArgAspArgIleAlaGlnLeuLeuLeu
TCCTGCCATATATTAAGGGTGGAAATAGTGAAATAAAAAGAATAGGAGGCTTGTAAGCACTGATCCAACAGGAAAGGCTGCATATTGGGCAAGTCAGGTCTCAGAGAACAGACCTGTGT
LeuProTyrIleLysGlyGlyAsnSerGluIleLysArgIleGlyGlyLeuValSerThrAspProThrGlyLysAlaAlaTyrTrpAlaSerGlnValSerGluAsnArgProValCys
GTAAGGCCATTATTCAAGGAAAACAGTTTGAAGGGTTGGTAGACACTGGAGCAGATGTCTCTATTATTGCTTTAAATCAGTGGCCAAAAAACTGGCCTAAACAAAAGGCTGTTACAGGAC
LysAlaIleIleGlnGlyLysGlnPheGluGlyLeuValAspThrGlyAlaAspValSerIleIleAlaLeuAsnGlnTrpProLysAsnTrpProLysGlnLysAlaValThrGlyLeu
TTGTCGGCATAGGCACAGCCTCAGAAGTGTATCAAAGTATGGAGATTTTACATTGCTTAGGGCCAGATAATCAAGAAAGTACTGTTCAGCCAATGATTACTTCAATTCCTCTTAATCTGT
ValGlyIleGlyThrAlaSerGluValTyrGlnSerMetGluIleLeuHisCysLeuGlyProAspAsnGlnGluSerThrValGlnProMetIleThrSerIleProLeuAsnLeuTrp
GGGGTCGAGATTTATTACAACAATGGGGTGCGGAAATCACCATGCCCGCTCCATTATATAGCCCCACGAGTCAAAAAATCATGACCAAGATGGGATATATACCAGGAAAGGGACTAGGGA
GlyArgAspLeuLeuGlnGlnTrpGlyAlaGluIleThrMetProAlaProLeuTyrSerproThrSerGlnLysIleMetThrLysMetGlyTyrIleproGlyLysGlyLeuGlyLys
AAAATGAAGATGGCATTAAAGTTCCAGTTGAGGCTAAAATAAATCAAGAAAGAGAAGGAATAGGGTATCCTTTTTAGGGGCGGTCACTGTAGAGCCTCCTAAACCCATACCACTAACTTG
AsnGluAspGlyIleLysValProValGluAlaLysIleAsnGlnGluArgGluGlyIleGlyTyrProPhe***
***AsnLysSerArgLysArgArgAsnArgValSerPheLeuGlyAlaValThrValGluProProLysProIleProLeuThrTrp
GAAAACAGAAAAACCGGTGTGGGTAAATCAGTGGCCGCTACCAAAACAAAAACTGGAGGCTTTACATTTATTAGCAAATGAACAGTTAGAAAAGGGTCACATTGAGCCTTCGTTCTCACC
LysThrGluLysProValTrpValAsnGlnTrpProLeuProLysGlnLysLeuGluAlaLeuHisLeuLeuAlaAsnGluGlnLeuGluLysGlyHisIleGluProSerPheSerPro
TTGGAATTCTCCTGTGTTTGTAATTCAGAAGAAATCAGGCAAATGGCATACGTTAACTGACTTAAGGGCTGTAAACGCCGTAATTCAACCCATGGGGCCTCTCCAACCCGGGTTGCCCTC
TrpAsnSerProValPheValIleGlnLysLysSerGlyLysTrpHisThrLeuThrAspLeuArgAlaValAsnAlaValIleGlnProMetGlyProLeuGlnProGlyLeuProSer
TCCGGCCATGATCCCAAAAGATTGGCCTTTAATTATAATTGATCTAAAGGATTGCTTTTTTACCATCCCTCTGGCAGAGCAGGATTGTGAAAAATTTGCCTTTACTATACCAGCCATAAA
ProAlaMetIleProLysAspTrpProLeuIleIleIleAspLeuLysAspCysPhePheThrIleProLeuAlaGluGlnAspCysGluLysPheAlaPheThrIleProAlaIleAsn
TAATAAAGAACCAGCCACCAGGTTTCAGTGGAAAGTGTTACCTCAGGGAATGCTTAATAGTCCAACTATTTGTCAGACTTTTGTAGGTCGAGCTCTTCAACCAGTGAGAGAAAAGTTTTC
AsnLysGluProAlaThrArgPheGlnTrpLysValLeuProGlnGlyMetLeuAsnSerProThrIleCysGlnThrPheValGlyArgAlaLeuGlnProValArgGluLysPheSer
AGACTGTTATATTATTCATTATATTGATGATATTTTATGTGCTGCAGAAACGAAAGATAAATTAATTGACTGTTATACATTTCTGCAAGCAGAGGTTGCCAATGCTGGACTGGCAATAGC
AspCysTyrIleIleHisTyrIleAspAspIleLeuCysAlaAlaGluThrLysAspLysLeuIleAspCysTyrThrPheLeuGlnAlaGluValAlaAsnAlaGlyLeuAlaIleAla
ATCCGATAAGATCCAAACCTCTACTCCTTTTCATTATTTAGGGATGCAGATAGAAAATAGAAAATTAAGCCACAAAAAATAGAAATAAGAAAAGACACATTAAAAACACTAAATGATTT
SerAspLysIleGlnThrSerThrProPheHisTyrLeuGlyMetGlnIleGluAsnArgLysIleLysProGlnLysIleGluIleArgLysAspThrLeuLysThrLeuAsnAspPhe
LTR
gag
prt
RT

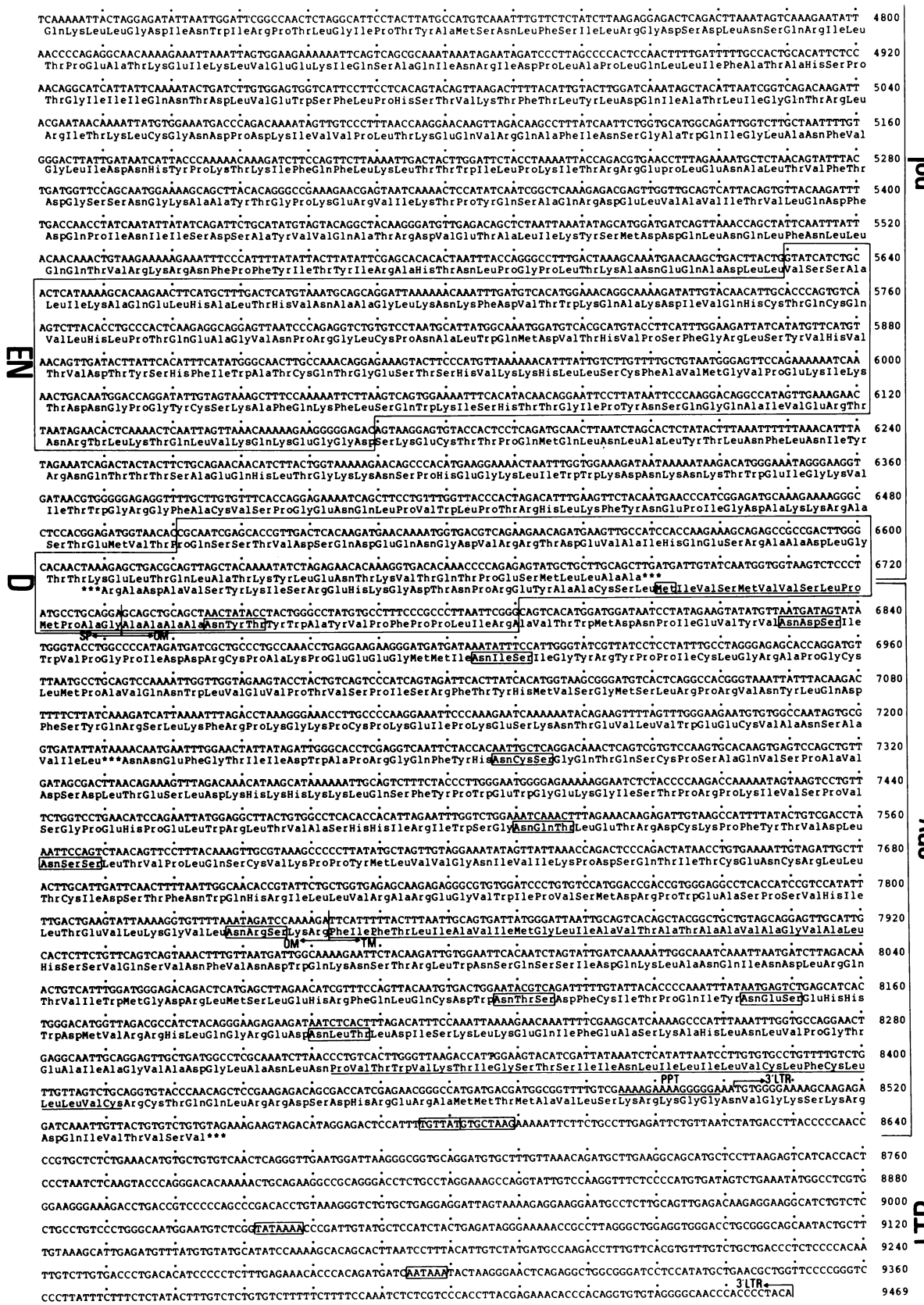

FIG. 1. Nucleotide sequence of human endogenous retrovirus genome HERV-K10(+). The DNA sequences of the coding strand, LTRs, and *gag*, protease (*prt*), *pol*, and *env* regions are shown. The sequences corresponding to the glucocorticoid responsive element (31), enhancer core (12), TATA box, and polyadenylation signal are boxed in the LTRs (24). The assumed primer-binding site (PBS) and polypurine tract (PPT) are underlined. The deduced amino acid sequences endoded by the six large open reading frames shown in Fig. 2 are given, and the first methionine in each reading frame is boxed. In the *gag* region, two amino acid sequences with the consensus sequences observed in the nucleic acid-binding protein (6, 7) are boxed. The sequence homologous to the C-terminal portion of cellular acid proteases (41) is boxed in the *prt* gene. In the *pol* gene, two conserved subregions, RT and EN (see the text), are boxed. Around the *pol-env* junction, a 290-bp sequence found in HERV-K8 but deleted in HERV-K10 is indicated in the box (D). The *env* gene is tentatively subdivided into SP, OM, and TM regions. Three hydrophobic amino acid clusters are underlined. Potential glycosylation sites in the *env* gene product are indicated in boxes.

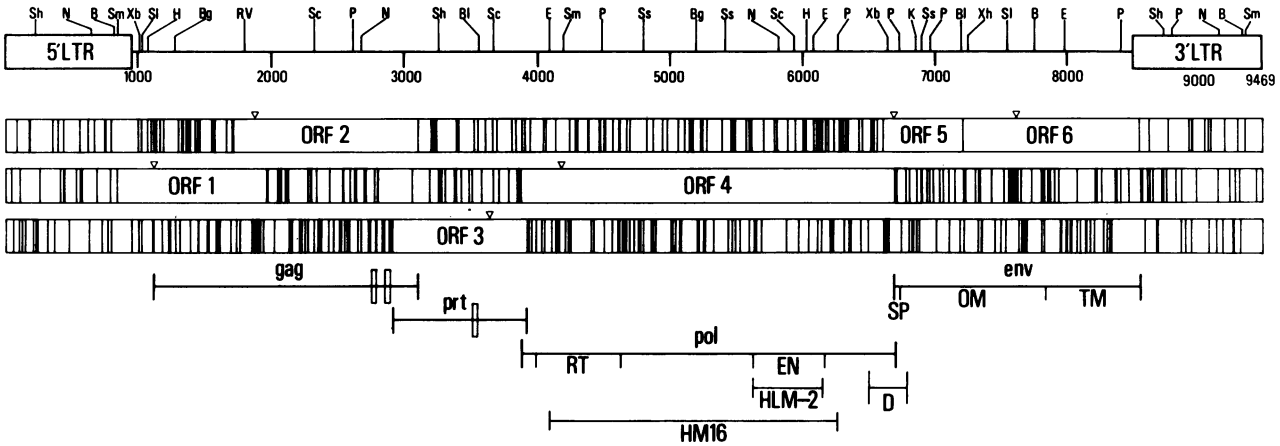

FIG. 2. Restriction map and open reading frames of human endogenous retrovirus genome HERV-K10(+). Stop codons in each phase of the coding strand are shown by vertical lines. The six large open reading frames (ORF1 to ORF6) and the *gag*, protease (*prt*), *pol*, and *env* genes are shown. The position (▽) of the first ATG codon in each reading frame is indicated. The locations corresponding to the two amino acid sequences with the consensus sequence observed in the nucleic acid-binding protein (6, 7) are boxed in the *gag* gene. The sequence homologous to the C-terminal portion of cellular acid proteases (41) is boxed in the *prt* gene. Two conserved subregions, RT and EN (see the text), are indicated in the *pol* gene. HLM-2 and HM16 are regions whose nucleotide sequences correspond to those reported by Callahan et al. (2) and Deen and Sweet (8), respectively. A 290-bp deletion (D) found in both HERV-K10 and HERV-K18 is indicated. The *env* gene is tentatively subdivided into SP, OM, and TM regions. The restriction enzyme abbreviations are as follows: Bl, *Bal*I; B, *Bam*HI; Bg, *Bgl*II; E, *Eco*RI; RV, *Eco*RV; H, *Hin*dIII; K, *Kpn*I; N, *Nsi*I; P, *Pst*I; Sl, *Sal*I; Sc, *Sca*I; Sm, *Sma*I; Se, *Spe*I; Sh, *Sph*I; Ss, *Ssp*I; Xb, *Xba*I; Xh, *Xho*I.

HERV-K22) each possess an additional 0.3-kb fragment 6.5 kb downstream from the 5′ end of HERV-K10 and HERV-K18. The HERV-K10 LTRs are 968 bp long and differ from each other by only two bases, whereas the HERV-K18 LTRs (969 bp) differ from one another in 42 positions. Therefore, it was decided to determine the complete nucleotide sequence of HERV-K10, possibly having fewer mutations introduced after integration as a provirus, followed by determination of the sequences of the 0.3-kb fragments present in HERV-K8 and HERV-K22.

HERV-K10 is 9,179 bp long with LTRs having segments corresponding to a TATA box, a polyadenylation signal, and an enhancer core (12) (Fig. 1). Immediately upstream from the putative enhancer core sequence is a segment corresponding to a glucocorticoid responsive element found in the MMTV LTR (31). A $tRNA_{1,2}^{Lys}$ with a CUU anticodon (39) and reported to be a putative primer tRNA in visna virus (37) and in the simian D-type virus SRV-1 (26) is considered to be the primer for the reverse transcription of HERV-K.

The additional sequences found in HERV-K8 and HERV-K22 are 290 bp long and are located at positions 6500 and 6501 from the 5′ end of HERV-K10 (Fig. 1). These 290-bp fragments differ from each other at only one position (6581), where A in HERV-K8 is replaced by G in HERV-K22. By comparing the amino acid sequences deduced from the nucleotide sequences of HERV-K with those of other retroviruses, it was concluded that HERV-K with the 290-bp fragment is a prototype in the HERV-K family (see below). A composite nucleotide sequence 9,469 bp long is shown in Fig. 1 as HERV-K10(+), which is made up of the complete HERV-K10 sequence and the 290-bp fragment of HERV-K8.

After translation of the HERV-K10(+) nucleotide sequence into the amino acid sequences, the positions of the stop codons in each reading frame were identified (Fig. 2). In the reading frame of the coding strand, several large open reading frames (ORF1 to ORF6) are present in the region corresponding to the *gag*, *prt*, *pol*, and *env* genes of the retrovirus genome. No open reading frame capable of encoding more than 150 amino acid residues was found on the noncoding strand.

Two-dimensional homology matrix comparisons of the nucleotide sequences of HERV-K10(+) and the Syrian hamster IAP (IAP-H18) (25) and of HERV-K10(+) and the D-type simian retrovirus SRV-1 (26) are shown in Fig. 3. The most homologous region in Fig. 3A and B was positioned in the ORF4 region of HERV-K10(+), which corresponds to the *pol* gene. The second homologous region located in ORF3 corresponds to the *prt* gene (see below).

HERV-K10(+) *gag* gene. Since the structural features and amino acid sequences of the gene products of HERV-K were totally unknown, an attempt was made to elucidate the organization of this genome by comparing the structures deduced from the nucleotide sequences of HERV-K10(+) with those deduced from the presumably closely related oncoviruses such as IAP-H18 (type A) (25), MMTV (type B) (4, 8), and SRV-1 (type D) (26). The N termini of the *gag* precursor proteins of the mammalian retroviruses so far examined have a Met followed by Gly. After removal of the N-terminal Met by a methionyl peptidase, the amino group of the Gly is modified by myristate (32). The codons for the Met-Gly residues located closest to the N terminus of ORF1 start at position 1112 of HERV-K10(+) (Fig. 1). The deduced amino acid sequence starting from these Met-Gly residues is homologous not only to the N-terminal sequence of the SRV-1 *gag* precursor protein (Fig. 4A) but also to those of IAP-H18 (25) and MMTV (10) (data not shown). Therefore, it was concluded that these Met-Gly residues are the N terminus of the HERV-K10(+) *gag* precursor protein. The deduced amino acid sequence of ORF2 is homologous to the C-terminal half of the SRV-1 *gag* precursor protein (Fig. 4A), and consequently, it was concluded that ORF1 and ORF2 form the HERV-K10(+) *gag* gene. The reading frame shift found in the HERV-K10(+) *gag* region differs so much

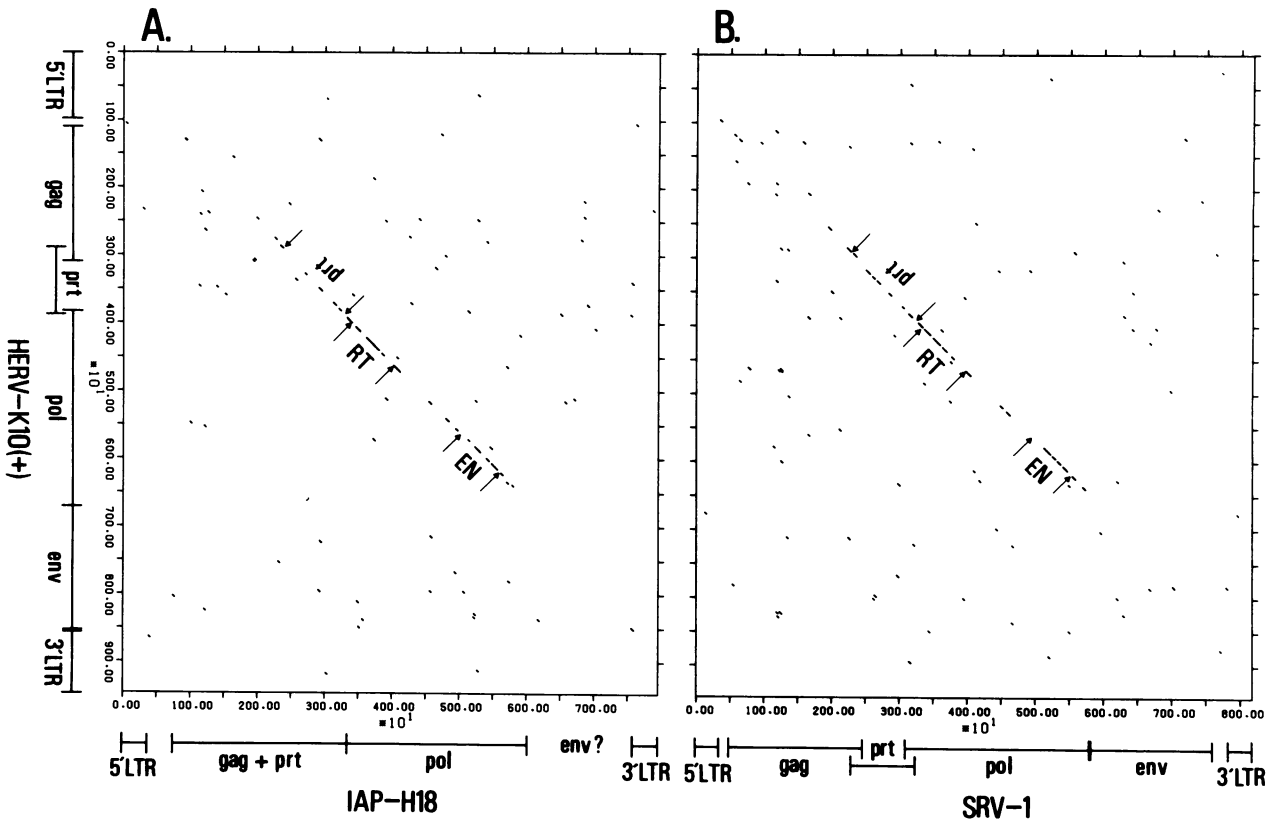

FIG. 3. Two-dimensional homology matrix comparisons of the nucleotide sequences of HERV-K10(+) and IAP-H18 (A) and HERV-K10(+) and SRV-1 (B). The nucleotide sequences of IAP-H18 and SRV-1 were taken from Ono et al. (25) and Power et al. (26), respectively. Each short line represents a region where there is at least 70% homology in 25 contiguous nucleotides between the two genes. The locations of the corresponding genes are indicated in the margins. Regions corresponding to the protease (*prt*) gene and the RT and EN subregions of the HERV-K10(+) genome are indicated in the matrices.

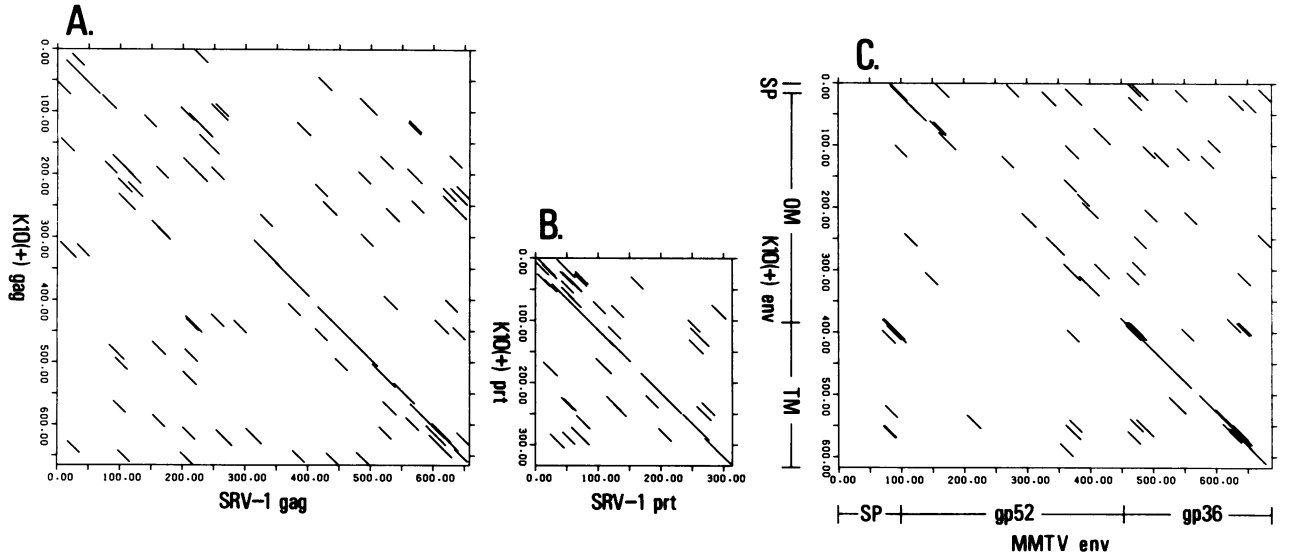

FIG. 4. Homology matrix comparisons of amino acid sequences. A computer program (40) was used to generate diagonal lines indicating segments of 20 residues with an average score of at least 65. (A) HERV-K10(+) and SRV-1 *gag* gene products; (B) HERV-K10(+) and SRV-1 *prt* gene products; (C) HERV-K10(+) and MMTV *env* gene products. The amino acid sequences of the gene products of SRV-1 and MMTV were deduced from the nucleotide sequences described by Power et al. (26) and Redmond and Dickson (27), respectively. The positions of each gene in the original nucleotide sequences are as follows: HERV-K10(+) *gag*, 1112 to 3108; HERV-K10(+) *prt*, 2913 to 3914; HERV-K10(+) *env*, 6691 to 8544; SRV-1 *gag*, 503 to 2476; SRV-1 *prt*, 2296 to 3237; MMTV *env*, 752 to 2815.

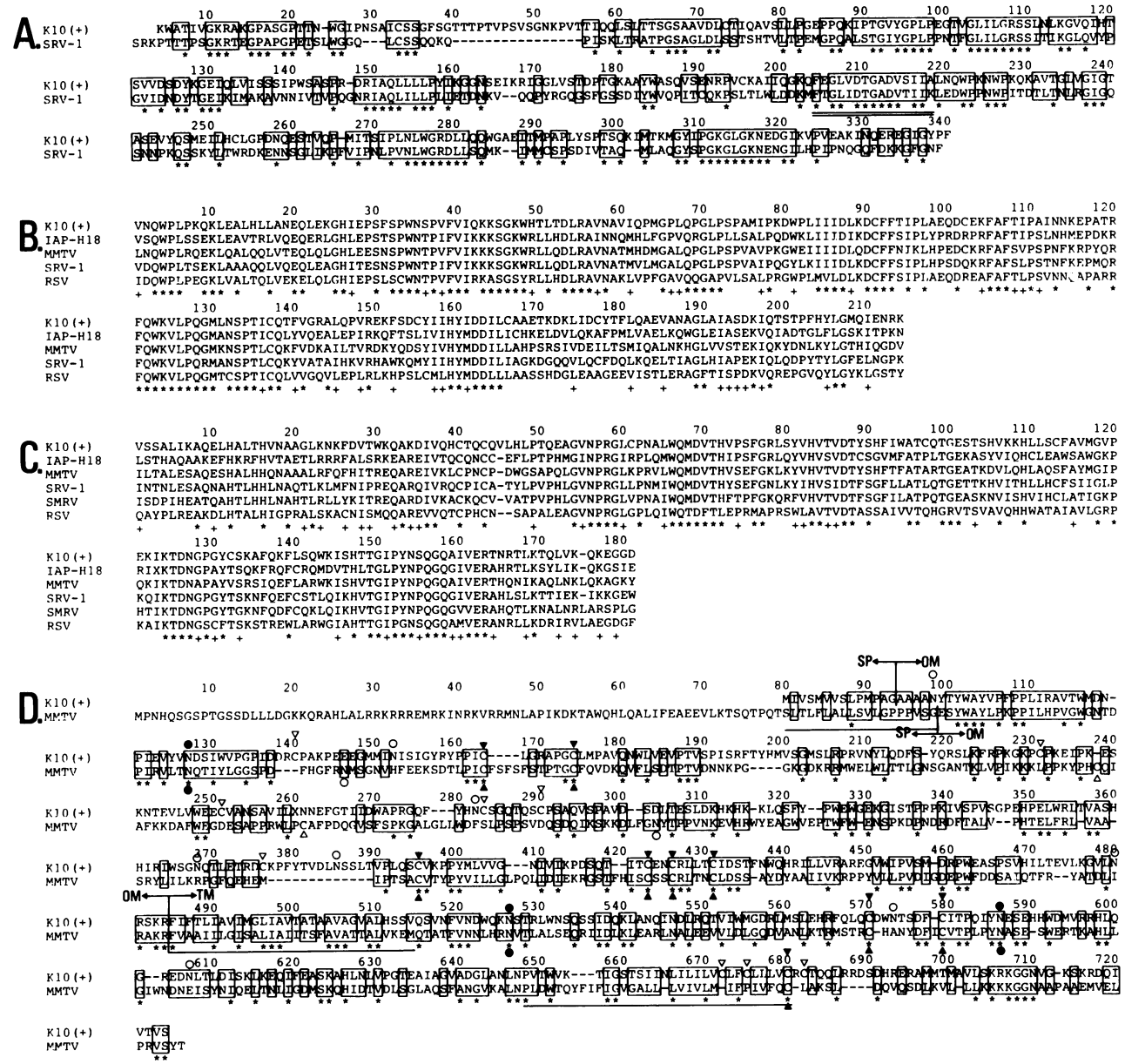

FIG. 5. Alignment of amino acid sequences of protease (*prt*), *pol*, and *env* gene products. (A) Alignment of *prt* gene products; (B) alignment of RT subregions of *pol* genes; (C) alignment of EN subregions of *pol* genes; (D) alignment of *env* gene products. Amino acids are represented by one-letter codes. Gaps (—) were inserted to make the similarity more evident. The aligned regions of the HERV-K10(+) *prt* and *env*, SRV-1 *prt*, and MMTV *env* gene products are indicated in the legend to Fig. 4. The aligned regions of IAP-H18 (25) correspond to nucleotide positions 3383 to 4024 for RT and 5046 to 5582 for EN. The aligned region of MMTV (8) corresponds to nucleotide positions 82 to 723 for RT. Amino acid sequences corresponding to the RT or EN region or both of MMTV, Rous sarcoma virus (RSV), and squirrel monkey retrovirus (SMRV) were obtained from Toh et al. (41). In panels A and D, the positions at which the aligned sequences have identical or chemically similar amino acids are boxed, and asterisks indicate the positions at which the aligned sequences have identical amino acids. Chemically similar amino acids are defined as pairs of residues belonging to the same group (33). The groups are as follows: A, S, T, P, and G; N, D, E, and Q; H, R, and K; M, L, I, and V; F, Y, and W. In panels B and C, asterisks and plusses indicate positions where aligned sequences, except that of Rous sarcoma virus, have identical or chemically similar amino acids, respectively. In panel A, the sequence homologous to the C-terminal portion of cellular acid proteases (41) is indicated by double underlines. In panel D, three hydrophobic amino acid clusters are underlined, and potential glycosylation sites (● and ○) and cysteine residues (▲ and △) are indicated; the closed symbols indicate the positions at which aligned asparagine or cysteine residues are identical.

from the reported retrovirus *gag* genes that it must have arisen either in the process of reverse transcription or as a result of postintegration mutation.

The mutually homologous region shown in Fig. 3A, made up of ca. 350 amino acid residues located in the C-terminal region, corresponds to the region encoding the major core shell structural protein and nucleic acid-binding protein. Starting at positions 2746 and 2854, two sequences of Cys-X_2-Cys-X_4-His-X_4-Cys commonly observed in the nucleic acid-binding protein region (6, 7) are present at a distance of

22 amino acid residues from each other. This is about twice the usual separation distance (7). The *gag* precursor protein of HERV-K10(+) was estimated to have 666 amino acid residues, which is almost the same number as that in the closely related SRV-1 protein (663 residues).

HERV-K10(+) *prt* gene. In addition to the *gag*, *pol*, and *env* genes, the retrovirus genome contains the gene encoding a protease (*prt*) which has a significant role in the processing of the polyprotein precursor into the mature form. ORF3 of HERV-K10(+), capable of encoding 334 amino acid residues, appears to be the *prt* gene because both its location and size are essentially the same as those of the SRV-1 *prt* gene, encoding 315 residues. The nucleotide sequence of the putative *prt* gene in HERV-K10(+) is homologous to those of SRV-1 and IAP-H18 (Fig. 3A and B), indicating that the putative *prt* gene products also resemble each other (Fig. 4B and 5A). At positions 3507 and 3551 of HERV-K10(+) (Fig. 1, 2, and 5A), spot homologies were observed for sequences in the C-terminal portions of several cellular acid proteases (41); therefore, ORF3 in HERV-K10(+) may reasonably be considered to be the *prt* gene. The size of the HERV-K10(+) *prt* gene, as well as that of the SRV-1 and IAP-H18 genes, was ca. twice the size noted in other retrovirus genomes (26).

HERV-K10(+) *pol* gene. The nucleotide sequence of ORF4, capable of encoding 937 amino acid residues, is highly homologous to both the IAP-H18 and SRV-1 *pol* genes (Fig. 3A and B); therefore, it was concluded that ORF4 in HERV-K10(+) is the *pol* gene. A comparison of the deduced amino acid sequences of the retrovirus *pol* gene products indicated two mutually conserved subregions (40, 41). One subregion (RT in Fig. 1, 2, and 3) with ca. 210 residues is located close to the N-terminal region and corresponds to the region bearing reverse transcriptase and RNase H activity in the multifunctional *pol* protein. The other subregion (EN), made up of ca. 180 amino acids, is positioned close to the C-terminal region and corresponds to the domain having endonuclease activity apparently involved in the integration of circular intermediate DNA into chromosomal DNA.

The deduced amino acid sequences corresponding to the RT and EN subregions of the *pol* gene products, which appear to be evolutionarily related (5, 25, 30, 37, 38, 41), are aligned in Fig. 5B and C, respectively. Based on these alignments, the sequence homology in each combination was calculated (Tables 1 and 2), and phylogenetic trees were constructed (Fig. 6A and B). In the EN subregion, HERV-K10(+) was significantly more homologous to MMTV (type B retrovirus) than to SRV-1 (type D retrovirus), squirrel monkey retrovirus (type D retrovirus), and IAP-H18 (type A retrovirus), whereas in the RT subregion, HERV-K10(+) was related to essentially the same extent to A-, B-, and D-type retroviruses.

TABLE 1. Amino acid sequence homology and divergence of RT subregion of *pol* gene product

Virus	Homology or divergence of[a]:				
	SRV-1 (D)[b]	MMTV (B)	HERV-K10(+)	IAP-H18 (A)	RSV[c] (avian C)
SRV-1		0.439	0.587	0.562	0.751
MMTV	0.645		0.612	0.665	0.741
HERV-K10(+)	0.556	0.542		0.621	0.665
IAP-H18	0.570	0.514	0.537		0.647
RSV	0.472	0.477	0.514	0.523	

[a] Based on the aligned amino acid sequences shown in Fig. 5B, homology and divergence were calculated (41) and are shown in the lower left and upper right halves of the table, respectively.
[b] The letters in parentheses specify retrovirus types.
[c] RSV, Rous sarcoma virus.

TABLE 2. Amino acid sequence homology and divergence of EN subregion of *pol* gene product

Virus	Homology or divergence of[a]:					
	SRV-1 (D)[b]	SMRV (D)	MMTV (B)	HERV-K10(+)	IAP-H18 (A)	RSV (avian C)
SRV-1		0.480	0.573	0.618	0.704	0.988
SMRV	0.619		0.650	0.699	0.762	0.973
MMTV	0.564	0.522		0.535	0.774	0.959
HERV-K10(+)	0.539	0.497	0.586		0.716	0.889
IAP-H18	0.494	0.467	0.461	0.489		1.082
RSV	0.372	0.378	0.383	0.411	0.339	

[a] Based on the aligned amino acid sequences shown in Fig. 5C, homology and divergence were calculated (41) and are shown in the lower left and upper right halves of the table, respectively. SMRV, Squirrel monkey retrovirus; RSV, Rous sarcoma virus.
[b] The letters in parentheses specify retrovirus types.

HERV-K10(+) *env* gene. The 290-bp fragment from positions 6501 to 6790 in HERV-K10(+) (Fig. 1) is not present in the HERV-K10 and HERV-K18 clones but is present in the HERV-K8 and HERV-K22 clones. The putative *env* region of HERV-K10(+), including the 290-bp fragment, has two open reading frames, ORF5 and ORF6. Since the nucleotide sequences corresponding to the termination codon found between ORF5 and ORF6 are CAA (Gln) in HERV-K8, HERV-K18, and HERV-K22, the TAA codon must have been generated by a C to T substitution in HERV-K10. Thus, ORF5 and ORF6 found separated in HERV-K10 are probably continuous in the prototype of the HERV-K family. Comparison of the amino acid sequences encoded by ORF5 and ORF6 with that encoded by the MMTV *env* gene (17, 27) (Fig. 4C) led to the conclusion that ORF5 and ORF6 are the HERV-K10(+) *env* gene. The deduced amino acid sequences of the HERV-K10(+) and MMTV *env* gene products are not homologous to that of the SRV-1 *env* gene product (data not shown), whereas the deduced amino acid

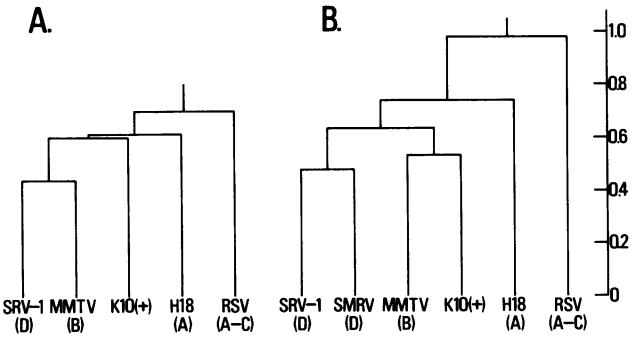

FIG. 6. Evolutionary trees of closely related oncovirus genomes. The evolutionary tree was constructed from corrected divergences (Tables 1 and 2) by the unweighted pair-group clustering method. (A) RT subregion in the *pol* gene (Fig. 5B); (B) EN subregion in the *pol* gene (Fig. 5C). The ordinate represents the average values of sequence divergences between the sequence pairs connected at each nodal point (36). The letters in parentheses at the bottom of the figure specify the following retrovirus types: A, type A; B, type B; A-C, avian type C; D, type D. These evolutionary trees agree well with others reported recently (5, 25, 30, 38, 41). RSV, Rous sarcoma virus; SMRV, squirrel monkey retrovirus.

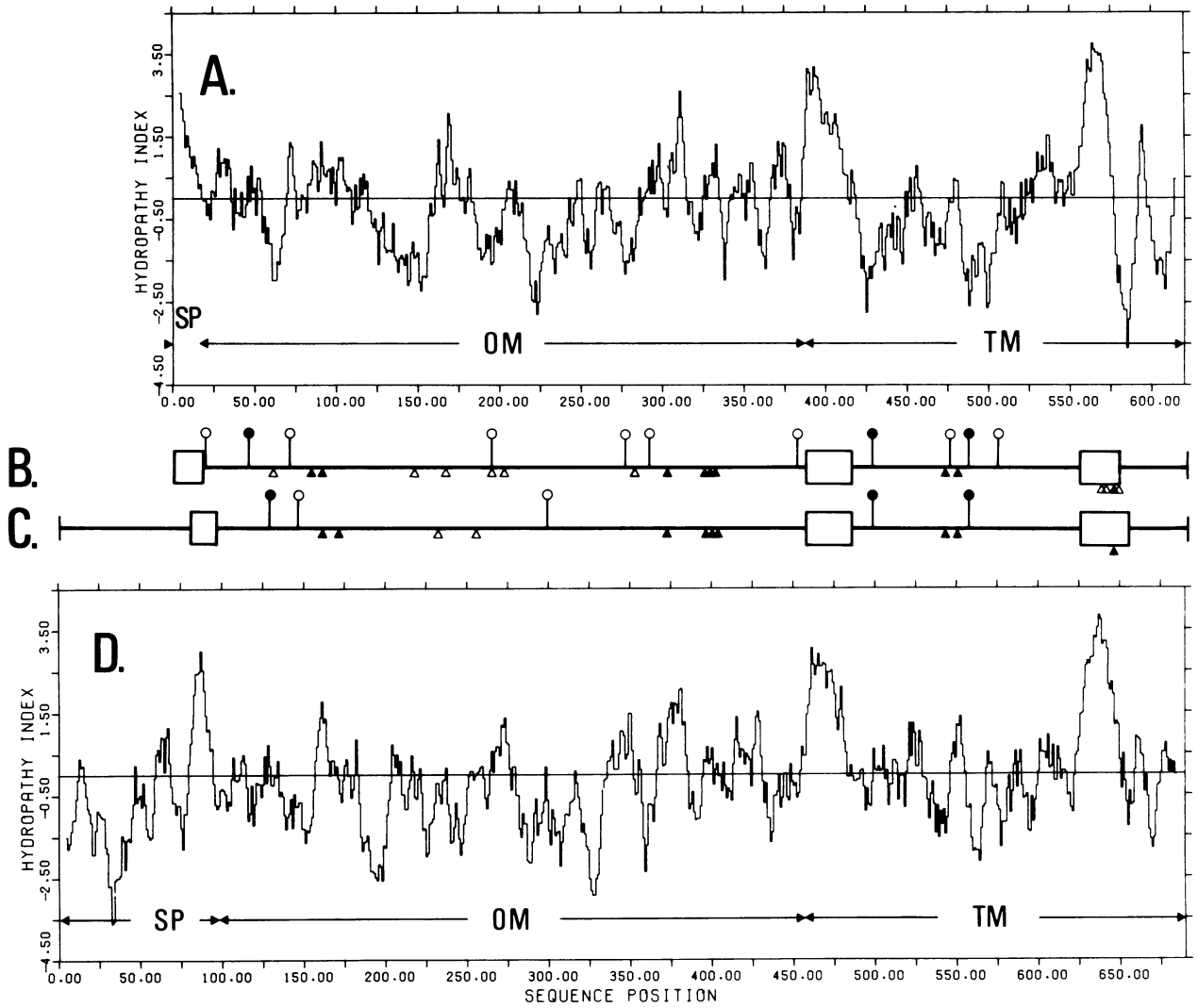

FIG. 7. Structural features of *env* gene products of HERV-K10(+) and MMTV. Hydropathy profiles of HERV-K10(+) (positions 6691 to 8544) (A) and MMTV (positions 752 to 2815 [26]) (D) *env* gene products. The profiles were drawn by the method of Kyte and Doolittle (13) with a cell length of nine amino acid residues. The positions of hydrophobic amino acid clusters (boxes), potential glycosylation sites (● and ○), and cysteine residues (▲ and △) in the HERV-K10(+) (B) and MMTV (C) *env* gene products are shown. The closed symbols indicate the positions at which aligned potential glycosylation sites or cysteine residues are identical, as shown in Fig. 5D.

sequence of the transmembrane protein region of the SRV-1 *env* gene product is significantly homologous to the corresponding sequence of the Moloney murine leukemia virus, a typical mammalian type C oncovirus (35) (our unpublished observation).

During biosynthesis of the *env* protein, proteolytic cleavages split the newly synthesized *env* precursor protein into a signal peptide (SP), an outer membrane protein (OM), and a transmembrane protein (TM). The homologous regions shown in Fig. 4C were located in the region of ca. 110 amino acid residues close to the N-terminal region and in the TM region consisting of ca. 230 residues of the HERV-K10(+) *env* gene product (Fig. 5D). Based on the aligned amino acid sequences of the MMTV and HERV-K10(+) *env* gene products (Fig. 5D), the proteolytic cleavage sites of the HERV-K10(+) *env* precursor protein were surmised. The estimated OM and TM encoded by the HERV-K10(+) *env* gene are 370 and 234 residues long, respectively, essentially the same lengths as the OM (358 residues) and TM (232 residues) encoded by the MMTV *env* gene (17, 27). Since the 14 amino acid residues starting with Met and located upstream from the putative OM encoded by the HERV-K10(+) *env* gene are rich in hydrophobic amino acids and are homologous to the sequence of the SP region encoded by the MMTV *env* gene, this region is assumed to be the SP encoded by the HERV-K10(+) *env* gene. The sequence comprising 34 amino acid residues starting from the N-terminal Met of the putative SP is not present in the HERV-K10 and HERV-K18 clones but is homologous to the sequence surrounding the SP-OM junction of the MMTV *env* gene product (Fig. 5D) (17, 27). Therefore, it was concluded that the HERV-K proviruses such as HERV-K8 and HERV-K22, each containing the additional 290-bp sequence, are prototypes of the HERV-K family.

The hydropathy profiles and positions of potential glycosylation sites and cysteine residues in the HERV-K10(+) and MMTV *env* gene products are shown in Fig. 7. Three markedly hydrophobic amino acid clusters were found, one in the *env* SP region and the other two in the *env* TM region of both HERV-K10(+) and MMTV. Seven *N*-

asparagine-type potential glycosylation sites are present in the *env* OM region of HERV-K10(+), and four are present in the TM region. For the HERV-K10(+) and MMTV *env* gene products, not only the locations of hydrophobic amino acid clusters, but also the positions of potential glycosylation sites and cysteine residues, necessary for cross-linking between OM and TM, are very much as expected (Fig. 7).

DISCUSSION

A group of human endogenous retrovirus genomes, tentatively designated HERV-K, were previously isolated from a human fetal liver gene library as sequences possessing homology to the Syrian hamster IAP *pol* gene (24). In the present research, determination of the nucleotide sequence of HERV-K10(+) indicated a close evolutionary relationship with types A, B, and D oncoviruses. This provirus was found to be most closely related to MMTV, known as a typical type B oncovirus.

Recently, human DNA sequences related to the MMTV genome have been isolated (3, 8, 42), and the local nucleotide sequences of some have been reported (2, 8). The nucleotide sequence of a clone (HLM-2) reported by Callahan et al. (2) is 525 bp long and corresponds to positions 5630 to 6151 (522 bp) of HERV-K10(+), which corresponds to the EN subregion of the *pol* gene (Fig. 1 and 2). With the HERV-K10(+) sequences as a standard, the 525-bp sequence has single-base insertions at five positions and two single-base deletions. The nucleotide sequence homology between HERV-K10(+) and HLM-2 in this region was calculated as 88%; consequently, it was concluded that HLM-2 is a member of the HERV-K family. Deen and Sweet determined the nucleotide sequence of the *pol* region of the MMTV-related human endogenous retrovirus genome HM16 (8). This sequence of 2,134 bp corresponds to positions 4084 to 6261 (2,178 bp) of HERV-K10(+) (Fig. 1 and 2). It is 44 bp shorter than the corresponding region of the HERV-K10(+) gene and has 23, 42, and 47 termination codons in the three reading frames, respectively. The nucleotide sequence of the HM16 *pol* region is 89% homologous to that of the HERV-K10(+) *pol* gene; this supports the possibility that HM16 is a member of the HERV-K family.

Even though the genes cloned from an identical gene library (14) are mutually homologous, the intactness of HERV-K10 seems to warrant attention in view of the extensive disruptions in HM16 caused by mutations. As in the case of HERV-K proviruses, multicopied endogenous retrovirus genomes appear to undergo a variety of mutations after integration into chromosomal DNA. For clarification of the organization of an endogenous retrovirus family, a presumably less mutated genome should be selected and studied. To determine the complete nucleotide sequence of a HERV-K genome, four clones with a greater number of common restriction fragments and standard size were first selected from among 26 clones. Finally, HERV-K10, appearing to have the fewest mutations introduced after integration, was selected, since the HERV-K10 LTRs (968 bp) differ by only two bases. As mentioned above, elucidation should be made of the organization of genes such as HERV-K.

During the maturation process of *env* protein, glycosylation converts OM, the MMTV *env* gene product comprising 358 amino acid residues (17, 27), into a glycoprotein, gp52, with an apparent molecular weight of 52,000. Antigens cross-reactive with polyclonal antibody raised against gp52 have been detected in human breast carcinoma tissue (23) and in the particulate fraction of human milk (9). Furthermore, it has been reported that a cell line (T47D) established from a human breast cancer produces retroviruslike particles and releases soluble glycoproteins, both with antigenicity toward the anti-gp52 antibody (11, 34). In this study, the *env* gene product of HERV-K10(+) was elucidated and found to be significantly related to the MMTV *env* gene product, not only with respect to the deduced amino acid sequence, but also in structural features. This indicates the possibility that the gp52-related antigen expressed in human breast cancer tissue and T47D cells is the *env* gene product of a potentially active HERV-K provirus.

ACKNOWLEDGMENTS

We are grateful to M. Kawakami for his encouragement throughout the course of this work and to H. Hayashida for kindly providing his computer program.

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

LITERATURE CITED

1. **Bonner, T. I., C. O'Connell, and M. Cohen.** 1982. Cloned endogenous retroviral sequences from human DNA. Proc. Natl. Acad. Sci. USA **79**:4709–4713.
2. **Callahan, R., I.-M. Chiu, J. F. H. Wong, S. R. Tronick, B. A. Roe, S. A. Aaronson, and J. Schlom.** 1985. A new class of endogenous human retroviral genomes. Science **228**:1208–1211.
3. **Callahan, R., W. Drohan, S. Tronick, and J. Schlom.** 1982. Detection and cloning of human DNA sequences related to the mouse mammary tumor virus genome. Proc. Natl. Acad. Sci. USA **79**:5503–5507.
4. **Chiu, I.-M., R. Callahan, S. R. Tronick, J. Schlom, and S. A. Aaronson.** 1984. Major *pol* gene progenitors in the evolution of oncoviruses. Science **223**:364–370.
5. **Chiu, I.-M., A. Yaniv, J. E. Dahlberg, A. Gazit, S. F. Skuntz, S. R. Tronick, and S. A. Aaronson.** 1985. Nucleotide sequence evidence for relationship of AIDS retrovirus to lentiviruses. Nature (London) **317**:366–368.
6. **Copeland, T. D., M. A. Morgan, and S. Oroszlan.** 1983. Complete amino acid sequence of the nucleic acid-binding protein of bovine leukemia virus. FEBS Lett. **156**:37–40.
7. **Covey, S. N.** 1986. Amino acid sequence homology in *gag* region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus. Nucleic Acids Res. **14**:623–633.
8. **Deen, K. C., and R. W. Sweet.** 1986. Murine mammary tumor virus *pol*-related sequences in human DNA: characterization and sequence comparison with the complete murine mammary tumor virus *pol* gene. J. Virol. **57**:422–432.
9. **Dion, A. S., D. C. Farwell, A. A. Pomenti, and A. J. Girardi.** 1980. A human protein related to the major envelope protein of murine mammary tumor virus: identification and characterization. Proc. Natl. Acad. Sci. USA **77**:1301–1305.
10. **Fasel, N., E. Buetti, J. Firzlaff, K. Pearson, and H. Diggelmann.** 1983. Nucleotide sequence of the 5′ noncoding region and part of the *gag* gene of mouse mammary tumor virus; identification of the 5′ splicing site for subgenomic mRNAs. Nucleic Acids Res. **11**:6943–6955.
11. **Keydar, I., T. Ohno, R. Nayak, R. Sweet, F. Simoni, F. Weiss, S. Karby, R. Mesa-Tejada, and S. Spiegelman.** 1984. Properties of retrovirus-like particles produced by a human breast carcinoma cell line: immunological relationship with mouse mammary tumor virus proteins. Proc. Natl. Acad. Sci. USA **81**:4188–4192.
12. **Khoury, G., and P. Gruss.** 1983. Enhancer elements. Cell **33**:313–314.
13. **Kyte, J., and R. F. Doolittle.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105–132.
14. **Lawn, R. M., E. F. Fritsch, R. C. Parker, G. Blake, and T. Maniatis.** 1978. The isolation and charactrization of linked δ- and β-globin genes from a cloned library of human DNA. Cell **15**:1157–1174.

15. **Lueders, K. K., and E. L. Kuff.** 1981. Sequence homologous to retrovirus-like genes of the mouse are present in multiple copies in the Syrian hamster genome. Nucleic Acids Res. **9**:5917–5930.
16. **Mager, D. L., and P. S. Henthorn.** 1984. Identification of a retrovirus-like repetitive element in human DNA. Proc. Natl. Acad. Sci. USA **81**:7510–7514.
17. **Majors, J. E., and H. E. Varmus.** 1983. Nucleotide sequencing of an apparent proviral copy of *env* mRNA defines determinants of expression of the mouse mammary tumor virus *env* gene. J. Virol. **47**:495–504.
18. **Martin, M. A., T. Bryan, S. Rasheed, and A. S. Khan.** 1981. Identification and cloning of endogenous retroviral sequences present in human DNA. Proc. Natl. Acad. Sci. USA **78**:4892–4896.
19. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65**:499–560.
20. **Needleman, S. B., and C. D. Wunsch.** 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**:443–453.
21. **Noda, M., M. Kurihara, and T. Takano.** 1982. Retrovirus-related sequences in human DNA: detection and cloning of sequences which hybridize with the long terminal repeat of baboon endogenous virus. Nucleic Acids Res. **10**:2865–2878.
22. **O'Connell, C. D., and M. Cohen.** 1984. The long terminal repeat sequences of a novel human endogenous retrovirus. Science **226**:1204–1206.
23. **Ohno, T., R. Mesa-Tejada, I. Keydar, M. Ramanarayanan, J. Bausch, and S. Spiegelman.** 1979. Human breast carcinoma antigen is immunologically related to the polypeptide of the group-specific glycoprotein of mouse mammary tumor virus. Proc. Natl. Acad. Sci. USA **76**:2460–2464.
24. **Ono, M.** 1986. Molecular cloning and long terminal repeat sequences of human endogenous retrovirus genes related to types A and B retrovirus genes. J. Virol. **58**:937–944.
25. **Ono, M., H. Toh, T. Miyata, and T. Awaya.** 1985. Nucleotide sequence of the Syrian hamster intracisternal A-particle gene: close evolutionary relationship of type A particle gene to types B and D oncovirus genes. J. Virol. **55**:387–394.
26. **Power, M. D., P. A. Marx, M. L. Bryant, M. B. Gardner, P. J. Barr, and P. A. Luciw.** 1986. Nucleotide sequence of SRV-1, a type D simian acquired immune deficiency syndrome retrovirus. Science **231**:1567–1572.
27. **Redmond, S. M. S., and C. Dickson.** 1983. Sequence and expression of the mouse mammary tumor virus *env* gene. EMBO J. **2**:125–131.
28. **Repaske, R., P. E. Steele, R. R. O'Neill, A. B. Rabson, and M. A. Martin.** 1985. Nucleotide sequence of a full-length human endogenous retroviral segment. J. Virol. **54**:764–772.
29. **Sagata, N., T. Yasunaga, K. Ohishi, J. Tsuzuku-Kawamura, M. Onuma, and Y. Ikawa.** 1984. Comparison of the entire genomes of bovine leukemia virus and human T-cell leukemia virus and characterization of their unidentified open reading frames. EMBO J. **3**:3231–3237.
30. **Sagata, N., T. Yasunaga, J. Tsuzuku-Kawamura, K. Ohishi, Y. Ogawa, and Y. Ikawa.** 1985. Complete nucleotide sequence of the genome of bovine leukemia virus: its evolutionary relationship to other retroviruses. Proc. Natl. Acad. Sci. USA **82**:677–681.
31. **Scheidereit, C., and M. Beato.** 1984. Contacts between hormone receptor and DNA double helix within a glucocorticoid regulatory element of mouse mammary tumor virus. Proc. Natl. Acad. Sci. USA **81**:3029–3033.
32. **Schultz, A. M., and S. Oroszlan.** 1983. In vivo modification of retroviral *gag* gene-encoded polyproteins by myristic acid. J. Virol. **46**:355–361.
33. **Schwartz, R. M., and M. O. Dayhoff.** 1978. Matrices for detecting distant relationships, p. 353–358. *In* M. O. Dayhoff (ed.), Atlas of protein sequence and structure, vol. 5, suppl. 3. National Biochemical Research Foundation, Washington, D.C.
34. **Segev, N., A. Hizi, F. Kirenberg, and I. Keydar.** 1985. Characterization of a protein, released by the T47D cell line, immunologically related to the major envelope protein of mouse mammary tumor virus. Proc. Natl. Acad. Sci. USA **82**:1531–1535.
35. **Shinnick, T. M., R. A. Lerner, and J. G. Sutcliffe.** 1981. Nucleotide sequence of Moloney murine leukemia virus. Nature (London) **293**:543–548.
36. **Socal, R. R., and P. H. A. Sneath.** 1963. Princples of numerical taxonomy. W. H. Freeman & Co., San Francisco.
37. **Sonigo, P., M. Alizon, K. Staskus, D. Klatzmann, S. Cole, O. Danos, E. Retzel, P. Tiollais, A. Haase, and S. Wain-Hobson.** 1985. Nucleotide sequence of the visna lentivirus: relationship to the AIDS virus. Cell **42**:369–382.
38. **Sonigo, P., C. Barker, E. Hunter, and S. Wain-Hobson.** 1986. Nucleotide sequence of Mason-Pfizer monkey virus: an immunosuppressive D-type retrovirus. Cell **45**:375–385.
39. **Sprinzl, M., J. Moll, F. Messner, and T. Hartmann.** 1985. Compilation of tRNA sequences. Nucleic Acids Res. **13**:1–49.
40. **Toh, H., H. Hayashida, and T. Miyata.** 1983. Sequence homology betwen retroviral reverse transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus. Nature (London) **305**:827–829.
41. **Toh, H., R. Kikuno, H. Hayashida, T. Miyata, W. Kugimiya, S. Inouye, S. Yuki, and K. Saigo.** 1985. Close structural resemblance between putative polymerase of a *Drosophila* transposable genetic element 17.6 and *pol* gene product of Moloney murine leukemia virus. EMBO J. **4**:1267–1272.
42. **Westley, B., and F. E. B. May.** 1984. The human genome contains multiple sequences of varying homology to mouse mammary tumor virus DNA. Gene **28**:221–227.