Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 3101–3105, May 1985
Biochemistry

Complete nucleotide sequence of an infectious clone of human T-cell leukemia virus type II: An open reading frame for the protease gene

(retrovirus/genome/DNA)

KUNITADA SHIMOTOHNO*, YURI TAKAHASHI*, NOBUAKI SHIMIZU*, TAKASHI GOJOBORI†, DAVID W. GOLDE‡, IRVIN S. Y. CHEN‡, MASANAO MIWA*, AND TAKASHI SUGIMURA*

*National Cancer Center Research Institute, Tsukiji, Chuo-ku, Tokyo 104, Japan; †National Institute of Genetics, Mishima, Shizuoka 411, Japan; and ‡Division of Hematology-Oncology, Department of Medicine, University of California School of Medicine, Los Angeles, CA 90024

Contributed by Takashi Sugimura, December 28, 1984

ABSTRACT **The entire nucleotide sequence of an infectious clone of human T-cell leukemia virus type II provirus was determined. This provirus consists of 8952 nucleotides. In addition to long terminal repeats and *gag*, *pol*, *env*, and *X*, a protease gene that is responsible for processing the gag precursor protein was found. The protease gene is encoded in a different frame from *gag* and *pol* and was located between the *gag* and *pol* open reading frames. The 5′ region of the protease gene overlaps the 3′ *gag* region. Coding regions of the provirus show about 60% homology with those of human T-cell leukemia virus type I at the nucleotide level. The evolutionary relationship between human T-cell leukemia virus types I and II is discussed.**

Human T-cell leukemia virus type I (HTLV-I) and human T-cell leukemia virus type II (HTLV-II) are typical exogenous human retroviruses and have some characteristics in common with other retroviruses (1–3). These two viruses are related. The *gag* proteins of these virions show immunological cross-reactivity (4). The nucleotide sequences of the *env* regions of the two viruses show about 65% homology (5), although their envelope proteins show low cross-reactivity (6). In addition, HTLV-II and HTLV-I have a sequence of about 1.6 kilobase pairs (kbp), called *X* or *pX*, between *env* and the 3′ long terminal repeat (LTR) (3, 7, 8). From comparison of this sequence in the two viruses, we and others previously predicted that this sequence might be translated (7, 8), and, in fact, proteins of 41 and 38 kDa were found to be encoded from this region in HTLV-I- and HTLV-II-infected cells, respectively (9–12).

To elucidate the functions of other regions of the HTLV-II genome, we have determined the entire nucleotide sequence of the provirus. The provirus examined was molecularly cloned from a patient (Mo) with hairy cell leukemia and was found to be replication competent (8, 13). Analysis of the nucleotide sequence indicated that the HTLV-II provirus has the structure LTR–*gag*–*protease*–*pol*–*env*–*X*-LTR in this order from the 5′ end of the genome (Fig. 1).

MATERIALS AND METHODS

DNAs and Sequencing. An infectious clone of HTLV-II provirus, λH6.0, was subcloned in pBR322 at the *Bam*HI site (13). The corresponding subclones, pH6-B5.0 and pH6-B3.5, which covered the 5′ and 3′ halves of the original provirus, respectively, were used for sequencing. The method of Maxam and Gilbert was mainly used for sequencing (14), and the M13 phage method (15) was used for sequencing part of the region of the *pol* gene.

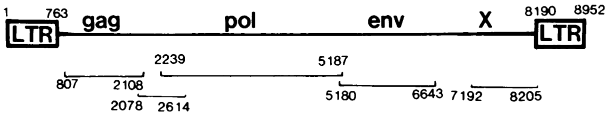


FIG. 1. Structure of HTLV-II provirus. Numbers indicate nucleotides from the 5′ end of the provirus. Bars indicate locations of open reading frames in the genome.

Materials. Radiolabeled compounds were purchased from Amersham. Restriction endonucleases, DNA polymerase, and polynucleotide kinase were from Takara Shuzo (Kyoto, Japan) or New England Biolabs.

RESULTS AND DISCUSSION

The nucleotide sequence of HTLV-II provirus consists of 8952 bases, as shown in Fig. 2. In addition to three major open reading frames, corresponding to *gag*, *pol*, and *env*, there are four large open reading frames. Three are located in the *X* region as reported previously (8) and the other is between the 3′ end of *gag* and the 5′ end of *pol*. This open reading frame was identified as the gene that codes for a protease that processes the precursor Gag protein to mature forms. The provirus has a genome structure (as shown in Fig. 4) different from that of any other retrovirus but similar to that of HTLV-I (3) and bovine leukemia virus (BLV) (16).

LTR and 5′ Noncoding Region. As we reported previously (17), the LTR has 763 bases, in which several functional domains, such as a promotor, enhancer, and terminator of transcription, are present. The LTR sequences of HTLV-I and HTLV-II show very low homology except in several stretches of oligonucleotides located in these functional domains. The $tRNA^{Pro}$ binding site is present two nucleotides downstream of the 5′ LTR. A short nucleotide sequence, 23 nucleotides, is present between the 3′ end of the tRNA binding site and the initiator of a Gag precursor protein.

***gag* Gene.** The precursor Gag protein of HTLV-I was shown to be cleaved to three peptides (18). From the amino acid sequence of the cleavage sites of the Gag precursor of HTLV-I, the proteolytic sites in the Gag precursor of HTLV-II were predicted to be localized between Phe-136 and Pro-137 and between Leu-350 and Val-351, counting from the NH_2 terminus of the Gag frame (Fig. 2). This prediction suggests that the Gag precursor is cleaved to 15-, 24-, and 9-kDa proteins, which show 55%, 85%, and 68% homology, respectively, with the corresponding proteins of HTLV-I.

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: HTLV-I and HTLV-II, human T-cell leukemia virus type I and type II; RSV, Rous sarcoma virus; Mo-MLV, Molony murine leukemia virus; BLV, bovine leukemia virus; kbp, kilobase pair(s); LTR, long terminal repeat.

```
LTR
TGACAATGGCGACTAGCCTCCCAAGCCAGCCACCCAGGGCGAGTCATCGACCCAAAAGGTCAGACCGTCTCACACAAACAATCCCAAGTAAAGGATCTGACGTCTCCCCCTTTTTTTAGG          120
AACTGAAACCACGGCCCTGACGTCCCTCCCCCCTAGGAACAGGAACAGCTCTCCAGAAAAAAATAGACCTCACCCTTACCCACTTCCCCTAGCGCTGAAAAACAAGGCTCTGACGATTAC          240
CCCCTGCCCATAAAATTTGCCTAGTCAAAATAAAAGATGCCGAGTCTATAAAAGCGCAAGGACAGTTCAGGAGGTGGCTCGCTCCCTCACCGACCCTCTGGTCACGGAGACTCACCTTGG          360
GGATCCATCCTCTCCAAGCGGCCTCGGTTGAGACGCCTTCCGTGGGACCGTCTCCCGGCCTCGGCACCTCCTGAACTGCTCCTCCCAAGGTAAGTCTCCTCTCAGGTCGAGCTCGGCTGC  5'LTR   480
CCCTTAGGTAGTCGCTCCCCGAGGGTCTTTAGAGACACCCGGGTTTCCGCCTGCGCTCGGCTAGACTCTGCCTTAAACTTCACTTCCGCGTTCTTGTCTCGTTCTTTCCTCTTCGCCGTC          600
ACTGAAAACGAAACCTCAACGCCGCCCTCTTGGCAGGCGTCCCGGGGCCAACATACGCCGTGGAGCGCAGCAAGGGCTAGGGCTTCCTGAACCTCTCCGGGAGAGGTCTATTGCTATAGG          720
                                                      LTR
CAGGCCCGCCCTAGGAGCATTGTCTTCCCGGGGAAGACAAACAATTGGGGGCTCGTCCGGGATTTGAATTCCTCCATTCTCACATTATGGGACAAATCCACGGGCTTTCCCCAACTCCAA          840
                                                     pbs                                MetGlyGlnIleHisGlyLeuSerProThrProI

TACCCAAAGCCCCCAGGGGGCTATCAACCCACCACTGGCTTAACTTTCTCCAGGCTGCTTACCGCTTGCAGCCTAGGCCCTCCGATTTCGACTTCCAGCAGCTACGACGCTTTCTAAAAC          960
leProLysAlaProArgGlyLeuSerThrHisHisTrpLeuAsnPheLeuGlnAlaAlaTyrArgLeuGlnProArgProSerAspPheAspPheGlnGlnLeuArgArgPheLeuLysL

TAGCCCTTAAAACGCCCATTTGGCTAAATCCTATTGACTACTCGCTTTTAGCTAGCCTTATCCCCAAGGGATATCCAGGAAGGGTGGTAGAGATTATAAATATCCTTGTCAAAAATCAAG        1,080
euAlaLeuLysThrProIleTrpLeuAsnProIleAspTyrSerLeuLeuAlaSerLeuIleProLysGlyTyrProGlyArgValValGluIleIleAsnIleLeuValLysAsnGlnV

TCTCCCCTAGCGCCCCCGCCGCCCCAGTTCCGACACCTATCTGCCCTACTACTACTCCTCCGCCACCTCCCCCCCTTCCCCGGAGGCCCATGTTCCCCCCCCTTACGTGGAACCCACCA        1,200
alSerProSerAlaProAlaAlaProValProThrProIleCysProThrThrThrProProProProProProProSerProGluAlaHisValProProProTyrValGluProThrT

CCACGCAATGCTTCCCTATCTTACATCCCCCAGGAGCCCCCTCAGCTCATAGGCCCTGGCAGATGAAAGACTTACAGGCCATCAAGCAGGAGGTCAGCTCCTCTGCTCTTGGCAGCCCCC        1,320
hrThrGlnCysPheProIleLeuHisProProGlyAlaProSerAlaHisArgProTrpGlnMetLysAspLeuGlnAlaIleLysGlnGluValSerSerSerAlaLeuGlySerProG

AGTTCATGCAGACCCTCCGGCTGGCGGTACAACAGTTTGACCCCACCGCCAAGGACTTACAAGATCTCCTCCAGTACCTATGCTCCTCCCTCGTAGTTTCCTTACACCATCAGCAGCTTA   gag  1,440
lnPheMetGlnThrLeuArgLeuAlaValGlnGlnPheAspProThrAlaLysAspLeuGlnAspLeuLeuGlnThrLeuCysSerSerLeuValValSerLeuHisHisGlnGlnLeuA

ACACACTAATTACCGAGGCTGAGACCCGCGGGATGACAGGCTACAACCCCATGGCAGGGCCCCTAAGAATGCAGGCTAATAACCCCGCCCAGCAAGGTCTTAGACGGGAGTACCAGAATC        1,560
snThrLeuIleThrGluAlaGluThrArgGlyMetThrGlyTyrAsnProMetAlaGlyProLeuArgMetGlnAlaAsnAsnProAlaGlnGlnGlyLeuArgArgGluThrGlnAsnL

TTTGGCTGGCTGCTTTCTCCACCCTGCCAGGCAATACCCGTGACCCCTCTTGGGCAGCTATCCTACAGGGGCTGGAGGAACCCTATTGCGCGTTCGTAGAGCGCCTTAACGTGGCCCTTG        1,680
euTrpLeuAlaALaPheSerThrLeuProGlyAsnThrArgAspProSerTrpAlaAlaIleLeuGlnGlyLeuGluGluProTyrCysAlaPheValGluArgLeuAsnValAlaLeuA

ACAACGGCCTCCCCGAGGGTACCCCCAAAGAGCCCATCTTACGTTCCCTAGCGTACTCAAACGCCAACAAAGAATGCCAAAAAATCTTACAAGCCCGCGGACACACTAACAGCCCCCTTG        1,800
spAsnGlyLeuProGluGlyThrProLysGluProIleLeuArgSerLeuAlaTyrSerAsnAlaAsnLysGluCysGlnLysIleLeuGlnAlaArgGlyHisThrAsnSerProLeuG

GGGAGATGCTCCGGACATGTCAGGCGTGGACACCCAAGGACAAAACCAAGGTCCTTGTGGTCCAACCACGGAGGCCCCCCCCACACAGCCCTGCTTTCGTTGTGGCAAGGTAGGACACT        1,920
lyGluMetLeuArgThrCysGlnAlaTrpThrProLysAspLysThrLysValLeuValValGlnProArgArgProProProThrGlnProCysPheArgCysGlyLysValGlyHisT

GGAGTCGGGACTGTACCCAGCCACGCCCCCCTCCTGGCCCCTGCCCCCTATGCCAAGATCCTTCTCACTGGAAAAGGGACTGCCCACAACTCAAACCCCCTCAGGAGGAAGGGGAACCCC        2,040
rpSerArgAspCysThrGlnProArgProProProGlyProCysProLeuCysGlnAspProSerHisTrpLysArgAspCysProGlnLeuLysProProGlnGluGluGlyGluProL

TCCTGTTGGATCTCCCTTCCACCTCAGGCACTACTGAGGAAAAAAACTCCTTAAGGGGGGAGATCTAATCTCCCCCCATCCCGATCAAGACATCTCGATACTCCCACTCATCCCCCTGCG        2,160
euLeuLeuAspLeuProSerThrSerGlyThrThrGluGluLysAsnSerLeuArgGlyGluIle
                                     GlyLysLysLeuLeuLysGlyGlyAspLeuIleSerProHisProAspGlnAspIleSerIleLeuProLeuIleProLeuAr

GCAGCAACAGCAACCAATTCTAGGGGTCCGGATCTCCGTTATGGGACAAACACCTCAGCCTACCCAAGCGCTACTTGACACAGGAGCCGACCTTACGGTTATACCCCAGACACTCGTGCC        2,280
                                                                              HisArgSerArgProTyrGlyTyrThrProAspThrArgAla
gGlnGlnGlnGlnProIleLeuGlyValArgIleSerValMetGlyGlnThrProGlnProThrGlnAlaLeuLeuAspThrGlyAlaAspLeuThrValIleProGlnThrLeuValPr   protease

CGGGCCGGTAAAGCTCCACGACACCCTGATCCTAGGCGCCAGTGGGCAAACCAACACCCAGTTCAAACTCCTCCAAACCCCCCTACACATATTCTTGCCCTTCCGAAGGTCCCCCGTTAT        2,400
ArgAlaGlyLysAlaProArgHisProAspProArgArgGlnTrpAlaAsnGlnHisProValGlnThrProProAsnProProThrHisIleLeuAlaLeuProLysValProArgTyr
oGlyProValLysLeuHisAspThrLeuIleLeuGlyAlaSerGlyGlnThrAsnThrGlnPheLysLeuLeuGlnThrProLeuHisIlePheLeuProPheArgArgSerProValIl

CCTTTCCTCCTGCCTCTTAGACACCCACAACAAATGGACCATCATTGGAAGGGACGCCCTACAACAATGCCAGGGGCTTCTATACCTCCCAGACGACCCCAGCCCCCACCAATTGCTGCC        2,520
ProPheLeuLeuProLeuArgHisProGlnGlnMetAspHisHisTrpLysGlyArgProThrThrMetProGlyAlaSerIleProProArgArgProGlnProProProIleAlaAla
eLeuSerSerCysLeuLeuAspThrHisAsnLysTrpThrIleIleGlyArgAspAlaLeuGlnGlnCysGlnGlyLeuLeuTyrLeuProAspAspProSerProHisGlnLeuLeuPr

AATAGCCACTCCAAACACCATAGGCCTCGAACACCTTCCCCCACCTCCCCAAGTGGACCAATTTCCTTTAAACCTGAGCGCCTCCAGGCCTTAAATGACCTGGTCTCCAAGGCCCTGGAG        2,640
AsnSerHisSerLysHisHisArgProArgThrProSerProThrSerProSerGlyProIleSerPheLysProGluArgLeuGlnAlaLeuAsnAspLeuValSerLysAlaLeuGlu
oIleAlaThrProAsnThrIleGlyLeuGluHisLeuProProProProGlnValAspGlnPheProLeuAsnLeuSerAlaSerArgPro

GCTGGTCACATTGAACCATACTCAGGACCAGGCAATAACCCCGTCTTCCCCGTTAAAAAACCAAATGGTAAATGGAGGTTCATTCATGACCTAAGAGCCACCAATGCCATTACTACCACC        2,760
AlaGlyHisIleGluProTyrSerGlyProGlyAsnAsnProValPheProValLysLysProAsnGlyLysTrpArgPheIleHisAspLeuArgAlaThrAsnAlaIleThrThrThr

CTCACCTCTCCTTCCCCAGGGCCCCCGATCTCACTAGCCTACCGACAGCCTTACCCCACCTACAGACCATAGATCTTACTGACGCCTTTTTCCAAATCCCCCTCCCCAAGCAGTACCAG        2,880
LeuThrSerProSerProGlyProProAspLeuThrSerLueProThrAlaLeuProHisLeuGlnThrIleAspLeuThrAspAlaPhePheGlnIleProLeuProLysGlnTyrGln

CCATACTTCGCCTTCACCATTCCCCAGCCATGTAACTATGGCCCCGGGACCAGATATGCATGGACTGTCCTTCCACAGGGGTTTAAAAACAGCCCCACCCTCTTCGAACAACAATTAGCA        3,000
ProTyrPheAlaPheThrIleProGlnProCysAsnTyrGlyProGlyThrArgTyrAlaTrpThrValLeuProGlnGlyPheLysAsnSerProThrLeuPheGluGlnGlnLeuAla

GCCGTCCTCAACCCCATGAGGAAAATGTTTCCCACATCGACCATTGTCCAATACATGGATGACATACTTTTAGCCAGCCCCACCAATGAGGAATTACAACAACTCTCCCAGCTAACCCTC        3,120
AlaValLeuAsnProMetArgLysMetPheProThrSerThrIleValGlnTyrMetAspAspIleLeuLeuAlaSerProThrAsnGluGluLeuGlnGlnLeuSerGlnLeuThrLeu

CAGGCACTGACCACGCATGGCCTTCCAATTTCCCAGGAAAAAACACAACAAACCCCAGGCCAAATACGCTTCTTAGGACAGGTCATCTCCCCTAATCACATTACATATGAGAGTACCCCT   pol  3,240
GlnAlaLeuThrThrHisGlyLeuProIleSerGlnGluLysThrGlnGlnThrProGlyGlnIleArgPheLueGlyGlnValIleSerProAsnHisIleThrTyrGluSerThrPro

ACTATTCCCATAAAATCCCAATGGACACTCACTGAATTACAAGTTATCCTAGGAGAGATCCAGTGGGTCTCTAAAGGAACACCCATCCTTCGCAAACACCTACAATCCCTATATTCTGCC        3,360
ThrIleProIleLysSerGlnTrpThrLeuThrGluLeuGlnValIleLeuGlyGluIleGlnTrpValSerLysGlyThrProIleLeuArgLysHisLeuGlnSerLeuTyrSerAla

CTTCACGGGTACCGGGACCCAAGAGCTTGTATCACCCTCACCCCACAACAACTCCATGCGTTACATGCCATTCAACAAGCTCTACAACATAACTGCCGTGGCCGCCTCAACCCCGCCCTA        3,480
LeuHisGlyTyrArgAspProArgAlaCysIleThrLeuThrProGlnGlnLeuHisAlaLeuHisAlaIleGlnGlnAlaLeuGlnHisAsnCysArgGlyArgLeuAsnProAlaLeu

CCTCTCCTTGGCCTCATCTCGTTAAGTACATCTGGTACAACATCTGTCATCTTTCAACCCAAGCAAAATTGGCCCCTGGCTTGGCTCCACACCCCCCACCCTCCGACCAGTTTATGTCCT        3,600
ProLeuLeuGlyLeuIleSerLeuSerThrSerGlyThrThrSerValIlePheGlnProLysGlnAsnTrpProLeuAlaTrpLeuHisThrProHisProProThrSerLeuCysPro

TGGGGTCACCTACTGGCCTGCACCATCTTAACTCTAGACAAATATACCCTACAACATTATGGCCAGCTCTGCCAATCTTTCCACCACAACATGTCAAAGCAAGCCCTTTGCGACTTCCTG        3,720
TrpGlyHisLeuLeuAlaCysThrIleLeuThrLeuAspLysTyrThrLeuGlnHisTyrGlyGlnLeuCysGlnSerPheHisHisAsnMetSerLysGlnAlaLeuCysAspPheLeu

AGGAACTCCCCTCATCCAAGTGTCGGCATCCTCATTCACCACATGGGTCGATTCCATAACCTTGGCAGCCAACCGTCTGGTCCGTGGAAGACTCTCTTACACCTCCCAACCCTTCTCCAG        3,840
ArgAsnSerProHisProSerValGlyIleLeuIleHisHisMetGlyArgPheHisAsnLeuGlySerGlnProSerGlyProTrpLysThrLeuLeuHisLeuProThrLeuLeuGln

GAACCACGACTCCTCAGGCCAATTTTCACCCTCTCCCCCGTCGTGCTTGACACGGCCCCCTGCCTTTTTTCCGATGGCTCCCCTCAAAAGGCAGCGTACGTTCTCTGGGACCAGACTATC        3,960
GluProArgLeuLeuArgProIlePheThrLeuSerProValValLeuAspThrAlaProCysLeuPheSerAspGlySerProGlnLysAlaAlaTyrValLeuTrpAspGlnThrIle

CTTCAACAGGACATCACTCCCCTGCCCTCTCACGAAACACATTCCGCACAAAAGGGGGAGCTCCTTGCACTTATCTGTGGACTACGTGCTGCCAAGCCATGGCCTTCCCTTAACATCTTT        4,080
LeuGlnGlnAspIleThrProLeuProSerHisGluThrHisSerAlaGlnLysGlyGluLeuLeuAlaLeuIleCysGlyLeuArgAlaAlaLysProTrpProSerLeuAsnIlePhe

TTAGACTCTAAATATTTAATCAAATACCTACATTCCCTCGCCATTGGGGCCTTCCTCGGCACTTCCTCCCATCAAACCCTCCAGGCGGCCTTGCCACCCCTACTGCAGGGCAAGACCATC        4,200
LeuAspSerLysTyrLeuIleLysTyrLeuHisSerLeuAlaIleGlyAlaPheLeuGlyThrSerAlaHisGlnThrLeuGlnAlaAlaLeuProProLeuLeuGlnGlyLysThrIle

TACCTCCACCATGTCCGCAGCCACACCAACCTCCCCGACCCAATTTCCACCTTCAATGAATACACAGACTCCCTTATCTTAGCTCCCCTTGTTCCCCTGACGCCCCAAGGCCTCCACGGC        4,320
TyrLeuHisHisValArgSerHisThrAsnLeuProAspProIleSerThrPheAsnGluTyrThrAspSerLeuIleLeuAlaProLeuValProLeuThrProGlnGlyLeuHisGly
```

(Fig. 2 continues on the next page.)

```
CTCACCCATTGCAATCAAAGGGCTCTAGTCTCTTTTGGCGCCACACCAAGGGAAGCCAAGTCCCTTGTACAGACTTGCCATACCTGTCAAACCATCAACTCACAACATCATATGCCTCGA     4,440
LeuThrHisCysAsnGlnArgAlaLeuValSerPheGlyAlaThrProArgGluAlaLysSerLeuValGlnThrCysHisThrCysGlnThrIleAsnSerGlnHisHisMetProArg

GGGTACATTCGCCGGGGCCTCTTGCCCAACCACATATGGCAAGGTGATGTAACCCATTATAAGTACAAAAAATACAAATACTGCCTCCACGTCTGGGTAGACACCTTCTCCGGTGCGGTT     4,560
GlyTyrIleArgArgGlyLeuLeuProAsnHisIleTrpGlnGlyAspValThrHisTyrLysTyrLysLysTyrLysTyrCysLeuHisValTrpValAspThrPheSerGlyAlaVal

TCCGTCTCCTGTAAAAAGAAAGAAACCAGCTGTGAGACTATCAGCGCCGTTCTTCAGGCCATTTCCCTCCTAGGGAAACCACTCCACATTAACACAGATAATGGGCCAGCCTTCCTATCA     4,680
SerValSerCysLysLysLysGluThrSerCysGluThrIleSerAlaValLeuGlnAlaIleSerLeuLeuGlyLysProLeuHisIleAsnThrAspAsnGlyProAlaPheLeuSer  pol

CAAGAATTCCAGGAGTTTTGTACCTCCTATCGCATCAAGCATTCTACCCATATACCATACAACCCCACCAGCTCAGGCCTGGTCGAGAGAACCAATGGTGTAATCAAAAACTTACTAAAT     4,800
GlnGluPheGlnGluPheCysThrSerTyrArgIleLysHisSerThrHisIleProTyrAsnProThrSerSerGlyLeuValGluArgThrAsnGlyValIleLysAsnLeuLeuAsn

AAATATCTACTAGACTGTCCTAACCTTCCCCTAGACAATGCCATTCACAAAGCCCTTTGGACTCTCAATCAGCTAAATGTCATGAACCCCAGTGGTAAAACCCGATGGCAAATCCACCAC     4,920
LysTyrLeuLeuAspCysProAsnLeuProLeuAspAsnAlaIleHisLysAlaLeuTrpThrLeuAsnGlnLeuAsnValMetAsnProSerGlyLysThrArgTrpGlnIleHisHis

AGTCCTCCACTACCACCCATTCCTGAAGCCTCTACCCCTCCCAAACCACCTCCCAAATGGTTCTATTATAAACTCCCCGGCCTTACCAATCAGCGGTGGAAAGGTCCATTGCAATCCCTC     5,040
SerProProLeuProProIleProGluAlaSerThrProProLysProProProLysTrpPheTyrTyrLysLeuProGlyLeuThrAsnGlnArgTrpLysGlyProLeuGlnSerLeu
  ↓
CAGGAAGCGGCCGGGGCAGCCTTGCTCTCCATAGACGGCTCCCCCGGTGGATCCCGTGGCGATTCCTGAAAAAAGCTGCATGCCCAAGACCAGACGCCAGCGAACTCGCCGAGCACGCC     5,160
GlnGluAlaAlaGlyAlaAlaLueLeuSerIleAspGlySerProArgTrpIleProTrpArgPheLeuLysLysAlaAlaCysProArgProAspAlaSerGluLeuAlaGluHisAla

GCAACAGACCACCAACACCATGGGTAATGTTTTCTTCCTACTTTTATTCAGTCTCACACATTTTCCACTAGCCCAGCAGAGCCGATGCACACTCACGATTGGTATCTCCTCCTACCACTC     5,280
AlaThrAspHisGlnHisHisGly
           MetGlyAsnValPhePheLueLeuLeuPheSerLeuThrHisPheProLeuAlaGlnGlnSerArgCysThrLeuThrIleGlyIleSerSerTyrHisSe

CAGCCCCTGTAGCCCAACCCAACCCGTCTGCACGTGGAACCTCGACCTTAATTCCCTAACAACGGACCAACGACTACACCCCCCCTGCCCTAACCTAATTACTTACTCTGGCTTCCATAA     5,400
rSerProCysSerProThrGlnProValCysThrTrpAsnLeuAspLeuAsnSerLeuThrThrAspGlnArgLeuHisProProCysProAsnLeuIleThrTyrSerGlyPheHisLy

GACTTATTCCTTATACTTATTCCCACATTGGATAAAAAAGCCAAACAGACAGGGCCTAGGGTACTACTCGCCTTCCTACAATGACCCTTGCTCGCTACAATGCCCCTACTTGGGCTGCCA     5,520
sThrTyrSerLeuTyrLeuPheProHisTrpIleLysLysProAsnArgGlnGlyLeuGlyTyrTyrSerProSerTyrAsnAspProCysSerLeuGlnCysProTyrLeuGlyCysGl

AGCATGGACATCCGCATACACGGGCCCCGTCTCCAGTCCATCCTGGAAGTTTCATTCAGATGTAAATTTCACCCAGGAAGTCAGCCAAGTGTCCCTTCGACTACACTTCTCTAAGTGCGG     5,640
nAlaTrpThrSerAlaTyrThrGlyProValSerSerProSerTrpLysPheHisSerAspValAsnPheThrGlnGluValSerGlnValSerLeuArgLeuHisPheSerLysCysGl

CTCCTCCATGACCCTCCTAGTAGATGCCCCTGGATATGATCCTTTATGGTTCATCACCTCAGAACCCACTCAGCCTCCACCAACTTCTCCCCCATTGGTCCATGACTCCGACCTTGAACA     5,760
ySerSerMetThrLeuLeuValAspAlaProGlyTyrAspProLeuTrpPheIleThrSerGluProThrGlnProProProThrSerProProLeuValHisAspSerAspLeuGluHi

TGTCCTAACCCCCTCCACGTCCTGGACGACCAAAATACTCAAATTTATCCAGCTGACCTTACAGAGCACCAATTACTCCTGCATGGTTTGCGTGGATAGATCCAGCCTCTCATCCTGGCA  env  5,880
sValLeuThrProSerThrSerTrpThrThrLysIleLeuLysPheIleGlnLeuThrLeuGlnSerThrAsnTyrSerCysMetValCysValAspArgSerSerLeuSerSerTrpHi

TGTACTCTACACCCCCAACATCTCCATTCCCCAACAAACCTCCTCCCGAACCATCCTCTTTCCTTCCCTTGCCCTGCCCGCTCCTCCATCCCAACCCTTCCCTTGGACCCATTGCTACCA     6,000
sValLeuTyrThrProAsnIleSerIleProGlnGlnThrSerSerArgThrIleLeuPheProSerLeuAlaLeuProAlaProProSerGlnProPheProTrpThrHisCysTyrGl

ACCTCGCCTACAGGCGATAACAACAGATAACTGCAACAACTCCATTATCCTCCCCCCTTTTTCCCTCGCTCCCGTACCTCCTCCGGCGACAAGACGCCGCCGTGCCGTTCCAATAGCAGT     6,120
nProArgLeuGlnAlaIleThrThrAspAsnCysAsnAsnSerIleIleLeuProProPheSerLeuAlaProValProProProAlaThrArgArgArgArgAlaValProIleAlaVa
                                                                                                             ↑
GTGGCTTGTCTCCGCCCTAGCGGCCGGAACAGGTATCGCTGGTGGAGTAACAGGCTCCCTATCTCTGGCTTCCAGTAAAAGCCTTCTCCTCGAGGTTGACAAAGACATCTCCCACCTTAC     6,240
lTrpLeuValSerAlaLeuAlaAlaGlyThrGlyIleAlaGlyGlyValThrGlySerLeuSerLeuAlaSerSerLysSerLeuLeuLeuGluValAspLysAspIleSerHisLeuTh

CCAGGCCATAGTCAAAAATCATCAAAACATCCTCCGGGTTGCACAGTATGCAGCCCAAAATAGACGAGGATTAGACCTCCTATTCTGGGAACAAGGGGGTTTGTGCAAGGCCATACAGGA     6,360
rGlnAlaIleValLysAsnHisGlnAsnIleLeuArgValAlaGlnTyrAlaAlaGlnAsnArgArgGlyLeuAspLeuLeuPheTrpGluGlnGlyGlyLeuCysLysAlaIleGlnGl

GCAATGTTGCTTCCTCAACATCAGTAACACTCATGTATCCGTCCTCCAGGAACGGCCCCCTCTTGAAAAACGTGTCATCACCGGCTGGGGACTAAACTGGGATCTTGGACTGTCCCAATG     6,480
uGlnCysCysPheLeuAsnIleSerAsnThrHisValSerValLeuGlnGluArgProProLeuGluLysArgValIleThrGlyTrpGlyLeuAsnTrpAspLeuGlyLeuSerGlnTr

GGCACGAGAAGCCCTCCAGACAGGCATAACCATTCTCGCTCTACTCCTCCTCGTCATATTGTTTGGCCCCTGTATCCTCCGCCAAATCCAGGCCCTTCCACAGCGGTTACAAAACCGACA     6,600
pAlaArgGluAlaLeuGlnThrGlyIleThrIleLeuAlaLeuLeuLeuLeuValIleLeuPheGlyProCysIleLeuArgGlnIleGlnAlaLeuProGlnArgLeuGlnAsnArgHi

TAACCAGTATTCCCTTATCAACCCAGAAACCATGCTATAATAGACCTGCTAGCTTCTGCAGCAAATCCCCTAGGTTCGTCCCCCTACCATTGACCCATCCACAGTCCTCTATACCAGATG     6,720
sAsnGlnTyrSerLeuIleAsnProGluThrMetLeu

AGTCGCCCCCGATGTCCAGCCCTAACTCGATTCTGAATAATTGCCTCAAATAGTTCCTCTAACCCCCGCTCACATTCCTCCCATAGGACCTTCTTTTCCCCTTCAGGAAATCCACATAAC     6,840

CCTGAAGCAAGTCACAAAACCCATCAAAACCCAGGAGTCCTATACACTCCAACTGCTGATGCCTTTCTTCCCTCTCCCGGCGCTTTTGATCCTTTTCCCGCAGGCGCTCCTTTCTGCGCC     6,960

GCTCCCGCTCCTCACGCTCCTGCAGAAGTTTTAAGATCTCCCGCTGCTCCTCCGCCAACAGTCTCCGACGAGAGTCTCGCACCTGCTCGCTGACCGATCCCGACCCCAGAGGGCGACCTT     7,080

TTGCTGTCCTTCTCGGTTCCTCTCCAGGGGGAGGGACACCAGATGTCAGACTCGCCTCTCCCTGGTCTCCTAACGGCAATCTCCTAAAATAGTCTAAAAAATCACACATAATTACAATCC     7,200
                                                                                                               LeuGlnSer

TGTCTCCTCTCAGCCCATTTCCTAGGATTTGGACAGAGCCTCCTATATGGATACCCCGTCTACGTGTTTGGCGATTGTGTACAGGCCGATTGGTGTCCCGTCTCAGGTGGTCTATGTTCC     7,320
CysLeuLeuSerAlaHisPheLeuGlyPheGlyGlnSerLeuLeuTyrGlyTyrProValTyrValPheGlyAspCysValGlnAlaAspTrpCySProValSerGlyGlyLeuCysSer

ACCCGCCTACATCGACATGCCCTCCTGGCCACCTGTCCAGAGCACCAACTCACCTGGGACCCCATCGATGGACGCGTTGTCAGCTCTCCTCTCCAATACCTTATCCCTCGCCTCCCCTCC     7,440
ThrArgLeuHisArgHisAlaLeuLeuAlaThrCysProGluHisGlnLeuThrTrpAspProIleAspGlyArgValValSerSerProLeuGlnTyrLeuIleProArgLeuProSer

TTCCCCACCCAGAGAACCTCAAGGACCCTCAAGGTCCTTACCCCTCCCACCACTCCTGTCTCCCCCAAGGTTCCACCTGCCTTCTTTCAATCAATGCGAAAGCACACCCCCTACCGAAAT     7,560
PheProThrGlnArgThrSerArgThrLeuLysValLeuThrProProThrThrProValSerProLysValProProAlaPhePheGlnSerMetArgLysHisThrProTyrArgAsn

GGATGCCTGGAACCAACCCTCGGGGATCAGCTCCCCTCCCTCGCCTTCCCCGAACCTGGCCTCCGTCCCCAAAACATCTACACCACCTGGGGAAAAACCGTAGTATGCCTATACCTATAC  x  7,680
GlyCysLeuGluProThrLeuGlyAspGlnLeuProSerLeuAlaPheProGluProGlyLeuArgProGlnAsnIleTyrThrThrTrpGlyLysThrValValCysLeuTyrLeuTyr

CAGCTTTCCCCACCCATGACATGGCCACTTATACCCCATGTCATATTCTGCCACCCCAGACAATTAGGAGCCTTCCTCACCAAGGTGCCTCTAAAACGATTAGAAGAACTTCTATACAAA     7,800
GlnLeuSerProProMetThrTrpProLeuIleProHisValIlePheCysHisProArgGlnLeuGlyAlaPheLeuThrLysValProLeuLysArgLeuGluGluLeuLeuTyrLys

ATGTTCCTACACACAGGGACAGTCATAGTCCTCCCGGAGGACGACCTACCCACCACAATGTTCCAACCCGTGAGGGCTCCCTGTATCCAGACTGCCTGGTGTACAGGACTTCTCCCCTAT     7,920
MetPheLeuHisThrGlyThrValIleValLeuProGluAspAspLeuProThrThrMetPheGlnProValArgAlaProCysIleGlnThrAlaTrpCysThrGlyLeuLeuProTyr

CACTCCATCTTAACAACCCCAGGTCTAATATGGACCTTCAATGACGGCTCACCAATGATTTCCGGCCCTTACCCCAAAGCAGGGCAGCCATCTTTAGTAGTTCAGTCCTCCCTATTAATC     8,040
HisSerIleLeuThrThrProGlyLeuIleTrpThrPheAsnAspGlySerProMetIleSerGlyProTyrProLysAlaGlyGlnProSerLeuValValGlnSerSerLeuLeuIle

TTCGAAAAATTCGAAACCAAAGCCTTCCATCCCTCCTATCTACTCTCTCATCAGCTTATACAATACTCCTCCTTCCATAACCTTCACCTTCTATTCGATGAATACACCAACATCCCTGTC     8,160
PheGluLysPheGluThrLysAlaPheHisProSerTyrLeuLeuSerHisGlnLeuIleGlnTyrSerSerPheHisAsnLeuHisLeuLeuPheAspGluTyrThrAsnIleProVal
                            ┌──LTR
TCTATTTTATTTAATAAAGAAGAGGCGGATGACAATGGCGACTAGCCTCCCGAGCCAGCCACCCAGGGCGAGTCATCGACCCAAAAGGTCAGACCGTCTCACACAAACAATCCCAAGTAA     8,280
SerIleLeuPheAsnLysGluGluAlaAspAspAsnGlyAsp

AGGCTCTGACGTCTCCCCCTTTTTTTAGGAACTGAAACCACGGCCCTGACGTCCCTCCCCCCTAGGAACAGGAACAGCTCTCCAGAAAAAAATAGACCTCACCCTTACCCACTTCCCCTA  3'LTR  8,400

GCGCTGAAAAACAAGGCTCTGACGATTACCCCCTGCCCATAAAATTTGCCTAGTCAAAATAAAAGATGCCGAGTCTATAAAAGCGCAAGGACAGTTCAGGAGGTGGCTCGCTCCCTCACC     8,520

GACCCTCTGGTCACGGAGACTCACCTTGGGGATCCATCCTCTCCAAGCGGCCTCGGTTGAGACGCCTTCCGTGGGACCGTCTCCCGGCCTCGGCACCTCCTGAACTGCTCCTCCCAAGGT     8,640

AAGTCTCCTCTCAGGTCGAGCTCGGCTGCCCCTTAGGTAGTCGCTCCCCGAGGGTCTTTAGAGACACCCGGGTTTCCGCCTGCGCTCGGCTAGACTCTGCCTTAAACTTCACTTCCGCGT     8,760

TCTTGTCTCGTTCTTTCCTCTTCGCCGTCACTGAAAACGAAACCTCAACGCCGCCCTCTTGGCAGGCGTCCCGGGGCCAACATACGCCGTGGAGCGCAGCAAGGGCTAGGGCTTCCTGAA     8,880
                                                                  LTR──┐
CCTCTCCGGGAGAGGTCTATTGCTATAGGCAGGCCCGCCCTAGGAGCATTGTCTTCCCGGGGAAGACAAACA                                                         8,952
```

FIG. 2. Complete nucleotide sequence of HTLV-II provirus. Amino acid sequences in the corresponding open reading frames are shown. LTRs and the primer binding site (pbs) are also indicated. The arrows at nucleotides 449 and 5043 indicate putative splice donor and acceptor sites, respectively. The arrows in the *gag* and the *env* genes show putative proteolytic cleavage sites. The underline in the *pol* gene shows the region to which the *pol* sequence of Moloney murine leukemia virus (Mo-MLV) shows the highest homology.

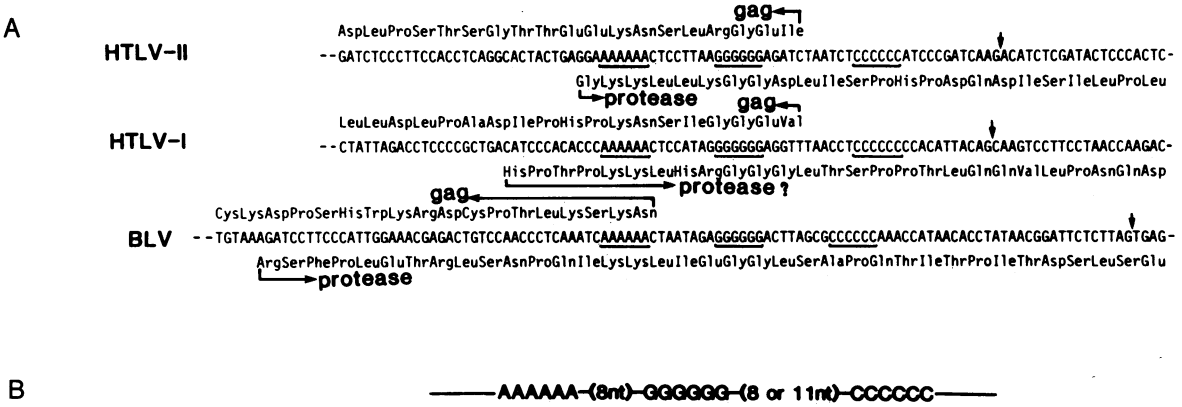
```
GGAAAAAAACTCCTTAAGGGGGGAGATCTAATCTCCCCCCATCCCGATCAAGACATCTCGATACTCCCACTCATCCCCCTGCGGCAGCAACAGCAACCAATTCTAGGG   108
GlyLysLysLeuLeuLysGlyGlyAspLeuIleSerProHisProAspGlnAspIleSerIleLeuProLeuIleProLeuArgGlnGlnGlnGlnProIleLeuGly

GTCCGGATCTCCGTTATGGGACAAACACCTCAGCCTACCCAAGCGCTACTTGACACAGGAGCCGACCTTACGGTTATACCCCAGACACTCGTGCCCGGCCGCTAAAG   216
ValArgIleSerValMetGlyGlnThrProGlnProThrGlnAlaLeuLeuAspThrGlyAlaAspLeuThrValIleProGlnThrLeuValProGlyProValLys

CTCCACGACACCCTGATCCTAGGCGCCAGTGGGCAAACCAACACCCAGTTCAAACTCCTCCAAACCCCCCTACACATATTCTTGCCCTTCCGAAGGTCCCCCCTTATC   324
LeuHisAspThrLeuIleLeuGlyAlaSerGlyGlnThrAsnThrGlnPheLysLeuLeuGlnThrProLeuHisIlePheLeuProPheArgArgSerProValIle

CTTTCCTCCTGCCTCTTAGACACCCACAACAAATGGACCATCATTGGAAGGGACGCCCTACAACAATGCCAGGGGCTTCTATACCTCCCAGACGACCCCAGCCCCCAC   432
LeuSerSerCysLeuLeuAspThrHisAsnLysTrpThrIleIleGlyArgAspAlaLeuGlnGlnCysGlnGlyLeuLeuTyrLeuProAspAspProSerProHis

CAATTGCTGCCAATAGCCACTCCAAACACCATAGGCCTCGAACACCTTCCCCCACCTCCCCAAGTGGACCAATTTCCTTTAAACCTGAGCGCCTCCAGGCCT        534
GlnLeuLeuProIleAlaThrProAsnThrIleGlyLeuGluHisLeuProProProProGlnValAspGlnPheProLeuAsnLeuSerAlaSerArgPro
```

FIG. 3. Amino acid sequence encoded from the open reading frame between *gag* and *pol* of HTLV-II. Underlines show where amino acid sequences are the same as in the proteases of Mo-MLV (---) and RSV (—) when the alignment is made for the best matches of the amino acid sequences. Numbers indicate nucleotides from the first base of the open reading frame. The first base corresponds to nucleotide 2078 in Fig. 2.

Protease Gene. A protease that is responsible for cleaving a precursor Gag protein is coded in a gene of retroviruses. This protease gene is at the 3′ end of the *gag* frame in avian retroviruses, such as Rous sarcoma virus (RSV), while it is at the 5′ end of the *pol* frame in murine retroviruses (19, 20). There is no detectable homology of the amino acid sequence in the 3′ region of *gag* or in the 5′ region of *pol* of HTLV-II provirus with the sequence of the protease domain of avian or murine retroviruses. However, an open reading frame from nucleotide 2078 to nucleotide 2611 located between the 3′ *gag* and 5′ *pol* frames can encode a sequence of 178 amino acids, which shows significant homology with the proteases of RSV and Mo-MLV (Fig. 3). There are several amino acid sequence clusters that are identical with some amino acid sequences of the protease domains of RSV or Mo-MLV. In BLV, an open reading frame located in a similar position between *gag* and *pol* was found to be a potential protease encoding gene (21), and it was confirmed that this open reading frame encodes a protein (S. Oroszlan, personal communication). A similar amino acid sequence was seen in the region between the *gag* and *pol* genes of HTLV-I, although this sequence was split by terminators and also showed a frame shift (3).

Since subgenomic mRNA of HTLV-II has not been characterized, it is not known how mRNA for the protease is synthesized. Because the first methionine residue in the frame is located at position 42 (underlined in the protease gene in Fig. 2) from the NH_2 terminus of this frame, it is unlikely that this methionine residue functions as an initiator of translation. The protease may be translated as a fused protein such as "Gag-protease" as in RSV. If this is so, rearrangement of the viral mRNA, involving a splicing or frameshift mechanism to make the mRNA, should be considered. In this regard, it is noteworthy that a putative splice acceptor site is present at nucleotide 2129 as indicated by the arrow in Fig. 4. Splice acceptor sites are also found at similar positions in HTLV-I and BLV. However, a possible splice donor site is not present close upstream from the acceptor site, but there is one at nucleotide 1910, 219 nucleotides from the acceptor site. If this is the case, a p19–p24–Δp15–protease precursor protein could be the translation product. Another possibility to be considered is that a frame shift occurs to suppress the termination of Gag. In yeast, some tRNAs are known to recognize a codon with four bases and cause a frame shift of the gene (22). There may be a similar mechanism in HTLV causing a frame shift from the *gag* gene to the protease gene, producing a Gag-protease fused protein. In fact, a longer *gag* product was found to be produced in BLV-infected cells (23). Interestingly, A-A-A-A-A-A, G-G-G-G-G-G, and C-C-C-C-C-C clusters are localized within the overlapping region of *gag* and the 5′ region of the protease gene and within the protease gene. These sequences may be important for rearrangement of viral mRNA to allow translation of the protease.

***pol* Gene.** The largest open reading frame for *pol* is located from nucleotide 2239 to nucleotide 5184 and can encode 982 amino acids. There is no homology of the first sequence of *pol* of HTLV-II, encoding 124 amino acids, with that of the protease gene of Mo-MLV, but the following sequence shows significant homology with the reverse transcriptase domain of Mo-MLV *pol* (data not shown). Therefore, the function of the first 124 amino acid residues is unclear, and this region may be lost during processing of the Pol precursor protein, or it may not be translated to amino acids because of rearrangement of mRNA for *pol* gene expression. In the COOH-terminal region of Pol, that is the nuclease domain, there is a consensus sequence, Gly-Lys-Pro-Leu-His-Ile-Asn-Thr-Asp-Asn-Gly-Pro-Ala-Phe-Leu-Ser, specific to A, B, D, and avian type C retroviruses but not to mammalian type C retrovirus (24). The amino acid sequences of the Pol



```
HTLV-II
AspLeuProSerThrSerGlyThrThrGluGluLysAsnSerLeuArgGlyGluIle
--GATCTCCCTTCCACCTCAGGCACTACTGAGGAAAAAAACTCCTTAAGGGGGGAGATCTAATCTCCCCCCATCCCGATCAAGACATCTCGATACTCCCACTC-
GlyLysLysLeuLeuLysGlyGlyAspLeuIleSerProHisProAspGlnAspIleSerIleLeuProLeu

HTLV-I
LeuLeuAspLeuProAlaAspIleProHisProLysAsnSerIleGlyGlyGluVal
--CTATTAGACCTCCCCGCTGACATCCCACACCCAAAAAACTCCATAGGGGGGGAGGTTTAACCTCCCCCCCCACATTACAGCAAGTCCTTCCTAACCAAGAC-
HisProThrProLysLysLeuHisArgGlyGlyGlyLeuThrSerProProThrLeuGlnGlnValLeuProAsnGlnAsp

BLV
CysLysAspProSerHisTrpLysArgAspCysProThrLeuLysSerLysAsn
--TGTAAAGATCCTTCCCATTGGAAACGAGACTGTCCAACCCTCAAATCAAAAAACTAATAGAGGGGGGACTTAGCGCCCCCCAAACCATAACACCTATAACGGATTCTCTTAGTGAG-
ArgSerPheProLeuGluThrArgLeuSerAsnProGlnIleLysLysLeuIleGluGlyGlyLeuSerAlaProGlnThrIleThrProIleThrAspSerLeuSerGlu

——AAAAAA-(8nt)-GGGGGG-(8 or 11nt)-CCCCCC——
```

FIG. 4. (*A*) Overlapping regions of *gag* and protease genes of HTLV-II. Sequences of the corresponding regions of HTLV-I (3) and BLV (16) are also listed. Vertical arrows indicate possible splice acceptor sites. The position of the arrow in HTLV-II indicates nucleotide 2129 in Fig. 2. Underlines show common oligonucleotides present in this region. (*B*) Consensus sequence present in the overlapping region of the *gag* and protease gene. nt, Nucleotides.

proteins of HTLV-I and -II show 61% homology when aligned for the best match of amino acid sequences (data not shown). However, the amino acid sequence from residue 117 to residue 281 (underlined in the *pol* gene in Fig. 2) in this open reading frame, which is thought to correspond to part of the domain of reverse transcriptase or RNase H, shows 82% homology.

***env* Gene.** The *env* gene, from nucleotide 5180 to nucleotide 6637, codes for 486 amino acids. There are five possible N-glycosylation sites. Four of these are located in the same positions in the Env proteins of HTLV-I and HTLV-II. These four glycosylation sites are located in the surface glycoprotein. The Env precursor protein of HTLV-I was cleaved to a surface glycoprotein, gp52, and a membrane protein, gp20 (E). The putative proteolytic site in the Env protein of HTLV-II is present between amino acid 308 and 309 from the NH_2 terminus of the *env* frame, leaving backbones of 35 and 19 kDa of those protein moieties (Fig. 2). The amino acid sequence homologies of the surface glycoproteins and membrane proteins of the two retroviruses are 63% and 73%, respectively. In the transmembrane protein, cysteine residues (at positions 389, 396, and 397 from the NH_2 terminus of the Env frame) are well conserved in retroviruses. These cysteine residues were suggested to be involved in S—S bridges between surface glycoprotein and transmembrane protein (25). Truncated mRNA for Env protein may be synthesized by splicing between a putative donor site at nucleotide 449 in 5′ LTR and a putative acceptor site at nucleotide 5043, which is located near the 3′ end of *pol* in HTLV-II (Fig. 2).

***X* Gene.** As previously reported, the *X* region has three open reading frames (8). However, one open reading frame, termed *Xc*, was found to be translated to a protein with a molecular mass of 38 kDa (9, 12). This may be compared with a protein of 41 kDa that was found in HTLV-I-infected cells (9–12). These proteins are translated from spliced mRNAs. The splice acceptor site for the mRNA was identified at nucleotide 7213 (26).

Evolutionary Relationship Between HTLV-I and HTLV-II. Nucleotide substitution, which occurs during evolution, is classified into synonymous (silent) or nonsynonymous (amino acid-altering) substitution, depending upon whether the nucleotide substitution causes an amino acid change or not. Comparing the nucleotide sequences of the *gag*, *pol*, *env*, and *X* genes in HTLV-II with those in HTLV-I, we estimated the number of synonymous and nonsynonymous substitutions for the four genes. As shown in Table 1, it is clear that for all four genes the number of synonymous changes is much larger than that of nonsynonymous changes. This suggests that, in all genes of the HTLV genome, amino acid changes are functionally constrained.

Comparing the numbers of synonymous substitutions among the four genes of HTLV, we found that the number of synonymous changes for the *X* gene is roughly less than half of those for the other three genes, *gag*, *pol*, and *env*. The number of synonymous substitutions per nucleotide site for both *gag* and *pol* genes is larger than 2.6, whereas that for the *X* gene is only 1.3. (Note that the number of synonymous substitutions for the *env* gene, 2.33, could be an underestimate.) One explanation of this observation is that the *X* gene may have been transduced into the HTLV genome quite recently. If this is the case, the number of synonymous substitutions for the *X* gene will be smaller than that for the other genes in HTLV, unless the rate of synonymous substitution varies a lot with the gene. Another explanation is that some constraints, other than amino acid changes, exist in the *X* gene. One of the most likely constraints may be the secondary structure of the *X* gene in the RNA viral genome. At present, both explanations seem to be equally plausible.

Table 1. Numbers of synonymous and nonsynonymous substitutions per nucleotide site for the genes of HTLV-II and HTLV-I

Gene	No. codons compared	Substitutions per site*	
		Synonymous	Nonsynonymous
gag	421	3.41	0.18
pol	895	2.66	0.28
env	482	2.33†	0.22
X	206	1.33	0.16

*For all genes except *env*, the numbers of synonymous and nonsynonymous substitutions were estimated by the method of Miyata and Yasunaga (27).

†Since the method of Miyata and Yasunaga was inapplicable to the synonymous substitutions for *env*, the method of Perler *et al.* (28) was used for the estimation.

We are grateful to Dr. N. Kato for helpful discussion about splice sites. This work was partly supported by the Grants-in-Aid for Cancer Research from the Ministry of Education, Science and Culture and from the Ministry of Health and Welfare of Japan and by a Grant-in-Aid from the Ministry of Health and Welfare for Comprehensive 10-Year Strategy for Cancer Control, Japan.

1. Reitz, M. S., Poiesz, B. J., Ruscetti, F. M. & Gallo, R. C. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 1887–1891.
2. Yoshida, M., Miyoshi, I. & Hinuma, Y. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 2031–2035.
3. Seiki, M., Hattori, S., Hirayama, Y. & Yoshida, M. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3618–3622.
4. Kalyanaraman, V. S., Sarngadharan, M. G., Robert-Guroff, M., Miyoshi, I., Blayney, D., Golde, D. W. & Gallo, R. C. (1982) *Science* **218**, 571–573.
5. Sodroski, J., Patarca, R., Perkins, D., Briggs, D., Lee, T.-H., Essex, M., Coligan, J., Wong-Staal, F., Gallo, R. C. & Haseltine, W. A. (1984) *Science* **225**, 421–424.
6. Clapham, P., Nagy, K. & Weiss, R. A. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 2886–2889.
7. Haseltine, W. A., Sodroski, J., Patarca, R., Briggs, D., Perkins, D. & Wong-Staal, F. (1984) *Science* **225**, 419–421.
8. Shimotohno, K., Wachsman, W., Takahashi, Y., Golde, D. W., Miwa, M., Sugimura, T. & Chen, I. S. Y. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 6657–6661.
9. Miwa, M., Shimotohno, K., Hoshino, H., Fujino, M. & Sugimura, T. (1984) *Gann* **75**, 752–755.
10. Kiyokawa, T., Seiki, M., Imagawa, K., Shimizu, F. & Yoshida, M. (1984) *Gann* **75**, 747–751.
11. Lee, T. H., Coligan, J. E., Sodroski, J. G., Haseltine, W. A., Salahuddin, S. Z., Wong-Staal, F., Gallo, R. C. & Essex, M. (1984) *Science* **226**, 57–61.
12. Slamon, D. J., Shimotohno, K., Cline, M. J., Golde, D. W. & Chen, I. S. Y. (1984) *Science* **226**, 61–64.
13. Chen, I. S. Y., McLaughlin, J., Gasson, J. C., Clark, S. C. & Golde, D. W. (1983) *Nature (London)* **305**, 502–505.
14. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
15. Schreir, P. H. & Cortese, R. (1979) *J. Mol. Biol.* **129**, 169–172.
16. Sagata, N., Yasunaga, T., Tsuzuku-Kawamura, J., Ohishi, K., Ogawa, Y. & Ikawa, Y. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 677–681.
17. Shimotohno, K., Golde, D. W., Miwa, M., Sugimura, T. & Chen, I. S. Y. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 1079–1083.
18. Oroszlan, S., Copeland, T. D., Kalyanaraman, V. S., Sarngadharan, M. G., Schultz, A. M. & Gallo, R. C. (1984) in *Human T-Cell Leukemia/Lymphoma Virus*, eds. Gallo, R. C., Essex, M. E. & Gross, L. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 101–110.
19. Schwartz, D. E., Tizard, R. & Gilbert, W. (1983) *Cell* **32**, 853–869.
20. Levin, J. G., Hu, S. C., Rein, A., Messer, L. I. & Gerwin, B. I. (1984) *J. Virol.* **51**, 470–478.
21. Sagata, N., Yasunaga, T. & Ikawa, Y. (1984) *FEBS Lett.* **178**, 79–82.
22. Roth, J. R. (1981) *Cell* **24**, 601–602.
23. Mamoun, R. Z., Astier, T., Guillemain, B. & Duplan, J. F. (1983) *J. Gen. Virol.* **64**, 1895–1905.
24. Chiu, I.-M., Callahan, R., Tronick, S. R., Schlom, J. & Aaronson, S. A. (1984) *Science* **223**, 364–370.
25. Lenz, J., Crowther, R., Sfraceski, A. & Haseltine, W. (1982) *J. Virol.* **42**, 519–529.
26. Wachsman, W., Shimotohno, K., Clark, S. C., Golde, D. W. & Chen, I. S. Y. (1984) *Science* **226**, 177–179.
27. Miyata, T. & Yasunaga, T. (1980) *J. Mol. Evol.* **16**, 23–36.
28. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dogson, J. (1980) *Cell* **20**, 555–566.