Vol. 139, No. 1, 1986 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
August 29, 1986 Pages 129-135

IDENTIFICATION OF A PROTEASE GENE OF HUMAN T-CELL LEUKEMIA VIRUS TYPE I (HTLV-I) AND ITS STRUCTURAL COMPARISON

Seok Hyun Nam and Masakazu Hatanaka

Institute for Virus Research, Kyoto University, Kyoto 606, Japan

Received July 14, 1986

SUMMARY: We determined the nucleotide sequence of a region between the gag and pol genes of a replication-competent proviral clone of a human T-cell leukemia virus type I (HTLV-I) from MT-2 cells. This region overlapping the gag and pol genes contains an open reading frame with a different phase from others. The deduced amino acid sequences show significant homology with the known protease gene of other retroviruses, and harbors highly conserved amino acid sequences that are well conserved in other retroviral protease domains. These results indicate that this open reading frame encodes a HTLV-I protease.

We have been studying the genomic structure of HTLV-I provirus and its functional role on cellular transforming mechanisms in a human T-lymphocyte cell line, MT-2 (1). The MT-2 cell was established by co-cultivation of peripheral blood lymphocytes of an ATL patient with normal cord leukocytes and produces the infectious HTLV-I virions (2). Southern blotting analysis revealed that at least eight proviruses are integrated in the chromosomes of MT-2 cells; among them, one provirus contains all information for the virus replication (3) and its transfection into human cells revealed the infectivity (4). As the post-translational processing of the viral structural protein is known to be necessary for the assembly of the virion, the involvement of a virus encoded Gag-protease has been recently suggested (5). Previously, Seiki et al, molecularly cloned a provirus of HTLV-I and sequenced the entire proviral genome (6). However, no continuous reading frame for a possible protease-

0006-291X/86 $1.50

coding gene was found in the nucleotide sequence of HTLV-I between the gag and pol genes which corresponds to the protease-coding region of human T-cell leukemia virus type II (HTLV-II), bovine leukemia virus (BLV), Rous sarcoma virus (RSV), Molony murine leukemia virus (M-MLV) and feline leukemia virus (FeLV)(7-11,13,22).

In this communication, we show that a replication-competent proviral clone of HTLV-I from MT-2 cells encodes a protease in an independent open reading frame between the gag and pol genes. The molecular structure of the protease of HTLV-I was compared to the ones of other retroviruses from the sequence analysis.

MATERIALS AND METHODS

Materials . The restriction endonucleases, T4 ligase, Klenow fragment of DNA polymerase I and M13 sequence kit were purchased from Takara shuzo, Co., Ltd. Alpha-^{32}P-dCTP (3000 Ci/mol) was obtained from Amersham, England and New England Nuclear, U.S.A.. Other enzymes and chemicals were described previously (12).

Nucleotide sequence determination . The replication-competent HTLV-I provirus DNA was cloned from MT-2 cells and termed HTLV1C (4). The appropriate fragments from HTLV1C DNA were sequenced by the procedure of Sanger et al.(14) with slight modification (15).

Homology matrix comparison and sequence alignment. Homology searches between amino acid sequences were made using the graphical matrix method, as described previously (16). With the aid of these matrices, the sequences were aligned by manual inspection.

RESULTS AND DISCUSSION

For analysis of nucleotide sequence of protease gene, a recombinant phage of the non-defective provirus was digested with SstI and SmaI. A 851 bp fragment produced by double-digestion of the restriction enzymes contained the gag-pol junction region. This fragment was re-digested with appropriate restriction enzymes and subcloned into the polylinker site of M13 phage (Fig. 1). As shown in Fig. 2, this region was found to contain an open reading frame of 702 bp that is on a different phase from

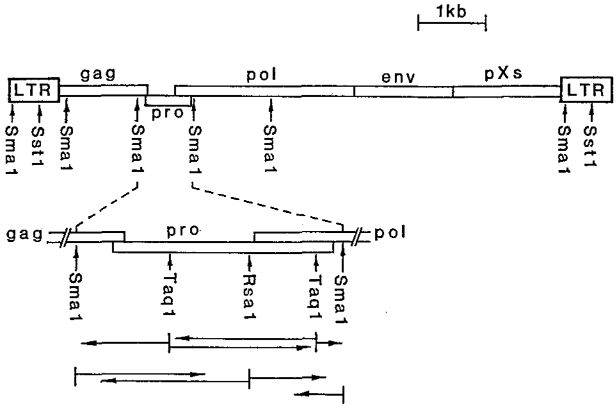

Fig. 1. Restriction map and sequencing strategy for the putative protease-coding region of HTLV1C. Arrows indicate the direction and extent of sequence analysis of individual fragment. The sequence is represented 5' to 3' from left to right.

```
           CCG.CCT.AAA.GCC.CAC.TAT.CCC.AGA.ACC.AGA.GCC.AGA.GGA.AGA.TGC.CCT.CCT.ATT.AGA.CCT
                       +--> protease
HTLV-1ATK [                 C
HTLV 1C   [CCC.CGC.TGA.CAT.TCC.ACA.CCC.AAA.AAA.CTC.CAT.AGG.GGG.GGA.GGT.TTA.ACC.TCC.CCC.CCC
                       His-Ser-Thr-Pro-Lys-Lys-Leu-His-Arg-Gly-Gly-Gly-Leu-Thr-Ser-Pro-Pro
                                                        gag <--+
                                C   C
           ACA.TTA.CAG.CAA.GTC.TTT.CTT.AAC.CAA.GAC.CCA.GCA.TCT.ATT.CTG.CCA.GTT.ATA.CCG.TTA
           Thr-Leu-Gln-Gln-Val-Phe-Leu-Asn-Gln-Asp-Pro-Ala-Ser-Ile-Leu-Pro-Val-Ile-Pro-Leu

           GAT.CCC.GCC.CGT.CGG.CCC.GTA.ATT.AAA.GCC.CAG.GTT.GAC.ACC.CAG.ACC.AGC.CAC.CCA.AAG
           Asp-Pro-Ala-Arg-Arg-Pro-Val-Ile-Lys-Ala-Gln-Val-Asp-Thr-Gln-Thr-Ser-His-Pro-Lys

                            T
           ACT.ATC.GAA.GCT.CTA.CTA.GAT.ACA.GGA.GCA.GAC.ATG.ACA.GTC.CTT.CCG.ATA.GCC.TTG.TTC
           Thr-Ile-Glu-Ala-Leu-Leu-Asp-Thr-Gly-Ala-Asp-MET-Thr-Val-Leu-Pro-Ile-Ala-Leu-Phe

                                O
           TCA.AGT.AAT.ACT.CCC.CTC.AAA.AAT.ACA.TCC.GTA.TTA.GGG.GCA.GGG.GGC.CAA.ACC.CAA.GAT
           Ser-Ser-Asn-Thr-Pro-Leu-Lys-Asn-Thr-Ser-Val-Leu-Gly-Ala-Gly-Gly-Gln-Thr-Gln-Asp

           CAC.TTT.AAG.CTC.ACC.TCC.CTT.CCT.GTG.CTA.ATA.CGC.CTC.CCT.TTC.CGG.ACA.ACG.CCT.ATT
           His-Phe-Lys-Leu-Thr-Ser-Leu-Pro-Val-Leu-Ile-Arg-Leu-Pro-Phe-Arg-Thr-Thr-Pro-Ile
                                                            *
                                                            A                   T
           GTT.TTA.ACA.TCT.TGC.CTA.GTT.GAT.ACC.AAA.AAC.AAC.TGG.GCC.ATC.ATA.GGT.CGC.GAT.GCC
           Val-Leu-Thr-Ser-Cys-Leu-Val-Asp-Thr-Lys-Asn-Asn-Trp-Ala-Ile-Ile-Gly-Arg-Asp-Ala

                            A                                   A
           TTA.CAA.CAA.TGC.CAG.GGC.GTC.CTG.TAC.CTC.CCT.GAG.GCA.AAA.GGG.CCG.CCT.GTA.ATC.TTG
           Leu-Gln-Gln-Cys-Gln-Gly-Val-Leu-Tyr-Leu-Pro-Glu-Ala-Lys-Gly-Pro-Pro-Val-Ile-Leu
                                                   +--> pol
                                            G
           CCA.ATA.CAG.GCG.CCA.GCC.GTC.CTT.GGA.CTA.GAA.CAC.CTC.CCA.AGG.CCC.CCC.CAA.ATC.AGC
           Pro-Ile-Gln-Ala-Pro-Ala-Val-Leu-Gly-Leu-Glu-His-Leu-Pro-Arg-Pro-Pro-Gln-Ile-Ser

           CAG.TTC.CCT.TTA.AAC.CAG.AAC.GCC.TCC.AGG.CCT.TGC.AAC.ACT.TGG.TCC.GGA.AGG.CCC.TGG
           Gln-Phe-Pro-Leu-Asn-Gln-Asn-Ala-Ser-Arg-Pro-Cys-Asn-Thr-Trp-Ser-Gly-Arg-Pro-Trp

                                C                       G
           AGG.CAG.GCC.ATA.TCG.AAT.CCT.ACA.CCG.GGC.CAG.GAA.ATA.ACC.CAG.TAT.TCC.CAG.TTA.AAA
           Arg-Gln-Ala-Ile-Ser-Asn-Pro-Thr-Pro-Gly-Gln-Glu-Ile-Thr-Gln-Tyr-Ser-Gln-Leu-Lys
                                                               protease <--+
           AGG.CCA.ATG.GAA.CCT.GGC.GAT.TCA.TCC.ACG.ACC.TGC.GGG.CCA.CTA.ACT.CTC.TAA.CCA.TAG
           Arg-Pro-MET-Glu-Pro-Gly-Asp-Ser-Ser-Thr-Thr-Cys-Gly-Pro-Leu-Thr-Leu-***

            T
           ACC.TCT.CAT.CAT.CTT.CCC.CCG.GGC.CCC.CTG.ACT.TGT.CCA.GCC.TGC.CAA.CCA.CAC.
```

Fig. 2. Nucleotide and translated amino acid sequences of the open reading frame of the putative protease-coding gene. Only the different nucleotides are shown in HTLV-I ATK. * amber stop codon in the sequence of HTLV-I ATK. O:nucleotide insertion in the revised sequence of the clone of HTLV-I ATK (19).

the ones of gag and pol genes, spanning nucleotide No.2052 to 2753 according to the sequence numbering by Seiki et al., (6). This open reading frame can encode a sequence of 234 amino acids which shows a significant homology with the proteases of HTLV-II and BLV (Fig.3). Moreover, there are two highly conserved amino acid regions that are identical with those of other retroviral protease domains (17). The fact that the deduced amino acid sequence bears a significant homology with other retroviral proteases suggest that this open reading frame possibly encodes a HTLV-I protease. Among these highly conserved amino acid regions, the second region contains a close similar amino acid sequence around the active site amino acids of the acid protease family, suggesting that the HTLV-I protease possesses a similar activity with those of acid proteases (17,23). One additional

```
HTLV-I    HSTPK KLHRG  GGLTS PPTLQ  QVF-- LNQDP  ASILP VIPLD
HTLV-II     GK KLLKG  GDLIS PHP--  ----- -DQD-  ISILP LIPLR
BLV       NPQIK KLI-E  GGLSA PQTVT  PITDP LSEAE  LECLL SIPLA

HTLV-I    PARRP VIKAQ  VDTQT SHPKT  IEA-- LLDTG  ADMTV LPIAL
HTLV-II   QQQQP ILGVR  ISVMG QTPQP  TQA-- LLDTG  ADLTV IPQTL
BLV       RSR-P SVAVY  LSGPW LQPSQ  NQALM LVDTG  AENTV LPQNW

HTLV-I    FSSNT PLKNT  SVLGA GGQTQ  DHFKL TSLPV  LI--- -RLPF
HTLV-II   VPGPV KLHDT  LILGA SGQTN  TQFKL LQTPL  HI--- -FLPF
BLV       LVRDY PRIPA  AVLGA GGVSR  NRYNW LQGPL  TLALK PEGPF

HTLV-I    RTTPI VLTSC  LVDTK NNWAI  IGRDA LQQCQ  GVLYL PEAKG
HTLV-II   RRSPV ILSSC  LLDTH NKWTI  IGRDA LQQCQ  GLLYL PDDPS
BLV       ITIP- ---KI  LVDTF DKWQI  LGRDV LSRLQ  ASISI PEEVR

HTLV-I    PPVIL PIQAP  AV-LG LEHLP  RPPEI SQFPL  NQNAS RPCNT
HTLV-II   PHQLL PIATP  NT-IG LEHLP  PPPQV DQFPL  NLSAS RP
BLV       PPMVG VLDAP  PSHIG LEHLP  VPPEV PQFPL  N

HTLV-I    WSGRP WRQAI  SNPTP GQEIT  QYSQL KRPME  PGDSS TTCGP

HTLV-I    LTL
```

Fig. 3. Alignments of amino acid sequences of the proteases and putative protease of HTLV-I. Amino acid residues are expressed as one-letter codes. The HTLV-II sequence is referred from Shimotohono et al., (7), and BLV sequence from Rice et al., (8) and Yoshinaka et al., (21). Gaps (-) were inserted to increase sequence similarity. Highly conserved regions among other retroviral proteases are indicated with lines.

short sequence Leu-Val-Asp-Thr is found at four amino acids before the second conserved region. This short sequence was conserved only in HTLV-I,HTLV-II and BLV. Thus, it may be considered to be characteristic to the HTLV family proteases (23).

This protease coding frame has not been identified in the nucleotide sequence of the HTLV-I in the original and other isolates (6,18). Recently Inoue et al.(19), reported one nucleotide insertion at position 2298 after reexamining the sequence of HTLV-I ATK. In spite of the correction, this region is devided by two open reading frames with an amber stop codon at position 2440. They suggested that HTLV-I protease may be synthesized by a translational readthrough of the amber codon in the middle of the frame or the clone is biologically inactive(19). In contrast, our HTLV1C was molecularly cloned from HTLV-I producing cell line MT-2 (3,20), and the experiment of HTLV1C DNA transfection revealed the expression of the processed gag and env proteins and the syncytium formation on the transfected cells (4). We compared our nucleotide sequence data with those corresponding region of HTLV-I ATK (6,19). In the putative protease-coding frame, 11 nucleotides differ from the sequence of HTLV-I ATK, resulting in the change of six amino acids. Among the differences, the A to G transition at position 2440 is most noteworthy; the biologically active HTLV1C contains a tryptophan codon, while the HTLV-I ATK clone does an amber codon. Apparently another clone HTLV-Ib and the closely related simian retrovirus (STLV) have the same TGG codon in this region(18,19). Taking together, the results suggest that the amber codon at position 2440 of HTLV-I ATK must be occurred by a point mutation(G to A) of the biologically active prototype HTLV-I and thus exclude the possibility that HTLV-I protease may

be synthesized by suppressing the amber codon in the middle of the frame.

Within the open reading frame of the putative HTLV-I protease gene, the first methionine residue is located at the 69th position from the amino-terminus, thus the protease may be translated as a Gag-fused protein. We found 76KD precursor protein in cell extract or in *in vitro* translation system using total virion mRNA of HTLV-I (20).To express the Gag-protease fused protein, two mechanisms may be considered; one is mRNA processing, the other is a frameshift by suppressing the termination codon of the *gag* gene. The purification and primary structure analysis of proteases of murine and bovine leukemia viruses, provided an evidence that the proteases are synthesized by suppression of the *gag* termination codon or a possible ribosomal frameshift (10.21). There may be a similar frameshift mechanism in HTLV-I to produce a Gag-protease fused protein as proved for BLV (21), but we cannot rule out the possibility of involvement of mRNA splicing event for the protease formation. Further analysis of the protease gene of HTLV-I and its expression is currently in progress.

ACKNOWLEDGEMENTS

We are very grateful to Dr Takashi Miyata of Kyushu University for supplying us the data of the homology matrix comparison and sequence alignment.

This investigation was supported by grants-in-aid from the Ministry of Education, Science, and Culture of Japan.

REFERENCES

1. Kobayashi, N., and Hatanaka, M. (1986) Cancer Reviews 1, 64-95.
2. Miyoshi, I., Kubonishi, I., Sumida, I., Miyamoto, K., and Sato, J. (1980) Gann 71, 155-156.
3. Kobayashi, N., Konishi, H., Sabe, H., Shigesada, K., Noma, T., Honjo, T., and Hatanaka, M. (1984) EMBO J. 3, 1339-1343.
4. Mori, K., Sabe, H., Siomi, H., Iino, T., Tanaka, A., Takeuchi, K., Hirayoshi, K., and Hatanaka, M. J. Gen. Virol. submitted.

5. Dickson, C., Eisenman, R., Fan, H., Hunter, E., and Teich, N. (1982) in Molecular Biology of Tumour Viruses, RNA Tumour Viruses (Weiss, R., et al.eds.) pp513-648, Cold Spring Harbor Laboratory, N.Y..
6. Seiki,M., Hattori, S., Hirayama, Y., and Yoshida, M. (1983) Proc. Natl. Acad. Sci. USA 80, 3618-3622.
7. Shimotohno, K., Takahashi, Y., Shimizu, N., Gojobori, T., Golde, D.W., Chen, I.S.Y., Miwa, M., and Sugimura, T. (1985) Proc. Natl. Acad. Sci. USA 82, 3101-3105.
8. Rice, N.R., Stephens, R.M., Burny, A., and Gilden, R.V. (1985) Virol. 142, 357-377.
9. Schwartz, D.E., Tizard, R., and Gilbert, W. (1983) Cell 32, 852-869.
10. Yoshinaka, Y., Katoh, K., Copeland, T.D., and Oroszlan , S. (1985) Proc. Natl. Acad. Sci. USA 82, 1618-1622.
11. Laprevotte, I., Hampe, A., Sherr, C.J., and Galibert, F. (1984) J. Virol. 50, 884-894.
12. Takeuchi, K., Kobayashi, N., Nam, S-H., Yamamoto, N., and Hatanaka, M. (1985) J. Gen Vriol. 66, 1825-1829.
13. Sagata,N., Yasunaga, T., Tuzuku-Kawamura, J., Ohnishi, K., Okawa,Y., and Ikawa, Y. (1985) Proc. Natl. Acad. Sci. USA 82, 677-681.
14. Sanger, F., Nicklen, S., and Coulson, A.R., (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
15. Iino, T., Takeuchi, K., Nam, S-H., Siomi, H., Kobayashi, N., and Hatanaka, M. (1986) J. Gen Virol. 67, in press.
16. Toh, H., Hayashida, H., and Miyata, T. (1983) Nature 305, 827-829.
17. Toh, H., Kikuno, R., Hayashida, H., Miyata, T., Kugimiya, W., Inouye,S., Yuki, S., and Saigo, K. (1985) EMBO J. 4, 1267-1272
18. Ratner, L., Josephs, S.F., Starcich, B., Hahn, B., Shaw, G.M., Gallo, R.C., and Wong-Staal, F. (1985) J. Virol. 54, 781-790.
19. Inoue, J-I., Watanabe, T., Sato, M., Oda, A., Toyoshima, K., Yoshida, M., and Seiki, M. (1986) Virol. 150, 187-195.
20. Kobayashi, N., Yamamoto, N., Koyanagi, Y., Schneider, J., Hunsmann, G., and Hatanaka, M. (1984) EMBO J. 3, 321-325.
21. Yoshinaka, Y., Katoh, I., Copeland, T.D., Smythers, G.W., and Oroszlan, S. (1986) J. Virol. 57, 826-832.
22. Sagata, N., Yasunaga, T., and Ikawa, Y. (1984) FEBS Lett. 178, 79-82.
23. Yasunaga, T., Sagata, N., and Ikawa, Y. (1986) FEBS Lett. 199, 145-150.