Volume 223, number 1, 69–73 FEB 05199 October 1987

Molecular cloning and sequencing of cDNA for rat cathepsin L

Kazumi Ishidoh*+, Takae Towatari*, Shinobu Imajoh+, Hiroshi Kawasaki+, Eiki Kominami*, Nobuhiko Katunuma* and Koichi Suzuki+

*Department of Enzyme Chemistry, Institute of Enzyme Research, The University of Tokushima, Tokushima 770, and +Department of Molecular Biology, Tokyo Metropolitan Institute of Medical Science, 3–18 Honkomagome, Bunkyo-ku, Tokyo 113, Japan

Received 28 August 1987

A near full-length cDNA for rat cathepsin L was isolated. The deduced protein comprises 334 amino acid residues (M_r 37 685) containing a typical signal sequence (N-terminal 17 residues), pro-peptide (96 residues), and the sequence for mature cathepsin L (221 residues). Rat cathepsin L shows 94% amino acid identity with mouse cysteine proteinase. Amino acid sequence homologies of rat cathepsin L with rat cathepsins H and B are 45 and 25%, respectively. These facts indicate that mouse cysteine proteinase is probably mouse cathepsin L and that cathepsin L is more closely related to cathepsin H than cathepsin B.

Cathepsin L; Cysteine proteinase; cDNA cloning; Amino acid sequence; Proenzyme; Lysosome

1. INTRODUCTION

Cathepsin L (EC 3.4.22.15), a lysosomal cysteine proteinase [1–3], plays an important role in intracellular protein degradation [4], especially in autophagy [5,6]. Understanding the primary structure of cathepsin L provides us with an important clue as to its structure-function relationship. However, the complete amino acid sequence of cathepsin L has not yet been determined because its purification on a large scale is extremely difficult due to its low tissue content and to instability during preparation [7]. Further, the lack of specific substrates makes the identification of cathepsin L difficult [8]. The partial amino acid sequence around the active site of cathepsin L from human and chicken liver [9,10] is highly homologous to that of rat cathepsin H [11]. Thus, to isolate cDNA clones for cathepsin L, we screened a cDNA library from rat kidney using oligonucleotide probes synthesized based on the amino acid sequence of rat cathepsin H, because at the time when we started this study, none of the peptide sequences of rat cathepsin L (cf. fig.2) had been determined. The isolated clone was identified as cathepsin L by partial amino acid sequences determined with the enzyme purified from rat liver.

Correspondence address: K. Suzuki, Department of Molecular Biology, Tokyo Metropolitan Institute of Medical Science, 3-18 Honkomagome, Bunkyo-ku, Tokyo 113, Japan

Abbreviations: bp, base pairs; HPLC, high-performance liquid chromatography; SDS-PAGE, SDS-polyacrylamide gel electrophoresis; Con A, concanavalin A; MCP, mouse cysteine proteinase

The nucleotide sequence presented has been submitted to the EMBL/GenBank database under the accession number Y00697

2. MATERIALS AND METHODS

2.1. *Procedures for molecular cloning of rat cathepsin L cDNA*

Preparation of mRNA from rat kidney, con-

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/87/$3.50

struction and screening of cDNA library, DNA sequencing, and RNA blot hybridization were carried out as in [12] except that double-strand cDNA was fractionated by agarose gel electrophoresis (>2 kbp) after ligation of an *Eco*RI linker. For screening of the cDNA library, two oligonucleotides were synthesized with an Applied Biosystems 380B DNA synthesizer (see fig.1A). Probes were labelled with [α-^{32}P]dCTP by the multiprime DNA labelling system (Amersham).

2.2. *Preparation and protein sequencing analysis of cathepsin L*

Cathepsin L from rat liver was purified to homogeneity on SDS-PAGE as in [3] with additional purification steps with a Con A-Sepharose column and HPLC on TSK gel G3000SW (LKB). The heavy and light chains [7] were separated by gel filtration after *S*-carboxymethylation with acetamide [14]. The purified heavy chain was digested with lysylendopeptidase (Wako) [15], and peptides were fractionated by reverse-phase HPLC (TSK gel/ODS 120T LKB). Amino acid sequences were determined with an Applied Biosystems 470A protein sequencer/Spectra Physics SP8100 HPLC system.

3. RESULTS AND DISCUSSION

3.1. *Isolation and identification of a cDNA clone for cathepsin L*

We expected to obtain clones for both cathepsins H and L, but not for cathepsin B, by screening a cDNA library with a probe, CATH-ACT, corresponding to the region (residues 16–30) around the active site Cys-26 of cathepsin H (fig.1A). The amino acid sequence of this region is highly conserved between cathepsins H and L, even from different animals (sequence homology >85%), whereas the sequence homology between cathepsins B and H from the same species, rat, is only 53% in this region (cf. fig.4). The other probe (CATH-H) derived from residues 98–114 of cathepsin H should be specific for cathepsin H, because the amino acid sequences of various cysteine proteinases thus far sequenced are fairly diverse in this region [11]. According to the above strategy, we first screened about 7×10^4 plaques from the λgt10 cDNA library of rat kidney with CATH-ACT, and positive clones were further classified according to whether they hybridized with CATH-H. Finally, seven clones positive to CATH-ACT, but not to CATH-H, were selected. These clones had inserts derived from the same mRNA as judged by the *Eco*RI digestion profiles (fig.1B). Therefore, we selected λN10, which had the longest insert, for sequence analysis. The nucleotide sequence of λN10 is shown in fig.2. A consensus polyadenylation signal was seen at 1280–1285 (double underlined) [16] together with a poly(A) sequence at the 3'-end. An open reading frame encoding 334 amino acid residues (M_r 37 685) is indicated under the nucleotide sequence. An in-phase stop codon, TGA, was found upstream of the initiation ATG. All the amino acid sequences that were determined for purified cathepsin L (underlined residues) were found in the deduced protein sequence. Therefore, we concluded that λN10 is a cDNA clone for rat cathepsin L.

The mRNA for cathepsin L is about 1.7 kb in length (fig.3) as judged by RNA blot hybridization analysis, indicating that λN10 (1.4 kbp) covers most of the mRNA.

3.2. *Post-translational modification of cathepsin L*

By N-terminal sequence analysis of the isolated heavy and light chains, their N-termini were assigned as Ile-1 and Asn-178, respectively, which demonstrates that the heavy and light chains of cathepsin L are derived from a single polypeptide. The peptide from Met-(−113) to Gln-(−1) is cleaved off during translocation to the lysosome. The deduced M_r values of the heavy (177 residues) and light (44 residues) chains are 19 710 and 5056, respectively, provided that proteolytic processing does not occur near the C-terminus of either chain. The C-termini of the light and heavy chains of cathepsin L have not yet been analyzed, since we could not purify cathepsin L enough for C-termini analyses. However, Asp-176 and Ser-177 were not detected by amino acid sequence analyses of the corresponding lysyl peptides. Therefore, we presume that the C-terminal residue of the heavy chain is Thr-175.

The mature cathepsin L shows 25 and 5 kDa bands on SDS-PAGE. The discrepancy in molecular mass of the heavy chain is mainly due to the presence of sugar chains. Two potential *N*-

```
                16                                             30
                ProValLysAsnGlnGlyAlaCysGlySerCysTrpThrPheSer
CATH-ACT    5'  CCTGTGAAGAACCAGGGCGCCTGTGGCTCCTGCTGGACCTTCTCC  3'
                98                                                 114
                GlnCysLysPheAsnProGluLysAlaValAlaPheValLysAsnValVal
CATH-H      5'  CAGTGCAAGTTCAACCCTGAGAAGGCTGTGGCCTTTGTGAAGAATGTGGTG  3'
```

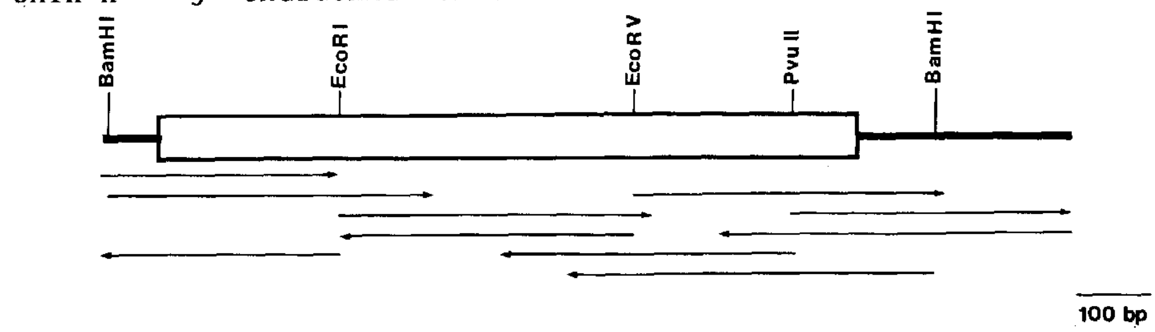

Fig.1. (A) Nucleotide sequences of probes used. Preferable nucleotide sequences were designated as in [13] from the amino acid sequence of rat cathepsin H [11]. CATH-ACT (45-mer) corresponds to the sequence around the active site Cys-26 (residues 16–30), and CATH-H (51-mer) encodes residues 98–114 of rat cathepsin H. (B) Restriction map and sequencing strategy of λN10. DNA sequencing was carried out by the dideoxy method in both directions (arrowed). White open box indicates the amino acid coding region. Arrows show the directions and lengths of sequencing.

```
-70                                                                                 -1
CGGCGACCTCCGGGGATCCGAGTTTGCAGACTACGTGTGTGCGCAGCTAGCCACCTCAGGTGTTTGAACC
                                                                              ***
1                                                                        100
ATGACCCCTTTACTCCTCCTGGCTGTCCTCTGCTTGGGAACAGCCTTAGCCACTCCAAAATTTGATCAAACGTTTAATGCACAGTGGCACCAGTGGAAGTCCACACACAGAAGACTGTAT
MetThrProLeuLeuLeuLeuAlaValLeuCysLeuGlyThrAlaLeuAlaThrProLysPheAspGlnThrPheAsnAlaGlnTrpHisGlnTrpLysSerThrHisArgArgLeuTyr
-113                     -100
                                                                                     200
GGCACGAATGAGGAAGAGTGGAGGAGAGCAGTGTGGGAGAAGAACATGAGGATGATCCAGCTACACAATGGGGAGTACAGCAACGGGAAGCACGGCTTTACCATGGAGATGAACGCCTTC
GlyThrAsnGluGluGluTrpArgArgAlaValTrpGluLysAsnMetArgMetIleGlnLeuHisAsnGlyGluTyrSerAsnGlyLysHisGlyPheThrMetGluMetAsnAlaPhe
                                                                      -50
                                                        300
GGTGACATGACCAATGAGGAATTCAGGCAGATAGTGAATGGCTATCGCCACCAGAAGCACAAGAAGGGAAGGTTATTTCAGGAACCTCTGATGCTGCAGATCCCCAAGACTGTGGACTGG
GlyAspMetThrAsnGluGluPheArgGlnIleValAsnGlyTyrArgHisGlnLysHisLysLysGlyArgLeuPheGlnGluProLeuMetLeuGlnIleProLysThrValAspTrp
                                                                                                          -1  1
                        400
AGAGAAAAGGGTTGTGTGACTCCTGTGAAGAATCAGGGCCAGTGTGGTTCTTGCTGGGCTTTTAGCGCATCGGGTTGCCTAGAAGGACAGATGTTCCTTAAGACTGGCAAACTGATCTCA
ArgGluLysGlyCysValThrProValLysAsnGlnGlyGlnCysGlySerCysTrpAlaPheSerAlaSerGlyCysLeuGluGlyGlnMetPheLeuLysThrGlyLysLeuIleSer
            500                                                                                                     600
CTGAGTGAACAGAACCTTGTGGACTGTTCTCACGATCAAGGCAATCAGGCCTGTAATGGAGGCCTGATGGATTTTGCTTTCCAGTACATTAAGGAAAATGGAGGTCTGGACTCAGAGGAG
LeuSerGluGlnAsnLeuValAspCysSerHisAspGlnGlyAsnGlnGlyCysAsnGlyGlyLeuMetAspPheAlaPheGlnTyrIleLysGluAsnGlyGlyLeuAspSerGluGlu
     50
                                                                                                   700
TCTTATCCCTATGAAGCAAAGGATGGATCTTGTAAATACAGAGCTGAGTATGCTGTGGCTAACGACACAGGGTTTGTGGATATCCCTCAGCAAGAGAAAGCCCTCATGAAGCCTGTAGCG
SerTyrProTyrGluAlaLysAspGlySerCysLysTyrArgAlaGluTyrAlaValAlaAsnAspThrGlyPheValAspIleProGlnGlnGluLysAlaLeuMetLysProValAla
                                    100                   *
                                                                                    800
ACGGTGGGGCCTATTTCTGTTGCCATGGATGCAAGCCATCCGTCTCTCCAGTTCTATAGTTCAGGTATCTACTATGAACCCAACTGTAGCAGCAAGGACCTCGACCATGGGGTTCTGGTG
ThrValGlyProIleSerValAlaMetAspAlaSerHisProSerLeuGlnPheTyrSerSerGlyIleTyrTyrGluProAsnCysSerSerLysAspLeuAspHisGlyValLeuVal
                                                                      150              *
                                                             900
GTTGGCTATGGTTATGAAGGAACAGATTCAAATAAGGATAAATACTGGCTTGTCAAAAACAGCTGGGGTAAAGAATGGGGTATGGATGGCTACATCAAAATAGCCAAAGACCGGAACAAC
ValGlyTyrGlyTyrGluGlyThrAspSerAsnLysAspLysTyrTrpLeuValLysAsnSerTrpGlyLysGluTrpGlyMetAspGlyTyrIleLysIleAlaLysAspArgAsnAsn
                                                                                                    200
                                    1000
CACTGCGGACTTGCCACCGCAGCCAGCTATCCTATCGTGAATTGATGGACAGCGATAATAAGGACTGACGGACACTACATCCGAAGGAGTTCATCTTAAAACTGACCAAACCCGTCTCTG
HisCysGlyLeuAlaThrAlaAlaSerTyrProIleValAsn***
                                        221
         1100                                                                                                        1200
AGTGAGACCATGGTACTTGAATCCTTCAGGATCCAAGTCACGATTTAAATTCTGTTGACATTTTTACATGGGTTAAATGTTACCACTACTTAAAACTCCTGTTATAAACAGCTTTATAAT
                                                                                                  1300
ATTGGACACTTAATGCTTAATTCTGATTCTGGAATATTTGTTTTATAAAAGTTGTATAAAACTTTCTTTACCTTTTAAAAATAAATTTTAGCTCAGTGCAAAAAAAAAA
```

Fig.2. Nucleotide sequence of rat cathepsin L (λN10) and the deduced amino acid sequence. Nucleotide sequence is numbered starting at the initiation codon ATG. Negative numbers show the upstream region. The deduced amino acid sequence for the precursor of cathepsin L, shown under the nucleotide sequence, is numbered beginning at the N-terminal residue of mature cathepsin L. Negative numbers indicate pre- and pro-sequences. Amino acid sequences determined using the purified enzyme are underlined. Broken lines indicate undetectable or ambiguous amino acids. Asterisks denote potential glycosylation sites. Arrows indicate cleavage sites in post-translational processing. The polyadenylation signal is double underlined. Triple stars indicate the termination codon.

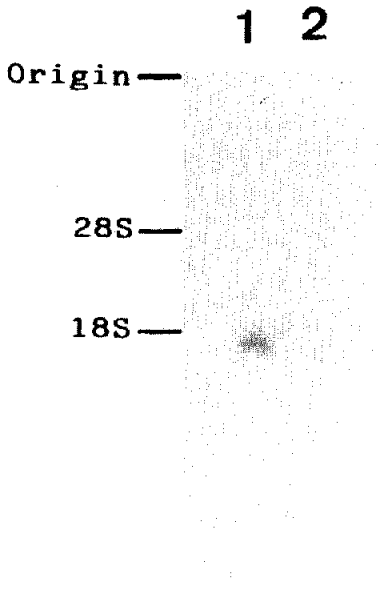

Fig.3. RNA blot hybridization analysis. Poly(A)$^+$ RNA from rat kidney (lane 1, 0.5 μg; lane 2, 0.1 μg) was analyzed. An *Eco*RI fragment (−70 to 263) of λN10 was used as a probe. Positions of rat ribosomal RNAs (28 S and 18 S) are arrowed.

glycosylation sites (Asn-108 and Asn-155) are found in the heavy chain. At least, Asn-108 is modified as judged from the sequencing profile of a peptide from the heavy chain (not shown). No potential glycosylation site exists in the pro-peptide region, whereas cathepsin B possesses Asn-linked sugar chain in its pro-peptide region [17].

3.3. *Amino acid sequence homology with other cysteine proteinases*

This is the first determination of the structure of complete cathepsin L. The deduced amino acid sequence of rat cathepsin L is, however, highly homologous (94%) to that of mouse cysteine proteinase (MCP), whose sequence was determined recently by cDNA cloning [18]. cDNA clones for MCP were isolated from the cDNA library of mouse macrophage cell line J774 using subtractive probes expressed in T-cell line S49.1 but lacking in L5187Y. The authors demonstrated that MCP is localized in lysosomes by immunofluorescence microscopy using an antibody against a recombinant fusion protein expressed in *E. coli*, and also that MCP presumably undergoes post-translational processing similar to that of cathep-

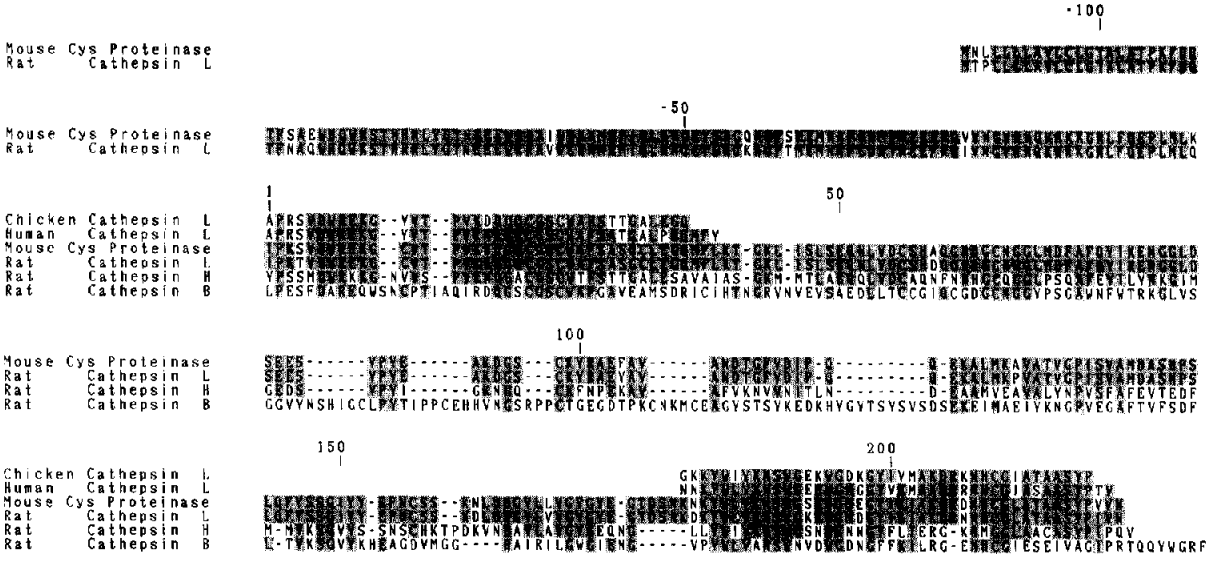

Fig.4. Comparison of amino acid sequence of rat cathepsin L with chicken cathepsin L [10], human cathepsin L [9], MCP [18], rat cathepsins H [11] and B [17]. Amino acid residues identical between rat cathepsin L and the other proteins are shaded. Residue numbers for rat cathepsin L are the same as in fig.2. Gaps, shown by dashes, were introduced for optimal alignment.

sins as judged from the results of Western blot analysis. They also stated that MCP is probably mouse cathepsin L on the basis of partial amino acid sequence homology with human cathepsin L (82.5% identical in the first 40 amino acids), although precise results at the protein level to support the identification have not been reported. The amino acid sequences of rat cathepsin L and MCP can be aligned without any gaps and most of the differences are substitutions between conserved amino acids (fig.4). Further, in the coding region, the nucleotide sequence of rat cathepsin L is 93% identical to that of MCP [18]. Thus, our results strongly support the hypothesis that MCP is probably mouse cathepsin L [9,18].

Partial amino acid sequences of chicken and human cathepsin L have been reported [9,10]. These show only 70–80% sequence homology with rat cathepsin L (fig.4). Significant sequence differences among cathepsin L from different species will be evaluated when the complete sequences of human and chicken are determined.

The overall sequence homologies of rat cathepsin L with other cysteine proteinases are rat cathepsins H and B, 45 and 25%, respectively (fig.4), when gaps introduced to maximize sequence homology were considered as mismatches. Thus, cathepsin L is more closely related to cathepsin H than cathepsin B.

The amino acid sequence of the precursor of cathepsin L will provide useful information to establish the regulation and activation mechanisms of cysteine proteinases.

ACKNOWLEDGEMENTS

We thank Drs S. Ohno and Y. Emori for helpful discussions on the techniques of cDNA cloning. This work was supported in part by research grants from the Ministry of Education, Science and Culture of Japan.

REFERENCES

[1] Towatari, T., Tanaka, K., Yoshikawa, D. and Katunuma, N. (1976) FEBS Lett. 67, 284–288.
[2] Kirschke, H., Langner, J., Wiederanders, B., Ansorge, S. and Bohley, P. (1977) Eur. J. Biochem. 74, 293–301.
[3] Towatari, T., Tanaka, K., Yoshikawa, D. and Katunuma, N. (1978) J. Biochem. 84, 659–671.
[4] Katunuma, N., Towatari, T., Kominami, E. and Hashida, S. (1981) in: Proteinases and their Inhibitors (Turk, V. and Vitale, L. eds) pp. 83–92, Pergamon, Oxford.
[5] Ii, K., Hizawa, K., Kominami, E., Bando, Y. and Katunuma, N. (1985) H. Histochem. Cytochem. 33, 1173–1175.
[6] Ohshita, T., Kominami, E., Ii, K. and Katunuma, N. (1986) J. Biochem. 100, 623–632.
[7] Bando, Y., Kominami, E. and Katunuma, N. (1986) J. Biochem. 100, 35–42.
[8] Barrett, A.J. and Kirshke, H. (1981) Methods Enzymol. 80, 535–561.
[9] Mason, R.W., Walker, J.E. and Northrop, F.D. (1986) Biochem. J. 240, 373–377.
[10] Wada, K. and Tanabe, T. (1986) FEBS Lett. 209, 330–334.
[11] Takio, K., Towatari, T., Katunuma, N., Teller, D.C. and Titani, K. (1983) Proc. Natl. Acad. Sci. USA 80, 3666–3670.
[12] Emori, Y., Kawasaki, H., Imajoh, S., Imahori, K. and Suzuki, K. (1987) Proc. Natl. Acad. Sci. USA 84, 3590–3594.
[13] Lathe, R. (1985) J. Mol. Biol. 183, 1–12.
[14] Takio, K., Towatari, T., Katunuma, N. and Titani, K. (1980) Biochem. Biophys. Res. Commun. 97, 340–346.
[15] Tsunasawa, S., Sugihara, A., Masaki, T., Sakiyama, F., Takeda, Y., Mawatani, T. and Narita, K. (1987) J. Biochem. 101, 111–121.
[16] Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211–214.
[17] Chan, S.J., Segundo, B.S., McCormick, M.B. and Steiner, D.F. (1986) Proc. Natl. Acad. Sci. USA 83, 7721–7725.
[18] Portnoy, D.A., Erickson, A.H., Kochan, J., Ravetch, J.V. and Unkeless, J.C. (1986) J. Biol. Chem. 261, 14697–14703.