Volume 259, number 1, 71-74 FEB 07924 December 1989

# Gene structure and 5'-upstream sequence of rat cathepsin L

Kazumi Ishidoh*°+, Eiki Kominami°+, Koichi Suzuki* and Nobuhiko Katunuma°

*Department of Molecular Biology, Tokyo Metropolitan Institute of Medical Science, Honkomagome 3-18-22, Bunkyo-ku, Tokyo 113, °Division of Enzyme Chemistry, Institute for Enzyme Research, The University of Tokushima, Tokushima 770 and +Department of Biochemistry, Juntendo University School of Medicine, Hongo 2-1-1, Bunkyo-ku, Tokyo 113, Japan

Received 10 October 1989

The structure of rat cathepsin L gene has been determined. The gene spans 8.5 kilobase pairs comprising 8 exons, and has an intron located near the active site cysteine residue. The gene structure does not correspond well to the functional units of the proteinase. These characteristics are found to be in common with the cysteine proteinase gene family. In the 5'-upstream region, one CAAT-box and four SP-1 binding sites, together with two AP-2 binding sites and CRE, but no typical TATA-box are found. Further, SP-1 and AP-2 binding sites and an octamer motif are also found in the 1st intron, suggesting a complex regulatory mechanism for the expression of the cathepsin L gene.

Cathepsin L; Gene structure; Lysosomal proteinase; Major excreted protein; Promoter-enhancer

## 1. INTRODUCTION

Cathepsin L is a typical cysteine proteinase which, together with cathepsins B and H, belongs to the papain superfamily [1]. These proteinases play a major role in intracellular proteolysis [2]. In addition, extracellular functions for cathepsins B and L have been reported; for example, cathepsins B and L have been implicated in tumor metastasis [3,4], and cathepsin L in emphysema [5]. Recently, a major protein excreted from ras-transformed NIH3T3 cells (MEP) was identified as mouse cathepsin L [6]. Secretion of cathepsin L from NIH3T3 cells is also observed upon treatment of cells with a tumor promoter, TPA and PDGF [7,8]. These factors affect the transcription level of the cathepsin L gene but not of the cathepsin H gene even though cathepsins L and H show very high amino acid sequence homology [1] and are presumably derived from a common ancestral gene. To elucidate the functional diversity and regulatory mechanism for gene expression, we analyzed the structure of the cathepsin L gene.

In this paper, we reported the structure of the rat cathepsin L gene and the nucleotide sequence of the 5'-upstream region and the 1st intron which probably regulates cathepsin L gene transcription.

*Correspondence address:* K. Ishidoh, Department of Biochemistry, Juntendo University School of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo 113, Japan

*Abbreviations:* MEP, major excreted protein; TPA, 4-*O*-tetradecanoyl phorbol-13-acetate; PDGF, platelet-derived growth factor; kbp, kilobase pairs; bp, base pairs; b, bases; AP-2, activator protein 2; CRE, cAMP regulatory element

## 2. MATERIALS AND METHODS

### 2.1. *Materials*

Materials used in this work were obtained from the following sources: restriction enzymes from Takara Shuzo, Toyobo and New England Biolab.; [$\alpha$-$^{32}$P]dCTP and [$\gamma$-$^{32}$P]ATP from Amersham and ICN; multi-prime labeling kit and nylon membrane Hybond N from Amersham; T7 DNA polymerase sequencing kit from Toyobo; other enzymes from Takara Shuzo.

### 2.2. *Methods*

*Bam*HI fragment of cDNA for rat cathepsin L [1] (from −57 to 1088) was labeled with [$\alpha$-$^{32}$P]dCTP by a multi-prime labeling kit and used as a probe. All procedures from screening to sequencing were carried out as previously described [9]. $S_1$ mapping analysis was carried out as described [10]; i.e. hybridization at 65°C overnight and digestion with 1000 U/ml of $S_1$ nuclease at 30°C for 30 min. Nucleotide sequence homology was sought in the Genebank.

## 3. RESULTS AND DISCUSSION

### 3.1. *Isolation and characterization of rat cathepsin L gene*

From the rat genomic library ($3.2 \times 10^5$ independent clones), 4 positive clones were isolated. Judging from restriction mapping, these inserts were derived from a single gene and overlapped one another. Genomic Southern hybridization analyses show that the cathepsin L gene is a single copy (data not shown), and that λGL1 and λGL6, spanning in total 18.5 kbp, encompass the rat cathepsin L gene (fig.1).

Nucleotide sequence analyses of these clones and nucleotide sequence comparison between the gene and cDNA show that the rat cathepsin L gene spans 8.5 kbp and comprises 8 exons. The pre- and pro-peptide regions are encoded by exon 2 and exons 2–4, respectively. Exons 4–8 encode the mature enzyme region and

*Published by Elsevier Science Publishers B.V. (Biomedical Division)*
00145793/89/$3.50 

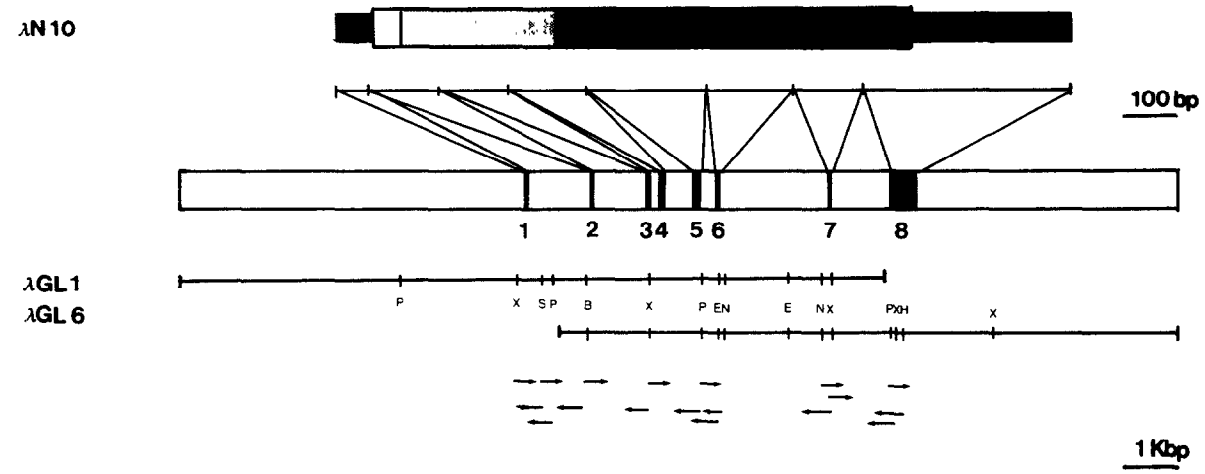



Fig.1. Restriction maps and schematic structure of the rat cathepsin L gene. λN10 is a full length cDNA clone for rat cathepsin L. Its schematic structure is shown at the top. Solid bars indicate 5′- and 3′-noncoding regions. The coding regions for the pro-peptide region and the mature enzyme region are stippled and horizontally shaded, respectively. The open area indicates the pre-peptide region. λGL1 and λGL6 span 18.5 kbp. Exons are shown by numbered filled areas. Arrows indicate the directions and lengths of sequencing. Restriction enzymes: *Pvu*II (P), *Xmn*I (X), *Sma*I (S), *Bam*HI (B), *Eco*RV (E), *Nco*I (N), *Hin*cII (H).

exon 1 codes only the 5′-upstream region (table 1 and fig.1). The rat cathepsin L gene structure does not correspond to its functional units. An intron insertion is found near the active site Cys residue (fig.2). These characteristics are found in common with gene structures of cysteine proteinases [9]. Each exon spans 101–225 bp in length, except for exon 8 (397 bp), and contains 43.2–51.2% GC, except for exon 1 (61.0%) and exon 8 (38.3%). All boundary sequences are consistent with the GT-AG rule [11] (data not shown).

The nucleotide sequence of the rat cathepsin L gene differs from that of the cDNA spanning ca. 1.4 kbp in 5 positions (table 1). Two alternations are located in the 3′-noncoding region. The adenine-12 in the cDNA sequence which is replaced by guanine in the gene sequence is a silent change. The cytosine in the gene sequence corresponding to guanine-93 in the cDNA sequence causes an amino acid change from Gln-(−16) to His. The guanine in the gene sequence corresponding to cytosine-712 in the cDNA sequence causes an amino acid change from Pro-125 to Ala. The amino acid residue for cytosine-712 determined at the protein level [12] is Ala rather than Pro, indicating that cytosine-712 has arisen from a mismatch during cDNA synthesis. The other cases are probably due to sequence polymorphism.

The gene structures of rat cathepsins H and L are not so similar in spite of their common ancestry [1]. Numbers and positions of introns are distinct (fig.2). Only two introns, including one immediately before the active site Cys residue, are located in the same positions. In the case of the serine proteinase superfamily, introns conserved at the amino acid sequence level are located immediately after the active site His residue and before the active site Ser residue [13]. Other introns are probably located in the surface loop regions of the three-dimensional structure [14] and are not conserved. In the cysteine proteinase superfamily, a similar situation is not observed. No intron position was found near the cleavage site where processing from a single-chain form to a two-chain form occurs, a region that is probably a loop out in the three-dimensional structure. Most

Table 1

Summary of the rat cathepsin L gene structure and differences in nucleotide sequences between the gene and its cDNA

| Exon | Nucleotide no. of cDNA | (bp) | GC-content (%) | Difference | |
|---|---|---|---|---|---|
| | | | | Gene | cDNA |
| 1 | −112–−12 | (101) | 61.0 | | |
| 2 | −11–126 | (137) | 49.6 | TTG (Leu) | TTA (Leu) [12] |
| | | | | CAC (His) | CAG (Gln) [93] |
| 3 | 127–249 | (123) | 51.2 | | |
| 4 | 250–396 | (147) | 49.0 | | |
| 5 | 397–621 | (225) | 48.0 | | |
| 6 | 622–784 | (163) | 48.5 | GCT (Ala) | CCT (Pro) [712] |
| 7 | 785–902 | (118) | 43.2 | | |
| 8 | 903–1299 | (397) | 38.3 | TTA | TGA [1027] (3′-noncoding) |
| | | | | TGT | TTT [1268] (3′-noncoding) |

Exons are numbered from 5′ to 3′ in the direction of transcription. Nucleotide numbering of the cDNA starts at the initiation codon ATG. Negative numbers indicate 5′-noncoding regions. the 5′-end of exon 1 represents a major protection site in $S_1$ mapping analysis. The 3′-end of exon 8 represents the poly(A)$^+$ addition site. Nucleotide sequences that differ between the gene and its cDNA are underlined. Numbers in brackets show the nucleotide number in the cDNA sequence. Amino acids in parentheses below the nucleotide sequences are deduced from the nucleotide sequence

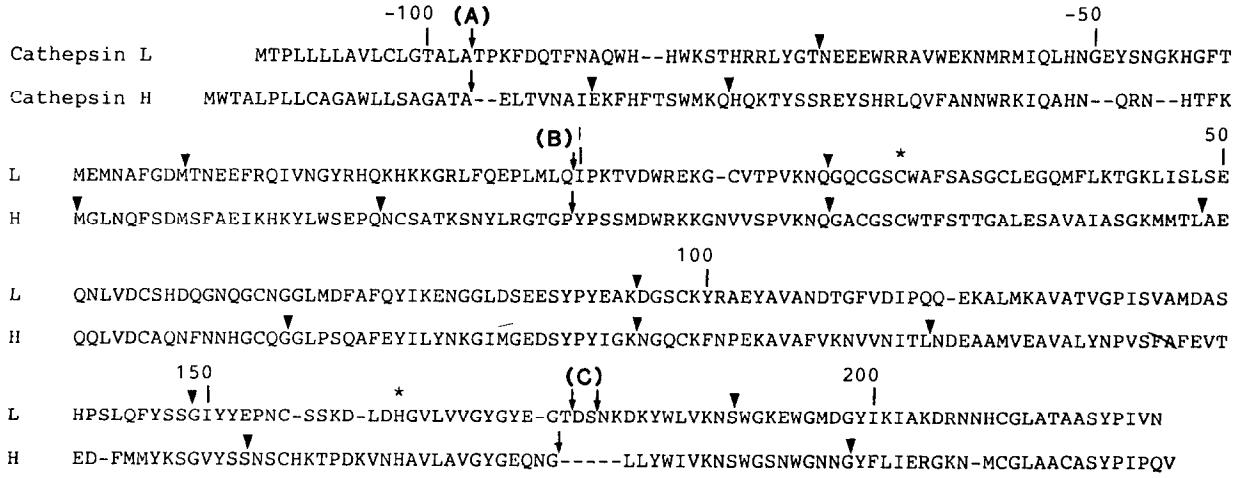
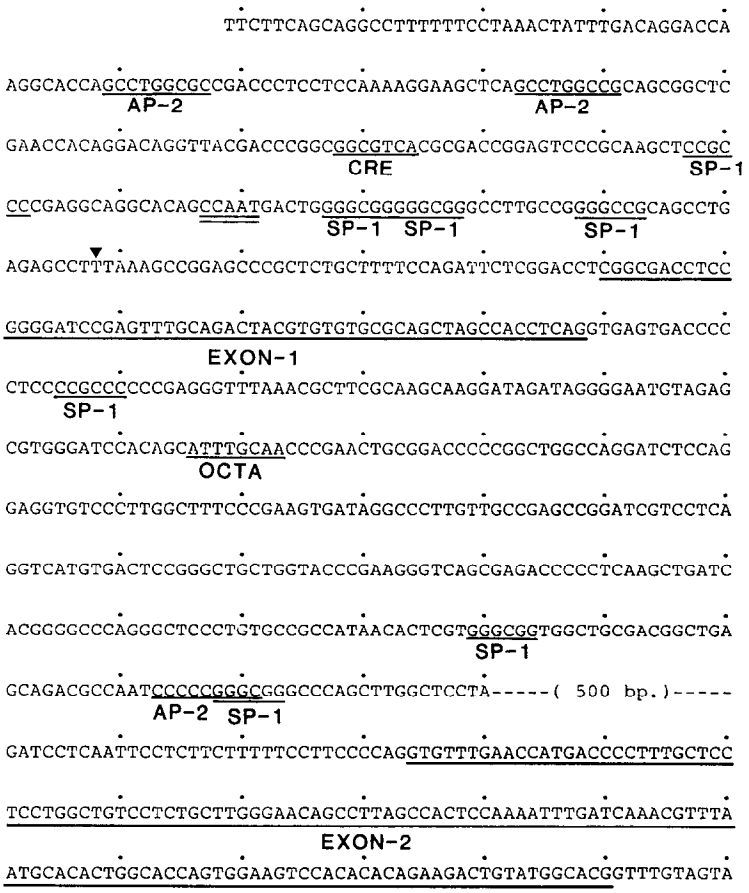



Fig.2. Comparison of intron positions between cathepsins L and H. Amino acids are numbered starting with the N-terminus of the mature enzyme region of cathepsin L. Negative numbers indicate the pre- and pro-peptide regions. Arrowheads indicate intron insertion positions. Arrows indicate the cleavage sites between the pre- and pro-peptide (A), pro-peptide and the mature enzyme (B) and heavy and light chains (C). Asterisks show the active site Cys- and His-residues.

A




B

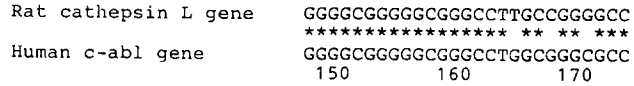



Fig.3. (A) Nucleotide sequence around the 1st exon. Thick lines below the nucleotide sequence indicate the cDNA sequences, namely, exon 1 and exon 2. Double thin underlining shows the CAAT-box. Single thin underlining indicates promoter-enhancer elements; SP-1, activator protein 2 binding sequence (AP-2), cAMP regulatory element (CRE) and octamer motif (OCTA). The arrowhead indicates a major protection position in $S_1$ mapping analysis. (B) Comparison of nucleotide sequences between an SP-1 binding sites cluster region (59–85 bp upstream from 5′-end of cDNA) in the rat cathepsin L gene and the c-abl gene [21]. Asterisks indicate identical nucleotides.

intron positions are located near Cys residues which are important in supporting three-dimensional structure (fig.2).

### 3.2. *Nucleotide sequence of 5'-upstream and 1st intron*

The nucleotide sequence of the rat cathepsin L gene around exon 1 is shown in fig.3A. $S_1$ mapping analysis (data not shown) revealed that the major transcriptional initiation site was at T, 42 b upstream of the end of the cDNA. In the 5'-upstream region, one CAAT-box and four SP-1 binding sites were found, but no typical TATA-box existed [15]. A typical TATA-box is also lacking in the 5'-upstream region of genes for lysosomal enzymes such as lysosomal human $\alpha$-galactosidase and human $\beta$-hexosaminidase $\alpha$- and $\beta$-chains [16–19], although the CAAT-box and SP-1 binding sites exist. These characteristics may be common among the 5'-upstream regions of lysosomal enzyme genes. In addition three SP-1 binding sites were found in the 1st intron. Since introns affect the transcriptional level in some cases [20], these SP-1 binding sites may also affect the transcriptional level of rat cathepsin L.

A search for nucleotide sequence homology in the Genebank shows that a region 59–85 bp upstream from the 5'-end of cDNA, where three SP-1 binding sites are clustered, is homologous to the promoter region of the ubiquitously expressed c-abl gene (fig.3B) [21]. This fact suggests that this region may be the basic promoter region of cathepsin L.

Various enhancer elements, i.e. AP-2, SP-1, CRE and an octamer motif [22,23], found in the 5'-upstream region and in the 1st intron may also regulate transcription of the cathepsin L gene. Ras-transformed NIH3T3 and normal NIH3T3 cells secrete cathepsin L as a major excreted protein (MEP) upon treatment with TPA and PDGF [6–8]. Troen et al. isolated the MEP gene and demonstrated that this gene spans more than 8 kbp and comprises more than 6 exons, judging from genomic Southern hybridization analysis [24]. These results are preliminary, but consistent with our present results. Nevertheless, the transcription level of cathepsin L in African green monkey kidney CV-1 cells and human epidermoid carcinoma A431 cells transfected with cathepsin L gene is not affected by treatment with TPA. The results suggest that the effect of TPA is cell-type-specific. AP-2, a tissue specific *trans*-acting factor may be responsible for tissue-specific induction of cathepsin L by TPA. In addition, the presence of CRE, a ubiquitous enhancer, suggests that the gene expression of cathepsin L may be affected by intracellular cAMP level. Transcriptional regulation of the cathepsin L gene is quite complex and apparently requires further precise analysis.

Knowledge of the cathepsin L gene will provide a clue to understanding the molecular evolution and functional diversity among cysteine proteinases and the mechanism for transcriptional regulation.

*Acknowledgements:* We thank Dr Yasufumi Emori for helpful discussion on gene cloning techniques and Drs Yoko Nadaoka and Hiroshi Kawasaki for the nucleotide sequence homology search in the Genebank. This work was supported in part by research grants from the Ministry of Education, Science and Culture of Japan, Yamanouchi Foundation for Research on Metabolic Disorders, The Tokyo Biochemical Research Foundation and Takeda Science Foundation.

## REFERENCES

[1] Ishidoh, K., Towatari, T., Imajoh, S., Kawasaki, H., Kominami, E., Katunuma, N. and Suzuki, K. (1987) FEBS Lett. 223, 69–73.
[2] Katunuma, N. (1989) in: Intracellular Proteolysis – Mechanisms and Regulation (Katunuma, N. and Kominami, E. eds) pp. 3–23.
[3] Sloane, B.F., Rozhin, J., Johnson, K., Taylor, H., Crissman, J.D. and Honn, K.V. (1986) Proc. Natl. Acad. Sci. USA 83, 2483–2487.
[4] Yagel, S., Warner, A.H., Nellans, H.N., Lala, P.K., Waghorne, C. and Denhardt, D.T. (1989) Cancer Res. 49, 3553–3557.
[5] Mason, R.W., Johnson, D.A., Barrett, A.J. and Chapman, H.A. (1986) Biochem. J. 233, 925–927.
[6] Troen, B.R., Gal, S. and Gottesman, M.M. (1987) Biochem. J. 246, 731–735.
[7] Gottesman, M.M. and Sobel, M.E. (1980) Cell 19, 449–455.
[8] Frick, K.K., Doherty, P.J., Gottesman, M.M. and Scher, C.D. (1985) Mol. Cell. Biol. 5, 2582–2589.
[9] Ishidoh, K., Kominami, E., Katunuma, N. and Suzuki, K. (1989) FEBS Lett. 253, 103–107.
[10] Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
[11] Mount, S.M. (1982) Nucleic Acids Res. 10, 459–472.
[12] Towatari, T. and Katunuma, N. (1988) FEBS Lett. 236, 57–61.
[13] Rogers, J. (1985) Nature 315, 458–459.
[14] Craik, C.S., Rutter, W.J. and Fletterick, R. (1983) Science 220, 1125–1129.
[15] Maniatis, T., Goodbourn, S. and Fischer, J.A. (1987) Science 236, 1237–1245.
[16] Quinn, M., Hantzopoulos, P., Fidanza, V. and Calhoun, D.H. (1987) Gene 58, 177–188.
[17] Bishop, D.F., Kornreich, R. and Desnick, R.J. (1988) Proc. Natl. Acad. Sci. USA 85, 3903–3907.
[18] Proia, R.L. and Soravia, E. (1987) J. Biol. Chem. 262, 5677–5681.
[19] Neote, K., Bapat, B., Dumbrille-Ross, A., Troxel, C., Schuster, S.M., Mahuran, D.J. and Gravel, R.A. (1988) Genomics 3, 279–286.
[20] Cohen, J.B. and Levinson, A.D. (1988) Nature 334, 119–124.
[21] Shtivelman, E., Lifshitz, B., Gale, R.P., Roe, B.A. and Canaani, E. (1986) Cell 47, 277–284.
[22] Jones, N.C., Rigby, P.W. and Ziff, E.B. (1988) Genes and Development 2, 267–281.
[23] Roesler, W.J., Vandenbark, G.R. and Hanson, R.W. (1988) J. Biol. Chem. 263, 9063–9066.
[24] Troen, B.R., Ascherman, D., Atlas, D. and Gottesman, M.M. (1988) J. Biol. Chem. 263, 254–261.