Volume 253, number 1,2, 103–107 FEB 07458 August 1989

Gene structure of rat cathepsin H

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

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

Received 7 June 1989

The gene structure of rat cathepsin H was determined. It comprises at least 12 exons of various lengths (32-433 bp) spanning in total more than 17.5 kbp. The gene structure does not correspond well to the functional unit of the proteinase. The region around the active site Cys residue, the most conserved region among cysteine proteinases, is split by an intron. This is a common characteristic among the gene structures of cysteine proteinases.

Cathepsin H; Cysteine proteinase; Gene structure

1. INTRODUCTION

Cathepsin H is a typical lysosomal cysteine proteinase, which, together with cathepsins B and L [1], plays a major role in lysosomal proteolysis. These proteins are translated as prepro-enzymes and exist in lysosomes as mature enzymes [2–4]. The amino acid sequences of various cysteine proteinases, including those of plant origin are highly homologous to each other [5]. Especially, the amino acid sequence of rat cathepsin H is 61.8% identical in the mature enzyme to that of aleurain in the barley aleurone cell. In the case of aleurain, the GC content of exons coding for the pre- and pro-peptide regions (exon 1–3) and the mature enzyme region (exon 4–8) is quite different. Whittier et al. [6], therefore, suggested that the gene for aleurain was constructed through some sort of recombination event. To investigate whether this recombination theory is also applicable to mammalian cysteine proteinases and to identify the characteristics of the gene structures common among cysteine proteinases, we isolated genomic clones for rat cathepsin H and determined the gene structure.

In this paper, we report the gene structure of cathepsin H and compare it with those of other cysteine proteinases.

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

* *Present address:* Department of Biochemistry, Juntendo University School of Medicine, 2-1 Hongo, Bunkyo-ku 113, Japan

Abbreviations: kbp, kilo base pairs; bp, base pairs; CANP, calcium activated neutral proteinase, the same as calpain; CP1, cysteine proteinase 1; CP2, cysteine proteinase 2

2. MATERIALS AND METHODS

2.1. *Materials*

The sources of materials used in this work are as follows: restriction enzymes from Takara Shuzo, New England Biolab, and Toyobo; [α-^{32}P]dCTP and [γ-^{32}P]ATP from Amersham or ICN; multiprime DNA labeling kit from Amersham; cloning vector λEMBL3 and in vitro packaging system from Stratagene Cloning Systems; other enzymes from Toyobo.

2.2. *Methods*

A rat genomic library was constructed by the method described previously [7] except that rat liver was used as the starting material. cDNA for rat cathepsin H labeled with [α-^{32}P]dCTP by the multiprime DNA labeling system, was used as a probe. The rat genomic library was screened by the probe under low stringent conditions [8]. Positive plaques were subjected to second and third screenings to isolate single plaques. DNA inserts of positive clones were digested with *Sal*I and

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

subcloned into pUC vectors. Restriction maps generated by digestion with appropriate restriction enzymes and Southern hybridization analyses revealed fragments containing exons. These fragments were further subcloned into pUC vectors. Nucleotide sequences were determined by the bi-directional dideoxy-sequencing method [9].

3. RESULTS AND DISCUSSION

3.1. *Isolation and characterization of the rat cathepsin H gene*

From the rat genomic library (6.4×10^5 independent clones), 13 clones were isolated. Judging from restriction maps and Southern hybridization analyses, two clones, λGH1002 and λGH109, spanning 21.7 kbp were identified as clones encompassing the rat cathepsin H gene (fig.1). Sequence analysis of these clones revealed the structure of the rat cathepsin H gene, consisting of 17.5 kbp, comprising at least 12 exons. Upstream of exon 1, only one GC box was found, but neither TATA nor CAAT boxes existed [10], suggesting the presence of one more exon (data not shown). The length of each exon varied from 32 bp (exon 2) to 433 bp (exon 12). The boundary sequences of the exon-intron junctions (table 1) determined by comparing the nucleotide sequences of the cDNA and genomic DNA, are consistent with the GT-AG rule [11]. The GC contents of exons also vary significantly from 34.4% (exon 2) to 68.8% (exon 1). Exon 1 codes mainly for the signal peptide, exons 2–4 code for the pro-peptide region, exon 5 codes for the junctional region of the pro-peptide and mature enzyme regions, and exons 6–12 encode the mature enzyme region.

Eight differences in nucleotide sequence were found between the cDNA and genomic DNA when about 1400 nucleotides were compared (table 2). Five differences were located in the 3′-non-coding region and three in the coding region, but none were found in the 5′-non-coding region. Two changes in the coding region (cytosine-471 and adenine-693) do not cause an amino acid change, but cytosine-860, which was guanine in the cDNA, causes an amino acid change from glycine to alanine. These differences are probably due to sequence polymorphism.

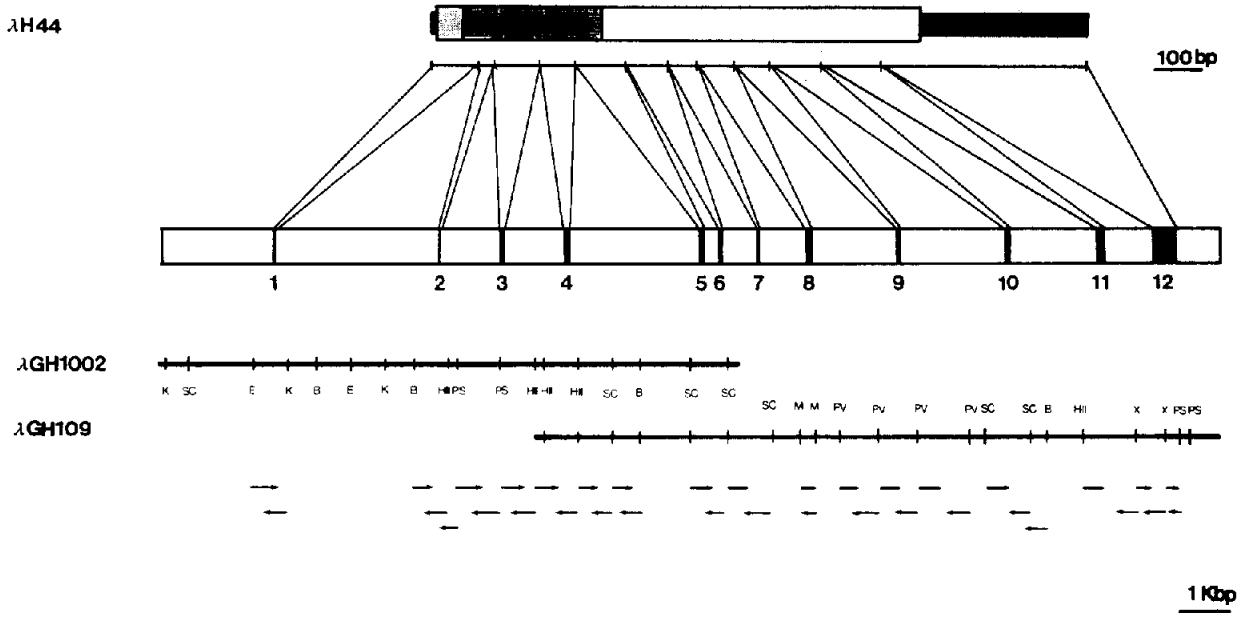

Fig.1. Restriction map and schematic structure of rat cathepsin H gene. λH44 is a full-length cDNA clone and its schematic structure is shown at the top [2]. Solid bars indicate 5′- and 3′-noncoding regions. The coding regions for the pre- and pro-peptides are stippled and horizontally shaded, respectively. The open area corresponds to the mature enzyme region. λGH1002 and λGH109, isolated using λH44 as probe, span 23.5 kbp. Exons are shown by the numbered filled areas. Arrows indicate the direction and length of sequencing. Restriction enzymes: *Kpn*I (K), *Sac*I (SC), *Eco*RI (E), *Bam*HI (B), *Hin*dIII (HIII), *Hin*cII (HII), *Pst*I (PS), *Mva*I (M), *Pvu*II (PV), *Xmn*I (X).

Table 1

Summary of rat cathepsin H gene structure

Exon	cDNA Nucleotide number	GC content (%)	Intron	Exon	Intron
1	−11– 85 (96 bp)	68.8	CGAGCAG	TGACCCC- - - - - - - - -	
				- - - - - - - - - GCCATAG	GTATGAG
2	86– 117 (32 bp)	34.4	TTTACAG	AAAAGTT- - - - - - - - -	
				- - - - - - - - - GAAACAG	GTCAGTG
3	118– 223 (106 bp)	49.1	GTTTTAG	CATCTTT- - - - - - - - -	
				- - - - - - - - - TTTAAAA	GTACTGG
4	224– 294 (71 bp)	42.3	TTTCCAG	TGGGATT- - - - - - - - -	
				- - - - - - - - - GCCTCAG	GTTAGTA
5	295– 399 (105 bp)	50.5	GCCTCAG	AATTGCT- - - - - - - - -	
				- - - - - - - - - GAACCAG	GTATGTC
6	400– 486 (87 bp)	57.5	TCTCTAG	GGGGCCT- - - - - - - - -	
				- - - - - - - - - GACCTTG	GTAAGAA
7	487– 542 (56 bp)	55.4	CCTCTAG	GCTGAGC- - - - - - - - -	
				- - - - - - - - - GCCAAGG	GTATGTT
8	543– 624 (82 bp)	54.9	CTTTCAG	AGGTCTC- - - - - - - - -	
				- - - - - - - - - AGGCAAG	GTAACTG
9	625– 693 (69 bp)	46.4	CGAACAG	AATGGTC- - - - - - - - -	
				- - - - - - - - - CACACTA	GTGAGTA
10	694– 800 (107 bp)	43.9	TTTCCAG	AATGATG- - - - - - - - -	
				- - - - - - - - - ACTCCAG	GTAAGCT
11	801– 926 (126 bp)	44.4	TTTCCAG	TAACTCC- - - - - - - - -	
				- - - - - - - - - ACAATGG	GTAAGAG
12	927–1359 (433 bp)	50.1	TACCCAG	GTACTTC	

The exons of the rat cathepsin H gene are numbered from 5′ to 3′ in the direction of transcription. The 5′-end of exon 1 was determined by S_1 mapping analysis [8]. The 3′-end of exon 12 represents the poly(A)$^+$ addition site

Table 2

Differences between the nucleotide sequences of the cDNA and genomic DNA

cDNA	Amino acid residue	Genomic DNA	Amino acid residue
471	45		
GGG	Gly	GGC	no change
693	118		
CTC	Leu	CTA	no change
860	174		
GGA	Gly	GCA	Ala
1023			
ACT	3′-noncoding	AAT	–
1119			
CGC	3′-noncoding	CGG	–
1174			
CCA	3′-noncoding	CCG	–
1203			
TTG	3′-noncoding	TTC	–
1207			
TTG	3′-noncoding	TTC	–

The numbers above the nucleotides and the amino acids are those of the cDNA [2]. Positions where differences occur are underlined

3.2. *The gene structure of rat cathepsin H and its comparison with other cysteine proteinase genes*

In the cathepsin H gene, intron break-points are not found at the junctions of the pre-peptide, propeptide and mature enzyme regions. The region around the active site cysteine, which is highly conserved among cysteine proteinases, is split by an intron. Further, five introns interrupt the two active site amino acid residues, Cys-26 and His-166. These facts indicate that the gene structure of this proteinase does not correspond well to the functional unit. These characteristics are also observed for the gene structures of the other cysteine proteinases thus far examined, e.g. aleurain, CP1, CP2 and CANP [6,12,13]. It should be noted that an intron-exon junction is always located immediately before or after the active site cysteine (fig.2).

The gene structures of cathepsin H and aleurain are significantly different, although they show

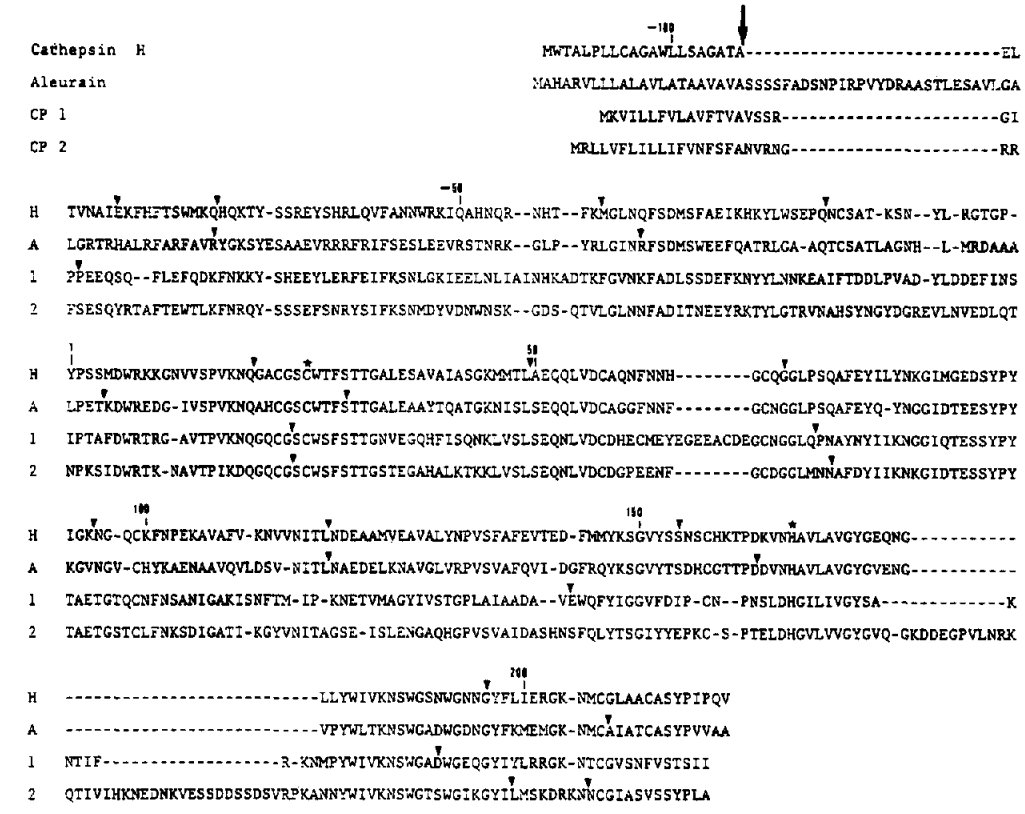

Fig.2. Comparison of intron positions of various cysteine proteinases. Amino acids are numbered starting with the N-terminus of mature rat cathepsin H and negative numbers indicate the pre- and pro-peptide. Amino acid sequences of aleurain [6], CP1 and 2 [12] are aligned for maximum identity to that of cathepsin H. The arrow indicates the junction between the pre- and pro-peptide. Arrowheads indicate intron positions. Asterisks show the active site residues.

61.8% amino acid sequence identity in the mature enzyme region. Cathepsin H is encoded by 12 exons, whereas the gene for aleurain comprises 8 exons. Only two intron-exon junctions are found in the same positions. This presumably indicates exon-intron rearrangement after fusion of the ancestral gene segments. Whittier et al. [6] have proposed that the aleurain gene arose by fusion of two ancestral genes. In the case of cathepsin H, however, clear evidence supporting the above hypothesis has not been obtained despite its high sequence homology to aleurain. Judging from the gene structure and GC content of the exons, the cathepsin H gene may be composed of 4 rather than 2 ancestral gene segments, i.e. exon 1, exons 2–4, exons 5–8 and exons 9–12, corresponding to the pre-peptide, pro-peptide and front and rear portions of the mature enzyme region, respectively.

To our knowledge, this is the first report on the gene structure of a mammalian cysteine proteinase. Further study of the gene structures of other cysteine proteinases may clarify their molecular evolution.

Acknowledgements: We thank H. Sorimachi and S. Ishida for the construction of the rat genomic library. We thank Drs S. Imajoh-Ohmi, Y. Emori and S. Ohno for helpful discussions on the techniques of gene cloning. This work was supported in part by research grants from the Ministry of Education, Science and Culture of Japan and the Yamanouchi Foundation for Research on Metabolic Disorders.

REFERENCES

[1] Katunuma, N. and Kominami, E. (1983) Curr. Top. Cell Regul. 22, 71–101.
[2] Ishidoh, K., Imajoh, S., Emori, Y., Ohno, S., Kawasaki, H., Minami, Y., Kominami, E., Katunuma, N. and Suzuki, K. (1987) FEBS Lett. 226, 33–37.

[3] Nishimura, Y. and Kato, K. (1987) Biochem. Biophys. Res. Commun. 148, 329–334.
[4] Kominami, E., Tsukahara, T., Hara, K. and Katunuma, N. (1988) FEBS Lett. 231, 225–228.
[5] Takio, K., Towatari, T., Katunuma, N., Teller, D.C. and Titani, K. (1983) Proc. Natl. Acad. Sci. USA 80, 3666–3670.
[6] Whittier, R.F., Dean, D.A. and Rogers, J.C. (1987) Nucleic Acids Res. 15, 2515–2535.
[7] Sorimachi, H., Emori, Y., Kawasaki, H., Kitajima, K., Inoue, S., Suzuki, K. and Inoue, Y. (1988) J. Biol. Chem. 263, 17678–17684.
[8] Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
[9] Hattori, M. and Sakaki, Y. (1986) Anal. Biochem. 152, 232–238.
[10] Brethnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978) Proc. Natl. Acad. Sci. USA 75, 4853–4857.
[11] Maniatis, T., Goodbourn, S. and Fischer, J.A. (1987) Science 236, 1237–1245.
[12] Pears, C.J., Mahbubani, H.M. and Williams, J.G. (1985) Nucleic Acids Res. 13, 8853–8866.
[13] Emori, Y., Ohno, S., Tobita, M. and Suzuki, K. (1986) FEBS Lett. 194, 249–252.