Volume 273, number 1,2, 195–199 FEBS 09040 October 1990

Gene structure of mouse cathepsin B

Marc Ferrara, Franck Wojcik, Houria Rhaissi, Sylvie Mordier, Marie-Paule Roux and Daniel Béchet

Unité de Recherches sur l'expression des protéases, SRV Theix, Institut National de la Recherche Agronomique, 63122 Ceyrat, France

Received 18 August 1990

The structure of a genomic DNA fragment encoding mouse cathepsin B was characterized. The genomic insert spans 15 kbp and contains 9 exons encoding the 339 amino acid residues of mouse preprocathepsin B. Intron break-points are not found at the junctions of the pre-peptide, pro-peptide and mature enzyme. Like other cysteine proteinase genes, the region around the cysteinyl active site is split by an intron, but in contrast with cathepsins L and H the intron break-point is located immediately after the active site.

Cathepsin B; Gene structure; Cysteine proteinase

1. INTRODUCTION

Members of the cysteine proteinase family include papain from papaya plant, aleurain from aleurone cells, proteinase CP1 and CP2 from *Dictyostelium discoideum* and cathepsins B, H and L from mammalian lysosomes [1–4]. Cathepsin B, together with cathepsins H and L, play an important role in intracellular protein catabolism [5] and may also be involved in tumor metastasis [6,7]. The primary structure of cathepsin B has been determined at the protein level [8] and immunological investigations on post-translational processing of cathepsin B have shown that this proteinase is synthesized as a preproenzyme form [9–11]. Cloning of the cDNA for cathepsin B has revealed the structure of its primary translation product [12–14]. Similar studies involving cDNA cloning have been reported for aleurain [3], rat cathepsin H [15] and rat cathepsin L [16]. The gene structure of rat cathepsin H [17] and cathepsin L [18] have been described recently. However, there is yet no data on cathepsin B gene structure.

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

2. MATERIALS AND METHODS

2.1. *Materials*

The sources of materials used in this work are as follows: restriction enzymes from Bethesda Research Laboratory and Boehringer Mannheim; *Bal*31 enzyme from Promega; T7 DNA polymerase sequencing kit from Pharmacia; (α-^{32}P)dATP, (α-^{35}S)dATP, Hybond N$^+$ membrane and Hyper-films MP from Amersham.

Correspondence address: M. Ferrara, Unité de Recherches sur l'expression des protéases, SRV Theix, Institut National de la Recherche Agronomique, 63122 Ceyrat, France

Abbreviations: kbp, kilobase pairs; bp, base pairs; CP1, cysteine proteinase 1; CP2, cysteine proteinase 2

2.2. *Methods*

2.2.1. Genomic library screening The genomic library used was from Clontech (BL 1011). It contained a partial *Mbo*I digest of mouse BALB/c genomic DNA cloned into the vector λEMBL3. We screened 5×10^5 plaque forming units (pfu) with the rat cDNA insert of the prCB3 clone (a generous gift of Chan and Steiner). This 1.15 kbp *Eco*RI insert was labeled by nick translation [19] to a specific activity of $2\text{–}4 \times 10^8$ dpm/μg.

2.2.2. Subcloning and sequencing DNA inserts from one phage clones (λ32) were digested with appropriate restriction enzymes and then subcloned into the pUC18 and M13 phage vectors. DNA sequences were determined by the dideoxynucleotide chain termination method of Sanger [20] using T7 DNA polymerase (Pharmacia).

3. RESULTS

3.1. *Isolation, mapping and sequence analysis of cathepsin B gene*

We screened 5×10^5 phages of the λEMBL3 mouse genomic library with the rat cathepsin B cDNA probe prCB3 [12]. Restriction endonuclease mapping showed that among the 6 positive clones obtained, 5 clones were different and one of them (λ32) was further characterized. Each digestion fragment which hybridized to the rat cDNA probe was separately subcloned into pUC18 plasmid and for each subclone restriction maps were completed. Using M13 clones, *Bal*31 digestion and 10 different oligonucleotide primers, the major part of the genomic clone λ32 spanning 15 kbp was sequenced. The clone λ32 contained all cathepsin B coding sequences. The structure of mouse cathepsin B gene, as revealed by restriction mapping, hybridization and DNA sequence analysis of λ32, is presented in Fig. 1.

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/90/$3.50

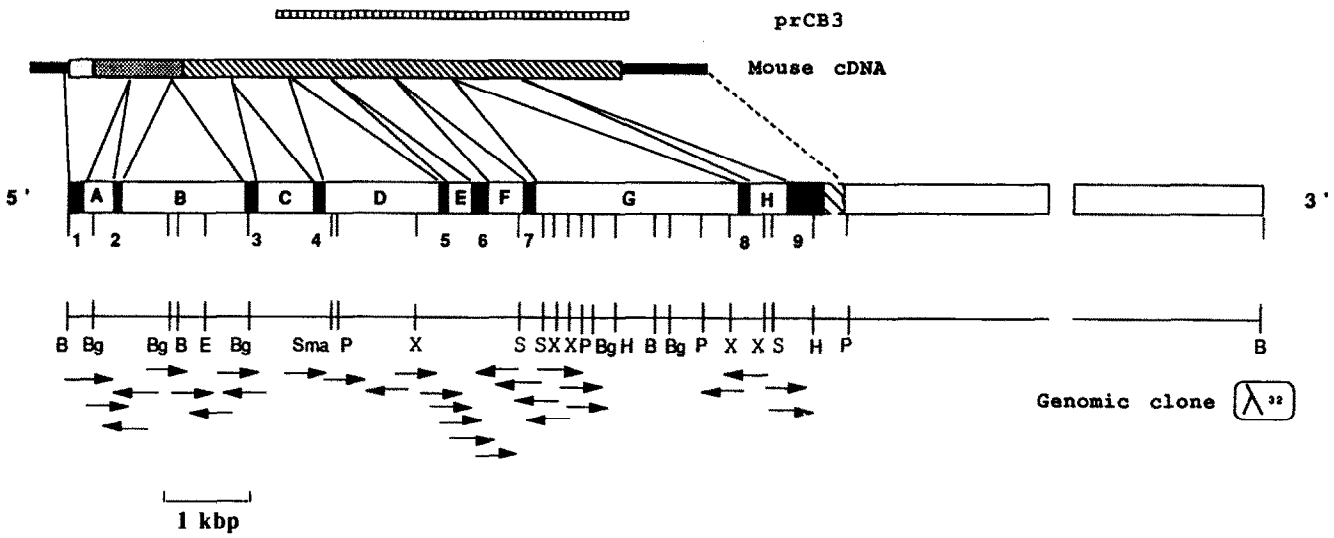

Fig. 1. Restriction map and schematic structure of mouse cathepsin B gene. The map was constructed from clone λ32. Each intron is alphabetically named from 5′ to 3′ end. Exons are represented by numbered filled areas. Arrows indicate direction and length of sequencing. Restriction enzyme sites are: B, *Bam*HI; Bg, *Bgl*II; E, *Eco*RI; H, *Hin*dIII; P, *Pst*I; Sma, *Sma*I; S, *Sst*I; X, *Xba*I. The structure of mouse cathepsin cDNA [13] is also shown. Solid bars indicate 5′- and 3′-non-coding regions. The coding regions for pre-peptide, pro-peptide and mature enzyme are represented by lightly dotted, heavily dotted and striped rectangles, respectively.

3.2. *Intron-exon structure*

Sequence analysis of the genomic clone λ32 showed that it contains 9 exons and 8 introns spanning a region of approximately 9 kbp. The length of each exon varied from 86 bp (exons 2 and 5) to more than 250 bp in exon 9 which size is not yet determined (Table I). A single initiation codon ATG is present in exon 1 which has a 15 bp untranslated region corresponding to the 3′-terminal sequence of the mRNA leader. Exon 1 encodes the signal peptide sequence of preprocathepsin B. The propeptide region is encoded by exons 1–3 and exons 3–9 encode the mature enzyme region.

Table I

Summary of mouse cathepsin B gene structure

Exon	Nucleotide number in cDNA		Size (bp)	GC content (%)
1	-	126	(>141)	51[a]
2	127	212	86	51
3	213	327	115	53
4	328	446	119	56
5	447	532	86	48
6	533	676	144	57
7	677	793	117	45
8	794	922	129	54
9	923	-	(>250)	47[b]

Nucleotide numbering of mouse cDNA starts at the initiation codon ATG.

[a]The genomic DNA fragment λ32 starts within Exon 1 (nucleotide −15 of mouse cDNA) and the exact size of this exon is not known (shown in parentheses). GC content was estimated for the 141 nucleotides that were sequenced in exon 1 [b]The exact size of exon 9 is not known (shown in parentheses). GC content was estimated for the 250 nucleotides that were sequenced in exon 9

Eleven differences in nucleotide sequence were found between cDNA and genomic DNA when 1110 bp were compared (Table II). The majority of these differences located in coding region are silent changes. Four differences are causing an amino acid change: for 3 of them the 'new' amino acid corresponds to the rat or the man cathepsin B sequence; only the guanine-479 which was adenine in the cDNA causes an amino acid change from Asn to Ser not found in the other cathepsin B proteins. These differences could be due to sequence polymorphism.

3.3. *Exon-intron boundary sequence*

Eight intron-exon boundary sites were determined and are shown in Table III. They were all in agreement with the GT/AG consensus rule [21]. For the cathepsin B gene, the intron splice phase [22] is type I for introns E, F, G and H (the intron interrupts the first and second bases of the codon), type II for introns B and D (the intron interrupts the second and third bases of the codon), and type 0 for introns A and C (this intron occurs between codons).

4. DISCUSSION

In this report we characterize the genomic structure of mouse cathepsin B gene. The clone (λ32) isolated contains all exons corresponding to the cDNA sequence except the leader region. So, this clone is defined as mouse cathepsin B gene. The gene analysed, containing at least 9 exons spanning approximately 9 kbp, shows the restricted size distribution typical for the most eukaryotic genes. A common characteristic observed

Table II

Differences in nucleotide sequences between the cDNA and the genomic DNA

Exon	cDNA nucleotide number	cDNA	Amino acid residue	Genomic DNA	Amino acid residue	Total difference in exon (%)
4	429	GGC	Gly	GGT	Gly	0.84
	479	AAC	Asn	AGC	Ser[b]	
5	520	GAT	Asp	AAT	Asn[a]	3.49
	529	ATA	Ile	GTA	Val[a]	
	543	CCG	Pro	CCA	Pro	
	591	CCG	Pro	CCA	Pro	
6	594	TGT	Cys	TGC	Cys	3.47
	600	GGG	Gly	GGA	Gly	
	603	GAG	Glu	GAA	Glu	
8	851	GTC	Val	GGC	Gly[a]	0.78
9	948	GAA	Glu	GAG	Glu	0.52[c]

[a]Amino acid residue found in rat or in human cathepsin B protein
[b]Amino acid residue not found in any cathepsin B protein so far examined
[c]The sequence published for mouse cathepsin B cDNA [13] encompasses only 191 bp of exon 9

for the gene structures of all cysteine proteinases so far examined (aleurain, CP1, CP2, cathepsins H, L) is that intron break-points are not found at the junctions of the pre-peptide, pro-peptide and mature enzyme regions. Accordingly, there is no evidence that the gene structure of cathepsin B corresponds to functional units (Fig. 2).

Important differences nevertheless distinguish gene structures of cysteine proteinases. Number and position of introns vary between cysteine proteinases. The genes encoding rat cathepsin H [17], rat cathepsin L [18], aleurain [21], CP1 and CP2 [24] contain 12, 8, 7, 5 and 4 introns, respectively, while the gene structure of mouse cathepsin B revealed at least 9 introns. In cathepsin B and H genes 5 introns interrupt the two active site Cys- and His-residues, instead of 2 in cathepsin L and CP1, 3 in aleurain and 1 in CP2 genes (Fig. 2). The region around the cysteinyl active site (Cys-29 in cathepsin B), which is highly conserved among cysteine proteinases, is always closed to an intron-exon junction. However, in cathepsin B and aleurain genes this junction is located immediately after the conserved sequence (Fig. 2) while in cathepsin H, cathepsin L and CP1 this junction is located immediately before. In CP2 gene an intron-exon junction interrupts the active site sequence [23]. The critical differences in both number and position of introns between cysteine proteinase genes supports the notion that the relation between the genes is not direct.

Because amino acid sequences of cysteine proteinases display a high degree of identity between each other, essentially in amino and carboxy terminal regions (con-

Table III

Splice junctions of the mouse cathepsin B gene

Consensus	5'donor AG*gtaagt.....IVS.....pppppppppcag* t	3'acceptor
IVS A	TGG CAG*gtaggc.....IVS A.....actctctttcag*	GCT GGA CGC Trp Gln Ala Gly Arg
IVS B	CCA GGA AG*gtgagt.....IVS B.....tctttgcctag*	G GTT GCG Pro Gly Ar g Val Ala
IVS C	TGT TGG*gtaagc.....IVS C.....atattaccacag*	GCA TTT GGG Cys Trp Ala Phe Glu
IVS D	GGG GAC GG*gtgagt.....IVS D.....tctctttttcag*	C TGT AAT Gly Asp Gl y Cys Asn
IVS E	CAT GTA G*gtgagt.....IVS E.....tccatctgtcag*	GC TGC TTA His Val G l y Cys Leu
IVS F	CAC TTT G*gtaagc.....IVS F.....ttctttttacag*	GG TAC ACT His Phe G l y Tyr Thr
IVS G	AAA TCA G*gtacca.....IVS G.....ttctttcttcag*	GA GTA TAC Lys Ser G l y Val Tyr
IVS H	GAT AAT G*gtgagt.....IVS H.....tctccatcccag*	GC TTC TTT Asp Asn G l y Phe Phe

The consensus sequence for splice donor and acceptor site is shown on the top line (p = pyrimidine). The residues surrounding the plice donor and acceptor sites for each cathepsin B exon are indicated. Note that the splice phase of introns A and C are type 0, the phases of introns E to H are type I, and the phases of introns B and D are type II. The exon sequences are shown in upper case letters; the intron sequences in lower case.

Fig. 2. Comparison of intron position in cysteine proteinases. Amino acid sequences of rat cathepsin H [17], rat cathepsin L [18], aleurain [21], CP1 and CP2 [24] are aligned for maximum identity to that of mouse cathepsin B. Amino acids are numbered starting from the N-terminus of the mature enzyme region of mouse cathepsin B. Negative numbers indicate pre- and pro-peptides regions. Arrows indicate intron insertion positions. Asterisks show the active site Cys- and His-residues.

taining the active site Cys- and His-residues, respectively [8,16]), it has initially been hypothesized [25] that cysteine proteinases derive from a common ancestral gene. In fact, judging from the gene structure and GC content of exons, Whittier et al. [24] have then suggested that the aleurain gene arose by fusion of 2 ancestral genes. Ishidoh et al. [17] have further proposed that cathepsin H gene is composed of 4 rather than 2 ancestral gene segments. For cathepsin B gene, the GC content of different exons is more homogeneous (see Table I) than for cathepsin H [17], cathepsin L [18] and aleurain [24]. In contrast to aleurain and cathepsins H and L, it is then difficult to ascribe different origins to the different exons of cathepsin B. We can only point out that the cathepsin B exon encoding the histidyl active site is 54% GC rich like the aleurain corresponding exon (55%) which is appreciably different from the corresponding exons in cathepsin L (43%) and cathepsin H (44%) (Table IV).

To discuss the relative variability of different regions of a protein family a functional divergence ratio (FDR) was defined by Creighton and Dardy [26] as 'the frequency of mutations that cause amino acid replacement at positions within the functional region of a protein relative to the frequency of replacement within the remainder of the protein'. Unlike the other protease family, the members of the cysteine proteinase family do not exhibit hypervariability of active site amino acid sequence [26]. This low FDR value observed for cathepsins reflects the preferential conservation during evolution of the functional region containing cysteine residue directly involved in the catalytic site of the enzyme. The comparison of cysteine proteinase gene structures demonstrates that the very conserved cysteinyl active site sequence has probably evolved in several ways. This confirms the essential role of this region for hydrolysing activity of cysteine proteinase and the critical role that cysteine proteinases play in vivo.

Table IV

Comparison between cysteine proteinase of the GC content of exons encoding active site regions

Exon encoding	Mouse cathepsin B (%)	Rat cathepsin L (%)	Rat cathepsin H (%)	Aleurain (%)
Cysteinyl active site	53	48	58	60
Histidyl active site	54	43	44	55

This report is, to our knowledge, the first description of cathepsin B gene structure. The region upstream the ATG initiation codon remains to be analysed. As reported before (see section 3), we have isolated four other clones by stringent screening of λEMBL3 genomic library with prCB3 insert. Restriction maps for these clones are clearly different from λ32 map. The analysis of these clones are in progress, but they already suggest the existence of several related cathepsin B gene sequences in BALB/c mouse genome.

Acknowledgements: We thank S.J. Chan for providing the rat cathepsin B cDNA clone prCB3; M. Renaud, J.C. Mercier and P. Gaye for valuable discussions during the initial stages of this work.

REFERENCES

[1] Mitchel, R.E.J., Chaiken, I.M. and Smith, E.L. (1970) J. Biol. Chem. 245, 3485–3492.
[2] Carne, A. and Moore, C.H. (1978) Biochem. J. 173, 73–83.
[3] Rogers, J.C., Dean, D. and Heck, R. (1985) Proc. Natl. Acad. Sci. USA 82, 6512–6516.
[4] North, M.J. and Harwood, J.M. (1979) Biochem. Biophys. Acta 566, 222–233.
[5] Natori, Y., Chiku, K., Oka, T. and Sato, A. (1989) in: Intracellular Proteolysis: Mechanisms and Regulations (Katunuma, N. and Kominami, E. eds) pp. 445–451, Japan Scientific Societies Press.
[6] Sloane, B.F., Honn, K.V., Sadler, J.G., Turner, W.A., Kimpson, J.J. and Taylor, J.D. (1982) Cancer Res. 42, 980–986.
[7] Sloane, B.F., Rozhin, J., Moin, K., Buck, M.R., Day, N.A. and Helmer, K.M. (1989) in: Intracellular Proteolysis: Mechanisms and Regulations (Katunuma, N. and Kominami, E. eds) pp. 527–533, Japan Scientific Societies Press.
[8] Takio, K., Towatari, T., Katunuma, N., Teller, D.C. and Titani, K. (1983) Proc. Natl. Acad. Sci. USA 80, 3666–3670.
[9] Steiner, D.F., Docherty, K. and Carroll, R. (1984) J. Cell Biochem. 24, 121–130.
[10] Nishimura, Y., Furuno, K. and Kato, K. (1988) Arch. Biochem. Biophys. 263, 107–116.
[11] Kominami, E., Katunuma, N. and Uchiyama, Y. (1989) in: Intracellular Proteolysis: Mechanisms and regulations (Katunuma, N. and Kominami, E. eds) pp. 52–60, Japan Scientific Societies Press.
[12] San Segundo, B., Chan, S.J. and Steiner, D.F. (1985) Proc. Natl. Acad. Sci. USA 82, 2320–2324.
[13] Chan, S.J., San Segundo, B., McCormick, M.B. and Steiner, D.F. (1986) Proc. Natl. Acad. Sci. USA 83, 7721–7725.
[14] Fong, D., Calhoun, D.H., Hsieh, W.T., Lee, B. and Wells, R.D. (1986) Proc. Natl. Acad. Sci. USA 83, 2909–2913.
[15] Ishido, K., Imajoh, S., Emori, Y., Ohno, S., Kawasaki, H., Minami, Y., Kominami, E., Katunuma, N. and Suzuki, K. (1987) FEBS Lett. 226, 33–37.
[16] Ishido, K., Towatari, T., Imajoh, S., Kawasaki, H., Kominami, E., Katunuma, N. and Suzuki, K. (1987) FEBS Lett. 223, 69–73.
[17] Ishido, K., Kominami, E., Katunuma, N. and Suzuki, K. (1989) FEBS Lett. 253, 103–107.
[18] Ishido, K., Kominami, E., Suzuki, K. and Katunuma, N. (1989) FEBS Lett. 259, 71–74.
[19] Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: a Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
[20] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463–5467.
[21] Breathnach, R. and Chambon, P. (1981) Annu. Rev. Biochem. 50, 349–383.
[22] Rogers, J. (1985) Nature 315, 458–459.
[23] Pears, C.J., Mahbubani, H.M. and Williams, J.G. (1985) Nucleic Acids Res. 13, 8853–8866.
[24] Whittier, R.F., Dean, D.A. and Rogers, J.C. (1985) Nucleic Acids Res. 15, 2515–2535.
[25] Barrett, A.J. (1984) Biochem. Soc. Trans. 12, 899–902.
[26] Creighton, T.E. and Dardy, N.J. (1989) Trends Biol. Sci. 14, 319–324.