Volume 226, number 1, 33–37 FEB 05402 December 1987

Molecular cloning and sequencing of cDNA for rat cathepsin H

Homology in pro-peptide regions of cysteine proteinases

Kazumi Ishidoh*+, Shinobu Imajoh*, Yasufumi Emori*, Shigeo Ohno*, Hiroshi Kawasaki*, Yasufumi Minami*, 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 +Department of Enzyme Chemistry, Institute of Enzyme Research, The University of Tokushima, Tokushima 770, Japan

Received 27 October 1987

A cDNA for rat cathepsin H was isolated and sequenced. The deduced protein comprising 333 amino acid residues is composed of a typical signal sequence (21 residues), a pro-peptide region (92 residues) and a mature enzyme region (220 residues). The amino acid sequence in the pro-peptide region, in particular, residues Phe-(–41) to Ser-(–29) of cathepsin H, is highly homologous to the pro-peptide regions of other cysteine proteinases. This homologous region may play a role in the processing of cysteine proteinases.

Cathepsin H; Cysteine proteinase; cDNA cloning; Propeptide region; Amino acid sequence

1. INTRODUCTION

Cysteine proteinases are widely distributed in plants and animals and participate in various cellular functions, e.g. intracellular protein breakdown, cell growth, biosyntheses of peptide hormones and tumor metastasis [1–3]. Cathepsins B, H and L are the most typical and well-characterized cysteine proteinases. The primary structures of cathepsins B and H have been determined at the protein level [4] and are highly homologous to that of papain. Studies on in vitro translation and immunological investigations have shown that cathepsins B and H are synthesized as preproenzymes and exist in lysosomes as mature enzymes after post-translational modifications [5–7]. Cloning of the cDNA for cathepsin B has revealed the structure of its primary translation product: 339 amino acid residues containing a signal peptide (17 residues), pro-peptide region (62 residues), mature enzyme region (254 residues) and C-terminal extension peptide (6 residues) [8]. Similar studies involving cloning of the cDNA for papain and rat cathepsin L [9,10] have also established the structure of the N-terminal extension peptides, i.e. a signal peptide and pro-peptide.

Here, we isolated cDNA for rat cathepsin H and determined its nucleotide sequence. The deduced amino acid sequence of cathepsin H is highly homologous to those of other cysteine proteinases in the pro-peptide region as well as in the mature enzyme region. The observation of this sequence homology in the pro-peptide region of cysteine

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

The nucleotide sequence presented here has been submitted to the EMBL/GenBank database under the accession number Y00708

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

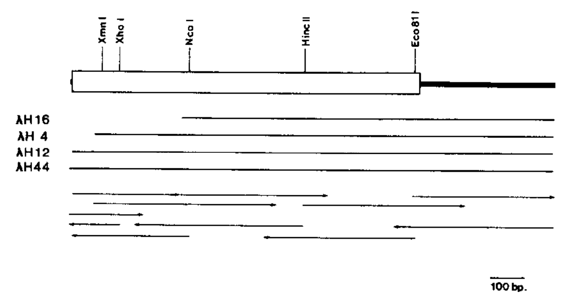

Fig.1. Restriction map and sequencing strategy of cDNA clones of cathepsin H. DNA sequencing was carried out using the dideoxy method in both directions. Arrows indicate the direction and length of sequencing. The open box indicates the amino acid coding region and thick lines indicate 5′- and 3′-non-coding regions.

proteinases may provide us with a clue to the understanding of the function of the pro-peptide region.

2. MATERIALS AND METHODS

Extraction of total RNA from rat kidney, selection of poly(A)$^+$ RNA and construction of a cDNA library were as described previously [10]. Synthetic oligonucleotide probes, CATH-ACT (5′-CCTGTGAAGAACCAGGGCGCCTGTGG-CTCCTGCTGGACCTTCTCC-3′, 45-mer) corresponding to the amino acid sequence around the active-site cysteine residue (from Pro-16 to Ser-30) and CATH-H (5′-CAGTGCAAGTTCAACCCT-GAGAAGGCTGTGGCCTTTGTGAAGAATG-

```
                                                                                                                    -1
                                                                                                                   GCG
1                                                      50                                                      100
ATGTGGACTGCGCTGCCCCTGCTGTGCGCTGGGGCCTGGCTGCTGAGTGCTGGGGCCACCGCTGAGCTGACCGTGAATGCCATAGAAAAGTTTCACTTTACGTCATGGATGAAACAGCAT
MetTrpThrAlaLeuProLeuLeuCysAlaGlyAlaTrpLeuLeuSerAlaGlyAlaThrAlaGluLeuThrValAsnAlaIleGluLysPheHisPheThrSerTrpMetLysGlnHis
-113                                         -100             ↑
                                   150                                                     200
CAAAAGACGTACAGCTCGAGGGAGTACAGCCACAGGCTGCAGGTGTTTGCCAACAACTGGAGGAAGATTCAAGCCCACAACCAGAGGAACCACACATTTAAAATGGGATTGAACCAGTTT
GlnLysThrTyrSerSerArgGluTyrSerHisArgLeuGlnValPheAlaAsnAsnTrpArgLysIleGlnAlaHisAsnGlnArgAsnHisThrPheLysMetGlyLeuAsnGlnPhe
                                                                             -50                *
     250                                                        300                                                      350
TCAGATATGAGCTTTGCCGAAATAAAACACAAATACCTCTGGTCAGAGCCTCAGAATTGCTCAGCCACCAAAAGTAACTACCTCCGTGGTACTGGCCCCTACCCATCCTCCATGGACTGG
SerAspMetSerPheAlaGluIleLysHisLysTyrLeuTrpSerGluProGlnAsnCysSerAlaThrLysSerAsnTyrLeuArgGlyThrGlyProTyrProSerSerMetAspTrp
                                                             *                                                  -1↑ 1
                                         400                                                      450
AGGAAGAAAGGAAATGTCGTTTCACCAGTGAAGAACCAGGGGGCCTGTGGCAGCTGCTGGACTTTCTCAACCACTGGGGCCCTAGAGTCAGCTGTGGCTATTGCCAGTGGGAAAATGATG
ArgLysLysGlyAsnValValSerProValLysAsnGlnGlyAlaCysGlySerCysTrpThrPheSerThrThrGlyAlaLeuGluSerAlaValAlaIleAlaSerGlyLysMetMet

                    500                                                    550                                                    600
ACCTTGGCTGAGCAGCAGCTGGTGGATTGTGCCCAGAACTTCAACAATCATGGCTGCCAAGGAGGTCTCCCCAGCCAGGCCTTCGAGTACATCCTGTACAACAAGGGCATCATGGGAGAG
ThrLeuAlaGluGlnGlnLeuValAspCysAlaGlnAsnPheAsnAsnHisGlyCysGlnGlyGlyLeuProSerGlnAlaPheGluTyrIleLeuTyrAsnLysGlyIleMetGlyGlu
     50
                                                     650                                                    700
GACAGCTACCCTTACATAGGCAAGAATGGTCAGTGCAAATTCAACCCAGAAAAGGCCGTCGCGTTCGTCAAGAATGTTGTCAACATCACACTCAATGATGAGGCTGCAATGGTAGAGGCT
AspSerTyrProTyrIleGlyLysAsnGlyGlnCysLysPheAsnProGluLysAlaValAlaPheValLysAsnValValAsnIleThrLeuAsnAspGluAlaAlaMetValGluAla
                                         100                                                *
                              750                                                     800
GTGGCCCTATACAATCCTGTGAGCTTTGCCTTTGAGGTGACTGAAGATTTTATGATGTATAAAAGTGGTGTCTACTCCAGTAACTCCTGTCATAAAACTCCAGATAAAGTAAACCATGCA
ValAlaLeuTyrAsnProValSerPheAlaPheGluValThrGluAspPheMetMetTyrLysSerGlyValTyrSerSerAsnSerCysHisLysThrProAspLysValAsnHisAla
                                                                       150
          850                                                        900                                                   950
GTCCTGGCTGTTGGCTATGGAGAACAGAATGGATTACTCTACTGGATTGTGAAAAACTCTTGGGGCTCCAACTGGGGGAACAATGGGTACTTCCTCATTGAGCGTGGAAAGAACATGTGT
ValLeuAlaValGlyTyrGlyGluGlnAsnGlyLeuLeuTyrTrpIleValLysAsnSerTrpGlySerAsnTrpGlyAsnAsnGlyTyrPheLeuIleGluArgGlyLysAsnMetCys
                              ↑                                                                                    200
                                        1000                                                   1050
GGCCTGGCTGCCTGTGCCTCCTACCCCATCCCTCAGGTATAAGCCACGGCTGCACAGGCCAACTGCTTGGCAGACAAAGGGAGGAACTGGTCCTACGATGAGAATGCCGCCCTGGAGAAA
GlyLeuAlaAlaCysAlaSerTyrProIleProGlnVal***

                    1100                                                   1150                                                  1200
GTTGTTCAGAAATCCACCCAGAGGCCCTCTCACTCCTGAGTCTAGACGCCTAAAGACAAGTAAGGAAGAACTTGACCAGCAACAAGCCCGCCCATGTGACGACATCACCAGCCATACGCT

                                                  1250                                                   1300
TTGTTTGAATATGGTTTTTAAATGACCCAAAACCATGTGGACCTAGAATCTTCTCTTTTCAGCTCTCTTCATATAGGGAAAGCTCCGATGGCTTACCTTTTCTATGTTGTATATTCAATA

                              1350                        1380
AATGCAGTGAACGCCTGCCTGGGCTGGGCCTAGACCCCTAAAAAAAAAAAAAAAAA
```

Fig.2. Nucleotide and deduced amino acid sequences of cDNA clones for rat cathepsin H. The nucleotide sequence is numbered starting at the initiation codon ATG and negative numbers show the 5′-non-coding region. The deduced amino acid sequence for the precursor of cathepsin H, shown under the nucleotide sequence, is numbered beginning at the N-terminal residue of the mature enzyme region. Negative numbers indicate pre- and pro-sequences. Arrows indicate cleavage sites in post-translational processing. Underlining indicates the polyadenylation signal. Single and triple asterisks indicate potential glycosylation sites and the termination codon, respectively.

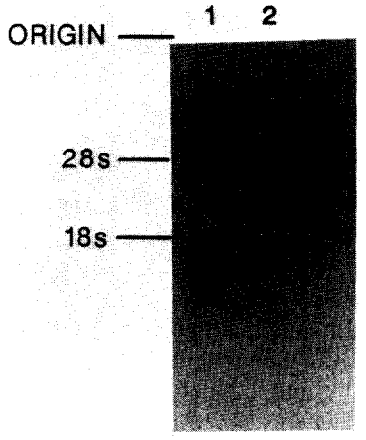

Fig.3. RNA blot hybridization analysis. Poly(A)$^+$ RNA from rat kidney (lane 1, 0.5 μg; lane 2, 0.1 μg) was analyzed. λH12 (8–1378) was used as a probe. Positions of rat ribosomal RNAs (28 S and 18 S) are indicated.

TGGTG-3′, 51-mer) corresponding to residues Gln-98 to Val-114 in cathepsin H were prepared as in [10]. RNA blot hybridization analysis was the same as in [10], except that λH12 was used as a probe. Nucleotide sequencing was carried out as in [11].

3. RESULTS

3.1. *Isolation of the cDNA clone for rat cathepsin H*

A rat kidney cDNA library (about 7.0×10^4 independent plaques) was screened with the synthetic oligonucleotide probes, CATH-ACT and CATH-H. cDNA inserts of 20 clones selected randomly from 105 recombinant clones which hybridized with both probes, were analyzed by agarose gel electrophoresis. Among them, three inserts (λH4, λH12, λH16) were sequenced (fig.1). Since even the longest insert (λH12) lacked the 5′-region of the mRNA, the library was re-screened with a fragment generated by digestion of λH12 with *Eco*RI and *Xmn*I (82 bp) under stringent conditions. The clone (λH44) which had the longest insert among 40 positive clones as judged from the restriction map of *Eco*RI and *Xho*I was sequenced.

3.2. *Nucleotide sequence analysis and deduced amino acid sequence of rat cathepsin H*

The nucleotide sequence of λH44 and the deduced amino acid sequence are shown in fig.2. Although a clone which contains a longer sequence at the 5′-region has not yet been obtained, we presume that Met at −113 is the initiation methionine for following reasons. Firstly, the N-terminal region from residues −113 to −93 is composed of hydrophobic amino acids, which is a typical feature of signal peptides [12]. Secondly, the lengths of the presumptive pre- and pro-peptides are nearly the same as those for other cysteine proteinases (cf. fig.4). Thirdly, assuming that Met-(−113) is the initiation methionine, the M_r values of preprocathepsin H and procathepsin H are 37103 and 35003, respectively. These values are consistent with the estimates made from the results of in vitro translation [6]. Lastly, RNA blot

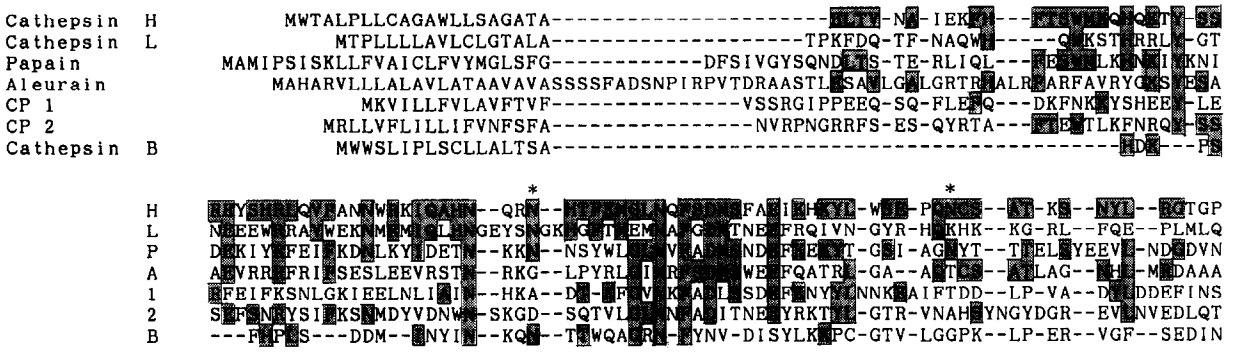

Fig.4. Comparison of amino acid sequence in the pre- and pro-peptide regions of rat cathepsin H with those of rat cathepsin L [10], papain [9], aleurain [16], CP1 [14], CP2 [15] and rat cathepsin B [8]. Amino acid residues identical between cathepsin H and other proteins are indicated by stippled boxes. Gaps, denoted by dashes, are introduced for maximal alignment using the computer program package IDEAS [18]. Cathepsin B is aligned manually because of its shorter length. Asterisks show potential glycosylation sites.

Table 1

Amino acid sequence homology of cathepsin H with other cysteine proteinases

	Pro-peptide region	Mature enzyme region	Total
Cathepsin L	24.5 (39.2)	44.2 (60.7)	38.7 (53.7)
Papain	34.0 (48.5)	30.9 (44.7)	30.6 (45.6)
Aleurain	29.0 (53.0)	61.8 (76.8)	48.3 (64.8)
CP1	20.8 (32.3)	36.7 (55.5)	31.1 (48.4)
CP2	22.4 (42.9)	34.1 (47.3)	30.0 (44.7)

Sequence identity (%) was calculated using a computer program package IDEAS (see fig.4) [18]. Values in parentheses were calculated assuming that conservative amino acids are identical. Gaps are regarded as mismatches. Cathepsin B was not included because of its shorter length. Values given as percentages

hybridization analysis shows that the size of the mRNA for cathepsin H is about 1.8 kb (fig.3), which is approximately the same as that of the insert of λH44.

The signal peptide and the pro-peptide are predicted to be residues −113 to −93 and −92 to −1, respectively, by comparing the deduced amino acid sequence with that of the mature enzyme [4] and by following the method in [12]. In the pro-peptide region, two potential glycosylation sites, Asn-(−44) and Asn-(−15) are found together with one additional site in the mature enzyme region which has been noted before [4]. These sites in the pro-peptide region are also found in the corresponding region of papain [9], although such sites do not exist in cathepsin L [10]. Whether or not these sites in the pro-peptide region are glycosylated remains unclear at present.

The deduced amino acid sequence of the mature enzyme region coincides completely with that determined at the protein level. A C-terminal extension peptide which is found in cathepsin B [8] cannot be found in cathepsin H.

A consensus polyadenylation signal is seen at 1317–1322 [13] followed by a poly(A) tail.

4. DISCUSSION

This is the first sequence determination of the primary translation product of cathepsin H. The 5′-termini of most clones obtained for cathepsin H (~40 clones) terminated around nucleotide no.12. Various attempts to obtain longer 5′-sequences have not been successful. mRNA having a high GC content often forms a stable secondary structure which inhibits efficient synthesis of cDNA. Presumably, this is one of the reasons why longer cDNA clones are not obtained.

Recently, cDNA cloning for various cysteine proteinases, including cathepsin B from human, mouse and rat [8], cathepsin L from rat [10], papain from *Carina papaya* fruit [9], cysteine proteinases 1 and 2 (CP1 and CP2) from *Dictyostelium discoideum* [14,15] and aleurain from barley aleurone cells [16], has revealed the amino acid sequences of their preproenzymes. As has already been discussed, the mature enzyme regions of these proteinases show very high sequence homology [4,16]. Their pro-peptide regions also show significant sequence homology when they are aligned as shown in fig.4. Although the sequence homology in the pro-peptide region is independent of that in the mature enzyme region and is clearly lower, it amounts to 21–34% when the pro-sequence of cathepsin H is compared with others (table 1). The homology increases to 32–53% when conservative amino acids are regarded as identical [17]. In particular, the region from Phe-(−41) to Ser-(−29) of cathepsin H is highly homologous to other cysteine proteinases except for cathepsin B.

The homology found in the pro-peptide region may be important when we consider the function of the pro-peptide in cysteine proteinases.

ACKNOWLEDGEMENT

This work was supported in part by research grants from the Ministry of Education, Science and Culture of Japan.

REFERENCES

[1] Katunuma, N. and Kominami, E. (1983) Curr. Top. Cell. Regul. 22, 71–101.
[2] Docherty, K., Hutton, J.C. and Steiner, D.F. (1984) J. Biol. Chem. 259, 6041–6044.
[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] Takio, K., Towatari, T., Katunuma, N., Teller, D.C. and Titani, K. (1983) Proc. Natl. Acad. Sci. USA 80, 3666–3670.
[5] Steiner, D.F., Docherty, K. and Carroll, R. (1984) J. Cell. Biochem. 24, 121–130.
[6] Nishimura, Y. and Kato, K. (1987) Biochem. Biophys. Res. Commun. 146, 159–164.
[7] Wiederanders, B. and Kirschke, H. (1986) Biomed. Biochim. Acta 45, 1421–1431.
[8] Chan, S.J., Segundo, B.S., McCormick, M.B. and Steiner, D.F. (1986) Proc. Natl. Acad. Sci. USA 83, 7721–7725.
[9] Cohen, L.W., Coghlan, V.M. and Dihel, L.C. (1986) Gene 48, 219–227.
[10] Ishidoh, K., Towatari, T., Imajoh, S., Kawasaki, H., Kominami, E., Katunuma, N. and Suzuki, K. (1987) FEBS Lett., in press.
[11] Hattori, M. and Sakaki, Y. (1986) Anal. Biochem. 152, 232–238.
[12] Von Heijne, G. (1983) Eur. J. Biochem. 133, 17–21.
[13] Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211–214.
[14] Williams, J.G., North, M.J. and Mahbubani, H. (1985) EMBO J. 4, 999–1006.
[15] Pears, C.J., Mahbubani, H.M. and Williams, J.G. (1985) Nucleic Acids Res. 13, 8853–8866.
[16] Rogers, J.C., Dean, D. and Heck, G.R. (1985) Proc. Natl. Acad. Sci. USA 82, 6512–6516.
[17] Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) in: Atlas of Protein Sequence and Structure, vol.5, suppl.3, pp.345–352, National Biomedical Research Foundation, Silver Spring, MD.
[18] Kanehisa, M. (1982) Nucleic Acids Res. 10, 183–196.