Molecular Cloning of cDNA for Rat Cathepsin C

CATHEPSIN C, A CYSTEINE PROTEINASE WITH AN EXTREMELY LONG PROPEPTIDE

Kazumi Ishidoh¶, Daisaku Muno, Noboru Sato, and Eiki Kominami

From the Department of Biochemistry, Juntendo University School of Medicine, Hongo 2-1-1, Bunkyo-ku, Tokyo 113, Japan

A cDNA for rat cathepsin C (dipeptidylaminopeptidase I) was isolated. The deduced amino acid sequence of cathepsin C comprises 462 amino acid residues: 28 NH₂-terminal residues corresponding to the signal peptide, 201 residues corresponding to the propeptide, and 233 COOH-terminal residues corresponding to the mature enzyme region. Four potential glycosylation sites were found, three located in the propeptide region, and one in the mature enzyme region. The amino acid sequence of mature cathepsin C has 39.5% identity to that of cathepsin H, 35.1% to that of cathepsin L, 30.1% to that of cathepsin B, and 33.3% to that of papain. Cathepsin C, therefore, is a member of the papain family, although its propeptide region is much longer than those of other cysteine proteinases and shows no significant amino acid sequence similarity to any other cysteine proteinase.

Lyosomal cysteine proteinases play important roles in intracellular protein degradation, antigen presentation, tumor metastasis, and muscular dystrophy (1-3), and their inhibition causes various pathogenic states. For example, the injection of leupeptin, a cysteine proteinase inhibitor, or chloroquine, a general lyosomal enzyme inhibitor, into young rat brain induces the subsequent formation of ceroid-lipofuscin-like granular aggregates (4). Thus, cysteine proteinases participate in various biological actions and homeostasis.

Recent cDNA cloning experiments have revealed the amino acid sequences of the cysteine proteinases cathepsins B, H, and L (5-7). As their amino acid sequences are similar to one another and also to that of papain, they belong to the papain family, although its propeptide region is much longer than those of other cysteine proteinases and shows no significant amino acid sequence similarity to any other cysteine proteinase.

A cDNA for rat cathepsin C comprises 462 amino acid residues: 28 NH₂-terminal residues corresponding to the signal peptide, 201 residues corresponding to the propeptide, and 233 COOH-terminal residues corresponding to the mature enzyme region. Four potential glycosylation sites were found, three located in the propeptide region, and one in the mature enzyme region. The amino acid sequence of mature cathepsin C has 39.5% identity to that of cathepsin H, 35.1% to that of cathepsin L, 30.1% to that of cathepsin B, and 33.3% to that of papain. Cathepsin C, therefore, is a member of the papain family, although its propeptide region is much longer than those of other cysteine proteinases and shows no significant amino acid sequence similarity to any other cysteine proteinase.

* This work was supported in part by Grant-in-Aid for Encouragement of Young Scientists 02770133, Grant-in-Aid for Cancer Research 02151027, Grant-in-Aid for Co-operative Research 63304034 and Grant-in-Aid for Scientific Research 02454159 from the Ministry of Education, Science and Culture of Japan, Grant 63-2 from the Ministry of Health and Welfare, Japan, and by Aid for Project Medical Research Grant 9002 from Juntendo University. The costs of publication of this paper (including "Materials and Methods," part of "Results," Figs. 1a-4a, and Tables 1a-3a) are presented in miniprint at the end of this paper. Miniprint is easily read with the aid of a standard magnifying glass. Full size photocopies are included in the microfilm edition of the Journal that is available from Waverly Press.

** Portions of this paper (including "Materials and Methods," part of "Results," Figs. 1a-4a, and Tables 1a-3a) are presented in miniprint at the end of this paper. Miniprint is easily read with the aid of a standard magnifying glass. Full size photocopies are included in the microfilm edition of the Journal that is available from Waverly Press.
Fig. 1. Schematic structure, restriction map, and sequencing strategy for cathepsin C. The schematic structure of AC14, a full-length cDNA clone of cathepsin C, is shown at the top. The 5'- and 3'-noncoding regions are indicated by the black bars. The signal peptide and propeptide regions are indicated by stippled and horizontal shadows, respectively. The coding region of the mature enzyme is indicated by the open bar. The restriction map of AC14 is depicted directly below the schematic structure. The restriction enzymes are abbreviated as follows: K, KpnI; M, MvaI; B, BamHI; E, EcoRI; H, HindIII. Arrows indicate the lengths and directions of sequencing.

Fig. 2. Nucleotide sequence and deduced amino acid sequence of cathepsin C. Nucleotide numbering starts at the initiation codon. The triple asterisks indicate the termination codon and an in-frame termination codon in the 5'-noncoding region. The double underlining indicates a poly(A)" addition signal. Amino acid numbering starts from the NH2 terminus of the mature enzyme. Negative numbers indicate the pre- and propeptide regions. The underlines and broken underlines indicate the amino acid sequences of peptides derived from the purified enzyme that were identified or not identified in our system at the protein level, respectively (see the Miniprint). Asterisks indicate potential glycosylation sites. Arrows indicate possible NH2 termini: (A), (B), and (C) indicate NH2 termini of the propeptide, the heavy chain, and the light chain of the mature enzyme, respectively.

(CAACCATGGG) corresponds to Kozak's rule (31) and an in-frame termination codon (−45 to −43) is found in the 5'-upstream sequence. The deduced amino acid sequence of the open reading frame codes for 462 amino acid residues. In the 3'-noncoding region, a poly(A)" additional signal was found at 1773–1778 (32).

Preprocathepsin C comprises 462 amino acid residues containing three functional domains. The 28 NH2-terminal amino acid residues correspond to the signal peptide in accordance with their hydropathy and prediction of the signal peptide processing site (33); the next 201 residues correspond to the propeptide region; the 233 COOH-terminal residues correspond to the mature enzyme region. The amino acid sequences derived from peptide fragments of the purified cathepsin C (see the Miniprint) were found in the deduced amino acid sequence of the cDNA. NH2-terminal analysis showed that purified cathepsin C is processed to a two-chain form: the amino acid sequences of the NH2 termini are LPESWDW and DPFPNFEL (see the Miniprint). Calculating from the amino acid sequence, the molecular weights of preprocathepsin C, procathepsin C, cathepsin C, and the heavy and light chains of cathepsin C are 52234.36, 49342.95, 26056.93, 18362.32, and 7712.63, respectively. Four potential N-glycosylation sites were found: three in the propeptide region and one in the mature enzyme region. Some of these sites must be glycosylated since the lysosomal enzymes bear mannose 6-phosphate as a recognition marker (34, 35).

A search for homology in the National Biomedical Research Foundation protein sequence data bank revealed no significant similarities to other proteins except cysteine proteinases.

RNA Blot Hybridization Analysis—RNA blot hybridization analyses revealed that two sizes of mRNAs for cathepsin C, 2.1 and 2.7 kb, are expressed in all tissues (Fig. 3). Judging from the nucleotide sequence length of the cDNA and the ratio of the two sizes of mRNA, the isolated cDNAs for cathepsin C probably correspond to the 2.1-kb transcript. The

Fig. 3. RNA blot hybridization analysis. Poly(A)" RNA (2.5 μg/lane) was electrophoresed on a 1% denatured agarose gel containing 2.2 M formamide. The RNA samples were prepared from testis (lane 1), heart (lane 2), skeletal muscle (lane 3), aorta (lane 4), adrenal gland (lane 5), submandibular gland (lane 6), large intestine (lane 7), small intestine (lane 8), stomach (lane 9), esophagus (lane 10), spleen (lane 11), pancreas (lane 12), liver (lane 13), kidney (lane 14), lung (lane 15), and brain (lane 16). The 5' side of the EcoRI fragment nucleotides −56 to 806 in Fig. 2) was labeled with [α-32P]dCTP by the multiprime labeling system. Prehybridization, hybridization, and washing conditions are described in the text. Total RNA was electrophoresed on the same gel as the standards. The positions of 28 and 18 S ribosomal RNA are indicated.
The amino acid sequences of cathepsins C and H. Identical amino acids are aligned to give maximal identity to that of cathepsin C. Identical amino acids between cathepsin C and the other cysteine proteinases are shadowed. Amino acid numbers start from the NH₂-terminal end of mature cathepsin C. Identical amino acids between cathepsins C and H are numbered from the NH₂ termini. The region of cathepsin C is different from the region that is similar to the propeptide region of other cysteine proteinases. The function of the similar regions in cathepsins H and C is still unknown; further expression studies using deleted- and site-directed mutants of the cDNA for cathepsins C and H in Escherichia coli, such as have been done with human cathepsin L (37), will provide clues to the function of this region.

The mRNA levels for cathepsin C are high in liver, spleen, small and large intestine, lung, and kidney, moderate in esophagus, stomach, and heart muscle, and low in the submandibular gland, aorta, and skeletal muscle. Only small amounts of mRNA for cathepsin C exist in brain, pancreas, adrenal gland, and testis. These results suggest that the enzyme plays an important role in the alimentary tract.

Discussion

In the mature enzyme region, the amino acid sequence of cathepsin C shares 39.5, 35.1, 30.1, and 33.3% identity with those of cathepsins H, L, B, and papain, respectively (Fig. 4A). Especially, around the active site cysteine, histidine, and asparagine residues, the amino acid sequence of cathepsin C is very similar to those of cathepsins H, L, B, and papain. Thus, cathepsin C belongs to the papain family. The tyrosine residue next to the active site cysteine in cathepsin C is a substitution for the tryptophan residue that is well conserved among other cysteine proteinases as part of the hydrophobic backbone (36). This tyrosine substitution, therefore, may affect substrate specificity.

Some amino acid sequence similarities have been reported for the propeptide regions of cysteine proteinases (7). The papain propeptide region, however, shows no sequence similarity to any other cysteine proteinase except cathepsin C. The region of cathepsin H that is similar to cathepsin C is different from the region that is similar to the propeptide regions of other cysteine proteinases. The function of the similar regions in cathepsins H and C is still unknown; further expression studies using deleted- and site-directed mutants of the cDNA for cathepsins C and H in Escherichia coli, such as have been done with human cathepsin L (37), will provide clues to the function of this region.

The papain family includes many proteinases: papain, cathepsins B, H, L, and calpain (39), and the calpain superfamily (40). The papain superfamily can be divided into two families, the papain family and the calpain family. The papain family includes many proteinases: papain, cathepsins B, H, L, and calpain (38), and the superfamily that includes viral cysteine proteinases that resemble trypsin-like serine proteinases rather than papain-like cysteine proteinase, even though they have a cysteine residue at the active site (40). The papain superfamily can be divided into two families, the papain family and the calpain family. The papain family includes many proteinases: papain, cathepsins B, H, L, and calpain (41), and potato cysteine proteinase from Schistosoma mansoni (41), T. brucei (12), tomato (13), and rice seed (42). Although cathepsin C differs in its oligomeric properties (10), slow inhibition by E-64 (18), and an extremely long propeptide in its primary structure, it is apparently a member of the papain family as discussed above.

This paper reports for the first time the structure of procathepsin C. Although cathepsin C has characteristics different from those of other cysteine proteinases, it is still a member of the papain family. The function of the long propeptide of cathepsin C is still unknown, but may provide a

3 D. Muno and E. Kominami, manuscript in preparation.
clue to understanding the function of the cysteine proteinase.

Acknowledgments—We thank Prof. Koichi Suzuki and Drs. Yoh Okamoto and Takashi Ueno for helpful discussions; Dr. Hirohisa Sorimachi for the amino acid sequence homology search in the NBRF database; the Division of Biochemical Analysis, Central Laboratory of Medical Science for determining the amino acid sequences of the peptides and synthesizing the oligonucleotide probes; and Toa Nenryo Kagyo K. K. (Tokyo, Japan) for the gift of the hydroxyapatite (fluoroapatite) HPLC column.

REFERENCES

1. Katunuma, N., and Kominami, E. (1989) Intracellular Proteolysis: Mechanisms and Regulations (Katunuma, N., and Kominami, E., eds.) pp. 3-23, Japanese Scientific Press, Tokyo, Japan.

2. Guagliardi, L. E., Koppelman, B., Blum, J. S., Marks, M. S., Cresswell, P., and Brodey, F. M. (1990) Science 234, 133-139.

3. Sloane, B. F., and Honn, K. V. (1984) Cancer Metastasis Rev. 3, 249-263.

4. Ivy, G. O., Schottler, F., Wenzel, J., Baudry, M., and Lynch, G. (1984) Science 226, 865-887.

5. Chan, S. J., San Segundo, B., McCormick, M. B., and Steiner, D. F. (1986) Proc. Natl. Acad. Sci. U. S. A. 83, 7721-7725.

6. Ishidoh, K., Majumdar, S., Ohno, S., Kawasaki, H., Minami, Y., Kominami, E., Katunuma, N., and Suzuki, K. (1987) FEBS Lett. 226, 35-37.

7. Ishidoh, K., Towatari, T., Majumdar, S., Kawasaki, H., Minami, Y., Kominami, E., Katunuma, N., and Suzuki, K. (1987) FEBS Lett. 226, 69-73.

8. Tokito, K., Towatari, T., Katunuma, N., Teller, D. C., and Titani, K. (1985) Proc. Natl. Acad. U. S. A. 80, 3666-3670.

9. McDonald, J. K., Zeitman, B. B., Reilly, T. J., and Ellis, S. (1969) J. Biol. Chem. 244, 2693-2709.

10. Liao, J. C., and Lenney, J. F. (1984) Biochem. Biophys. Res. Commun. 124, 909-916.

11. Suzuki, M., DeLeon, J. C., and Lazo, J. S. (1987) Biochemistry 26, 4215-4219.

12. Mottram, J. C., North, M. J., Barry, J. D., and Coombs, G. H. (1989) FEBS Lett. 258, 211-215.

13. Schaffer, M. A., and Fisher, R. L. (1988) Plant Physiol. 87, 431-439.

14. Doughty, M. J., and Gruneisen, E. I. (1988) Thromb. Haemostasis 59, 373-377.

15. Lynh, G. W., and Pfueffer, S. L. (1988) Thromb. Haemostasis 59, 373-377.

16. D'Agrosa, R. M., and Calhoun, J. W. (1988) Biochem. Biophys. Res. Commun. 157, 770-775.

17. McDonald, J. K., Ellis, S., and Reilly, T. J. (1966) J. Biol. Chem. 241, 1494-1501.

18. Barret, A. J., Kembhavi, A. A., Brown, M. A., Kirschke, H., Wright, C. G., Tatsumi, M., and Hanada, K. (1985) Biochem. J. 21, 280-293.

19. Barret, A. J. (1972) Anal. Biochem. 47, 280-293.

20. Laemmli, U. K. (1970) Nature 227, 680-686.

21. Crestfield, A. M., Moore, S., and Stein, W. H. (1963) J. Biol. Chem. 238, 227.

22. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., and Rutter, W. J. (1979) Biochemistry 18, 5294-5299.

23. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, pp. 97-402, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

24. Gubler, U., and Hoffman, B. J. (1983) Gene (Amst.) 26, 263-269.

25. Lathe, R. (1985) J. Mol. Biol. 183, 1-12.

26. Hanahan, D. (1983) J. Mol. Biol. 166, 557-580.

27. Hattori, M., and Sakaki, Y. (1986) Anal. Biochem. 152, 232-238.

28. Chomczynski, P., and Sacchi, N. (1987) Anal. Biochem. 162, 156-159.

29. Kyte, J., and Doolittle, R. F. (1982) J. Mol. Biol. 157, 109-132.

30. Kanehisa, M. (1986) Nucleic Acids Res. 10, 183-196.

31. Kozak, M. (1987) Nucleic Acids Res. 15, 8125-8148.

32. Proudnfoot, N. J., and Brownlee, G. G. (1976) Nature 263, 211-214.

33. von Heijne, G. (1985) J. Mol. Biol. 184, 99-105.

34. von Figura, K., and Hasilik, A. (1986) Annu. Rev. Biochem. 55, 167-193.

35. Kornfeld, S. (1986) J. Clin. Invest. 77, 1-6.

36. Dufour, E. (1986) Biochimie (Paris) 70, 1353-1342.

37. Smith, S. M., and Gottesman, M. M. (1989) J. Biol. Chem. 264, 20487-20492.

38. Ritonja, A., Popovic, T., Kotnik, M., Machleidt, W., and Turk, V. (1988) FEBS Lett. 228, 341-345.

39. Ohno, S., Emori, Y., Majumdar, S., Kawasaki, H., Kirisama, M., and Suzuki, K. (1984) Nature 312, 556-570.

40. Bazan, J. F., and Fletterick, R. J. (1988) Proc. Natl. Acad. Sci. U. S. A. 85, 7872-7876.

41. Klinkert, M. Q., Felleisen, R., Link, G., Ruppel, A., and Beck, E. (1989) Mol. Biochem. Parasitol. 33, 113-122.

42. Watanabe, H., Abe, K., Emori, Y., Hosoyama, H., and Arai, S. (1991) J. Biol. Chem. 266, 18971-18992.

43. Cohen, L. W., Coghlan, V. M., and Dihel, L. C. (1986) Gene (Amst) 48, 219-227.
peptides were isolated by the same HPLC system. The 5-carboxyfluorescein-labeled cDNA C (5'G) was also digested with 200 ng of proteinase K from Staphylococcus aureus, 1 nMol of 5'G probe and digested cDNA C and the digestions were resolved by 5% non-denaturing polyacrylamide gel electrophoresis. cDNA C, the cDNA was resuspended in 10 mM Tris-HCl (pH 8.0) for 1 hour at room temperature. This resuspended cDNA was subjected to reverse phase HPLC [Fig. 6; peptide 15-11] with a linear gradient of 0 to 80% acetonitrile in 0.1% TFA.

The amino acid sequences of the separated peptides, and heavy and light chain sequences were determined using Applied Biosystems 477A peptide sequencer.

The 16316 molecular cloning of cDNA for Rat Cathepsin C

**Figure 1:** HPLC-PAGE of purified cathepsin C.

Purified cathepsin C was loaded onto a 15% SDS-PAGE gel containing 2-mercaptoethanol. After electrophoresis, the gel was stained with Coomassie Brilliant Blue C-250. The number on the left shows the molecular weight standards. 20,000 (phosphorylase b), 116,000 (bovine serum albumin), 66,000 (carbonic anhydrase), 49,000 (insulin), 30,000 (trypsin inhibitor).

**Figure 2:** Isolation of cathepsin C digested with trypsin endopeptidase. Purified cathepsin C was digested with trypsin endopeptidase, and the resultant peptides were separated on a reverse phase column (TSK805) with a linear gradient of acetonitrile to 80% (0-15% TFA). Peptide A (1) eluted at 33.50%. Peptide B, which consisted of undigested or aggregated peptides, eluted at 40.50%. Peptide A was subjected to protein sequencing. Fraction A was further digested with trypsin.

**Figure 3:** Separation of tryptic peptides of cathepsin C. Peptides were separated on a reverse phase column (TSK805) by HPLC. Peptides were eluted at 33.5% and 40.5% acetonitrile, respectively. Peptides A and B were further subjected to protein sequencing.

**Figure 4:** Separation of tryptic peptides of cathepsin C. Peptides were separated on reverse phase column (TSK805) by HPLC. Peptides were eluted at 33.5% and 40.5% acetonitrile, respectively.
Table 1a. Summary of calpain C purification.

| Steps | Total protein (mg) | Total activity (U/mg) | Specific activity (U/mg) | Yield (%) | Purification (fold) |
|-------|-------------------|-----------------------|--------------------------|-----------|---------------------|
| [1]   | 20229             | 4763                  | 607                      | 100       | 1                   |
| [2]   | 10910             | 1054                  | 964                      | 122       | 11                  |
| [3]   | 1932              | 4781                  | 246                      | 65        | 2.2                 |
| [4]   | 228.5             | 1465                  | 6.47                     | 16        | 24                  |
| [5]   | 88.9              | 932                   | 12.4                     | 8.3       | 46                  |
| [6]   | 0.7              | 407                   | 185.2                    | 7.0       | 686                 |
| [7]   | 0.8              | 426                   | 426                      | 4.3       | 1060                |

Table 2a. Amino acid sequences of peptides from calpain C.

Peptide IA: LQGAGTAYGAKILYQDK
Peptide ID: LQGAGTAYGAKILYQDK
Peptide IB: LQGAGTAYGAKILYQDK
Peptide IC: LQGAGTAYGAKILYQDK
Peptide ID: LQGAGTAYGAKILYQDK
Peptide IC: LQGAGTAYGAKILYQDK
Peptide ID: LQGAGTAYGAKILYQDK
Peptide IC: LQGAGTAYGAKILYQDK
Peptide ID: LQGAGTAYGAKILYQDK
Peptide IC: LQGAGTAYGAKILYQDK
Peptide ID: LQGAGTAYGAKILYQDK
Peptide IC: LQGAGTAYGAKILYQDK

Table 2b. Nucleotide sequences of oligonucleotide probes.

| Probe | Nucleotide sequence | Mixtures |
|-------|---------------------|----------|
| [C-1] | GTGCAACTAGTTAGCA    | 204 mixes|
| [C-2] | GTGCAACTAGTTAGCA    | 204 mixes|
| [C-3] | GCTGCAACTAGTTAGCA   | 204 mixes|
| [C-4] | GCTGCAACTAGTTAGCA   | 204 mixes|
| [C-5] | GCTGCAACTAGTTAGCA   | 204 mixes|
| [C-6] | GCTGCAACTAGTTAGCA   | 204 mixes|