Nucleotide and predicted amino acid sequences of cloned human and mouse procathepsin B cDNAs

(cysteine proteinases/cathepsin B gene/precursor processing/lysosomal sorting)

SHU JIN CHAN, BLANCA SAN SEGUNDO*, MARY BETH MCCORMICK, AND DONALD F. STEINER

The Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biology, The University of Chicago, 920 East 58th Street, Chicago, IL 60637

Contributed by Donald F. Steiner, July 14, 1986

ABSTRACT  Cathepsin B is a lysosomal thiol proteinase that may have additional extralysosomal functions. To further our investigations on the structure, mode of biosynthesis, and intracellular sorting of this enzyme, we have determined the complete coding sequences for human and mouse procathepsin B by using cDNA clones isolated from human hepatoma and kidney phage libraries. The nucleotide sequences predict that the primary structure of procathepsin B contains 339 amino acids organized as follows: a 17-residue NH2-terminal prepeptide sequence followed by a 62-residue propeptide region, 254 residues in mature (single chain) cathepsin B, and a 6-residue extension at the COOH terminus. A comparison of procathepsin B sequences from three species (human, mouse, and rat) reveals that the homology between the propeptides is relatively conserved with a minimum of 68% sequence identity. In particular, two conserved sequences in the propeptide that may be functionally significant include a potential glycosylation site and the presence of a single cysteine at position 59. Comparative analysis of the three sequences also suggests that processing of procathepsin B is a multistep process, during which enzymatically active intermediate forms may be generated. The availability of the cDNA clones will facilitate the identification of possible active or inactive intermediate processing forms as well as studies on the transcriptional regulation of the cathepsin B gene.

Cathepsin B is a member of a superfamily of structurally similar tissue proteinases having a catalytic unit of ~25 kDa that contains an active center made up of side chains derived from a cysteine residue located in its NH2-terminal region and a more COOH-terminally located histidine residue (1). These thiol proteinases are structurally and functionally closely related to papain as well as to aminidase, both plant enzymes (2). Cathepsin B also shows significant amino acid sequence homology to the proteolytic domain of the cytosolic calcium-dependent proteases (3), indicating that further evolutionary diversification has occurred within this proteolytic superfamily. Mature cathepsin B and the related thiol cathepsins H and L as well as a number of other exo- or endoproteinases have been localized to the lysosomes in various cells, indicating that these proteinases are involved in protein turnover (4).

Recent biosynthetic studies in our laboratory have indicated that cathepsin B is derived in biosynthesis from a larger precursor form, or procathepsin B, which in its glycosylated state has a molecular size of ~40 kDa (5). In isolated islets of Langerhans, this precursor form is either secreted into the medium or is slowly converted to material similar in size and immunological properties to mature cathepsin B and localized in lysosomal and secretion granule fractions; e.g., Docherty et al. (6) have identified both 31-kDa and 38-kDa cathepsin B-like proteins in purified secretion granules from a rat insulinoma and in normal rat islet granule fractions (7). A functional role for (pro)cathepsin B in the secretory vesicles has not been established, but it may be involved in prohormone conversion or, alternatively, in peptide hormone degradation (8). In addition, the secretion of higher molecular weight latent or active forms of cathepsin B, or closely related enzymes, has also been observed from a wide variety of tumors in vivo and in vitro (9–11), and this has led to speculations that overproduction of the enzyme may play some role in the transformed phenotype of some malignant tumors (12, 13).

To investigate these manifold questions surrounding the biosynthesis, intracellular targeting, secretion, and possible extralysosomal functions of cathepsin B, we have cloned cDNAs encoding the precursor from several species. In this paper, we report the structures of cDNAs encoding nearly full-length mRNAs for both the human and mouse precursors. We show that these contain a propeptide, or signal peptide, region for segregation of the precursor into the lumen of the rough endoplasmic reticulum as well as a lengthy presequence on the NH2-terminal side of the catalytic domain, which exhibits several interesting features.

MATERIALS AND METHODS

Materials. Restriction endonucleases, T4 polynucleotide kinase, T4 DNA ligase, and Escherichia coli DNA polymerase (Klenow) were obtained from New England Biolabs or Boehringer Mannheim. Plasmid vector pGEM2 DNA was purchased from Promega Biotec (Madison, WI). Nitrocellulose filter circles were obtained from Schleicher & Schuell. Radioactive nucleotides were purchased from Amersham.

Isolation of cDNA Clones. A human hepatoma cDNA library, cloned into λgt11, was obtained from J. DeWet (San Diego, CA) and has been described (14). A λgt10 human kidney cDNA library was obtained from G. Bell (Chiron, Emeryville, CA). The libraries were grown in 150-mm media plates at a density of 40,000 plaques per plate in E. coli strains Y1088 or BNN102 (15). Duplicate nitrocellulose filter lifts were prepared, hybridized with a nick-translated 950-base-pair (bp) EcoRI rat cathepsin B cDNA fragment isolated from λcB3 (16), and washed under reduced-stringency conditions (17). After autoradiography, selected positive clones were plaque-purified and phage DNA was isolated, digested with EcoRI to release the cloned cDNA fragment, and subcloned into plasmid vector pGEM2 for further analysis.

Sequence Analysis. DNA sequences were determined by the chemical degradation procedure of Maxam and Gilbert (18) and the dideoxynucleotide chain-termination method of Sanger (19) after subcloning into M13. Specific primers for dideoxynucleotide sequencing were synthesized on an Ap-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: bp, base pair(s).

*Present address: Department of Genética Molecular, Centro de Investigación y Desarrollo de Barcelona, Barcelona, Spain.
and length the mouse showed clones positive plaques within sequence from J. to reactive weakly thus digested To isolate human described. Approximately thesizer probe hybridization clones sentative into separated Biosystems (Foster City, CA) model 380A DNA synthesizer and purified by polyacrylamide gel electrophoresis in 7 M urea (20).  

RESULTS

To isolate human preprocathepsin B cDNA clones, we screened a λgt11 human hepatoma cDNA library obtained from J. DeWet, using cloned rat cathepsin B cDNA as the hybridization probe and reduced-stringency conditions as described. Approximately 60 positive signals were observed from 900,000 plaques in the initial screening. Although the strength of the autoradiographic signals was variable, the positive plaques on duplicate filters could be consistently separated into two classes: strongly reactive clones and more weakly reactive ones. We plaque-purified three representative clones from each class, extracted phage DNA, digested with EcoRI, and subcloned the cDNA inserts into the plasmid form for further analysis.

Restriction endonuclease mapping revealed that the cDNA inserts from the strongly reactive clones were overlapping, and thus originated from a single mRNA species. In addition, DNA sequence analysis revealed that the clones were 93% homologous to the rat preprocathepsin B cDNA sequence within the coding region, and these clones, designated λmCB14, λmCB54, and λmCB58 were identified as encoding mouse preprocathepsin B mRNA (Fig. 1).

Similar restriction mapping of the three weakly reactive clones showed that the cDNA inserts from these also overlapped and were derived from a second distinct mRNA species. DNA sequence analysis revealed an extended open reading frame in which the deduced amino acid sequence was in agreement with the published sequence for mature human cathepsin B (21). Based on these results, clones λhCB3, λhCB4, and λhCB8 were identified as encoding human preprocathepsin B mRNA (Fig. 2).

Because we are interested in comparing the expression of tumor form(s) of preprocathepsin B with preprocathepsin B in normal human tissues, we also screened λgt10 normal human kidney cDNA library with rat cathepsin B cDNA. One clone (λhCB79) containing a 2000-bp insert was isolated, restriction-mapped, and sequenced by using the strategy illustrated in Fig. 2.

The nucleotide sequence and deduced amino acid sequence for human preprocathepsin B cDNA is shown in Fig. 3. In comparing the kidney and hepatoma clones, no sequence differences were detected in the 5' overlapping or 3' untranslated regions, and a single nucleotide change was found in the coding region. The change, a G to D transition at position 120, however, resulted in a silent substitution in the codon for arginine and may reflect an allelic variation or a cloning artifact. In a Southern blot of human genomic DNA digested with several restriction enzymes, hybridization with labeled phCB79 revealed a simple fragmentation pattern consistent with the presence of a single copy gene (data not shown). These results also are essentially in agreement with the recently reported partial sequence of a human cathepsin B cDNA clone (30). We conclude that both human tumor and

![Fig. 1. Restriction map and sequencing strategy for mouse preprocathepsin B cDNA. The map was constructed from overlapping clones λmCB14, λmCB54, and λmCB58 as shown. Arrows indicate 5' to 3' direction and length of each sequenced fragment. UT, untranslated.](image)

![Fig. 2. Restriction map and sequencing strategy for human preprocathepsin B cDNA. The map was constructed from hepatoma cDNA clones λhCB3, λhCB4, and λhCB8, and from human kidney cDNA clone λhCB79, spanning the regions indicated. Arrows indicate 5' to 3' direction and length of each sequenced fragment. UT, untranslated.](image)
FIG. 3. Nucleotide and predicted amino acid sequences of human and mouse preprocathepsin B cDNAs. The composite human and mouse sequences were constructed from sequence data from the overlapping clones shown in Figs. 1 and 2. Nucleotides and predicted amino acid residues in rat preprocathepsin B cDNA that differ from the mouse sequence are shown below it. The complete human 3' untranslated region and a portion of the mouse and rat sequences are given. Arrows indicate potential cleavage sites for posttranslational processing.
normal tissue preprocathepsin B mRNAs are transcribed from the single cathepsin B gene.

The predicted primary structure of human preprocathepsin B contains 339 amino acids, including a 17-residue predominantly hydrophobic sequence at the NH$_2$ terminus. Such signal sequences function to sequester the nascent protein within the endoplasmic reticulum and are usually rapidly removed after synthesis (22). We identified a potential cleavage site at alanine-17 based on data from other known prepeptide sequences, which indicate that cleavage often occurs after the sequence Ala-X-Ala (23). Following the prepeptide, the structure of human procathepsin B consists of a 62-residue NH$_2$-terminal propeptide extension connected to the 254-residue mature single chain form of mature cathepsin B and is terminated by a 6-residue COOH-terminal peptide. Mature cathepsin B has also been isolated in a two-chain form, and this form from human liver has been sequenced by Ritonja et al. (21). In comparison, the cDNA-derived sequence predicts that the two-chain form is generated by cleavage at two sites between residues 126 and 129, coupled with the loss of a dipeptide. Otherwise, the two sequences are in agreement except for an asparagine for aspartic acid substitution at residue 228. The cleavage sites required to generate mature cathepsin B from preprocathepsin B are indicated by arrows in Fig. 3.

Outside the coding sequence, human preprocathepsin B cDNA contains 391 nucleotides in the 3' untranslated region, including a canonical hexanucleotide polyadenylation signal, AAATAAA (24), located 16 bp upstream from a stretch of poly(A). We also sequenced 191 bp in the 5' untranslated region for a total of 1995 nucleotides (Fig. 3). Since an RNA blot of human liver total RNA hybridized with labeled preprocathepsin B cDNA revealed a single band of ~2300 nucleotides, we conclude that the 5' untranslated region contains ~400 nucleotides (data not shown).

The composite sequence of mouse preprocathepsin B cDNA derived from clones AbmCB24, AbmCB54, and AbmCB58 is also given in Fig. 3. Like human and rat preprocathepsin B (for which we have now obtained the complete coding sequence), the primary structure of mouse preprocathepsin B contains 339 residues. As noted earlier, the mouse and rat sequences are strongly homologous, with 90.3% sequence identity within the coding region. This conserved homology extends into the 5' untranslated regions, which were compared, and for ~90 bp in the 3' untranslated region, as shown in Fig. 3. Downstream from this segment, however, the two sequences abruptly diverge and the mouse preprocathepsin B cDNA contains a much longer 3' untranslated region, exceeding 1500 nucleotides. The evolutionary origin for the extended 3' untranslated region in mouse is unknown, but it may be due to mutational loss of a polyadenylation signal sequence or the insertion of an additional exon in the genomic sequence.

**DISCUSSION**

In this report, we present the complete coding sequences for human and mouse preprocathepsin B from cDNA clones isolated from human hepatoma and kidney phage libraries. The mouse preprocathepsin B cDNA clones were obtained from screening the hepatoma library after infiltration of this tissue with host mouse reticuloendothelial cells (14). The calculated molecular masses for human and mouse procathepsin B predicted from the coding sequences are 35.9 and 35.5 KDa, respectively, and with allowance for the addition of carbohydrate moieties, these are close to the molecular masses of the observed biosynthetic form in inlets and to the secreted forms from human tumor cells (9-11).

Together with the coding sequence for rat preprocathepsin B, which we have recently completed (ref. 16; B.S.S., S.J.C., and D.F.S., unpublished data), the availability of the mouse and human sequences provided an opportunity to compare the cathepsin B structural gene from three mammalian species and to search for conserved features that may be functionally important. The nucleotide and amino acid sequence homologies between different regions of human, mouse, and rat preprocathepsin B are summarized in Table 1. As shown, mature cathepsin B contained the highest percentage of sequence identity followed by the NH$_2$-terminal proregion and the prepeptide. A direct comparison of the primary structures in the latter two regions, however, reveals that many of the amino acid substitutions are conservative (Fig. 4). Thus, changes in the prepeptide chain retain its overall hydrophobic character. Similarly, in the propeptide region the majority of the substitutions involve residues with chemically analogous side chains.

One conserved residue in the propeptide that may be of interest is the cysteine at position 59. It is possible that this additional thiol amino acid plays a role in regulating the enzymatic activity of procathepsin B by forming a disulfide bond with the active-site cysteine-152, although the oxidation state of the cysteines in either pro- or mature cathepsin B has not yet been determined. Another conserved feature that may be functionally important is that all three procathepsin B sequences contain a potential second glycosylation site at residue 38 with the identical recognition sequence Asn-Thr-Thr. Mature cathepsin B contains a single glycosylation site at asparagine-289. Glycosylation with mannose 6-phosphate has been shown to be an important sorting signal for routing proteins into lysosomes, but the mechanisms involved in this process, including substrate specificity, have not been completely elucidated (25). In preliminary experiments, we have found that rat procathepsin B contains a larger carbohydrate moiety than that reported for the mature enzyme (26), and this may be due in part to glycosylation at both sites (D.F.S., unpublished results).

A comparison of the primary structures of human, rat, and mouse procathepsin B also provided clues on the possible processing pathway for this enzyme. In particular, the residue preceding the NH$_2$-terminal leucine in mature cathepsin B is different in all three sequences (Fig. 4). This suggests that the initial cleavage in procathepsin B may occur further upstream in the propeptide, followed by stepwise removal of the NH$_2$-terminal extension, possibly by an amino dipeptidase activity. Within this context, it is noteworthy that the second NH$_2$-terminal residue in cathepsin B is a conserved proline, which would not be a substrate for amino

<table>
<thead>
<tr>
<th>Table 1. Homology between different regions of human, rat, and mouse procathepsin B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Prepeptide</td>
</tr>
<tr>
<td>Human/rat</td>
</tr>
<tr>
<td>Human/mouse</td>
</tr>
<tr>
<td>Rat/mouse</td>
</tr>
<tr>
<td>Propeptide</td>
</tr>
<tr>
<td>Human/rat</td>
</tr>
<tr>
<td>Human/mouse</td>
</tr>
<tr>
<td>Rat/mouse</td>
</tr>
<tr>
<td>Cathepsin B</td>
</tr>
<tr>
<td>Human/rat</td>
</tr>
<tr>
<td>Human/mouse</td>
</tr>
<tr>
<td>Rat/mouse</td>
</tr>
<tr>
<td>COOH-terminal peptide</td>
</tr>
<tr>
<td>Human/rat</td>
</tr>
<tr>
<td>Human/mouse</td>
</tr>
<tr>
<td>Rat/mouse</td>
</tr>
</tbody>
</table>

Cathepsin B sequences compared are the single-chain forms.
dipeptidase. Further processing steps include removal of the COOH-terminal hexapeptide and cleavage at residues 126 and 129 (coupled with the loss of a dipeptide) to generate the two-chain form. The latter cleavages probably occur within the lysosome, since a mixture of single- and two-chain cathepsin B has been isolated from this organelle and both forms are enzymatically active (1).

In the foregoing scheme, the processing of preprocathepsin B is postulated to be a multistep process during which intermediates may be formed that possess biological activity. The availability of the cDNA clones should facilitate the identification and isolation of such intermediates. For example, studies are in progress on the generation of antibodies to synthetic peptides corresponding to different segments of procathepsin B to be used to immunoprecipitate and characterize biosynthetic intermediates. The cDNA clones will also be used to test the function of specific residues via the techniques of in vitro mutagenesis and subsequently assaying the mutated genes for activity by transfection into cells. In particular, it would be of interest to introduce substitutions into cysteine-59 as well as the two potential glycosylation sites.

We have used these preprocathepsin B cDNA clones as hybridization probes to investigate the distribution of cathepsin B mRNA in various tissues (27). In addition, these cDNA clones can also be used to investigate the transcriptional regulation of this gene in normal and tumor cells. Although increased cathepsin B-like activity has been consistently reported in metastatic tumors, the molecular identity of this activity has not been fully elucidated. A major caveat in interpreting these studies is that the tumors are often infiltrated with macrophages that contain substantial amounts of cathepsin B (28). However, by using the cathepsin B cDNA for hybridization in situ, it should be possible to assay the expression of this gene directly in individual cells (29).

We would like to thank Cathy Christopherson for her assistance in the preparation of this manuscript. This work was supported in part by National Institutes of Health Grants AM 13914 and AM 20595.

Fig. 4. Homology between signal peptides and propeptide regions of human, rat, and mouse preprocathepsin B. Amino acid residues that are identical in all three sequences are boxed; asterisk indicates potential glycosylation site; arrows indicate possible cleavage sites.