Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 7721–7725, October 1986
Biochemistry

Nucleotide and predicted amino acid sequences of cloned human and mouse preprocathepsin 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 preprocathepsin B by using cDNA clones isolated from human hepatoma and kidney phage libraries. The nucleotide sequences predict that the primary structure of preprocathepsin B contains 339 amino acids organized as follows: a 17-residue NH_2-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 processive 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 NH_2-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 actinidin, 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 prepeptide, or signal peptide, region for segregation of the precursor into the lumen of the rough endoplasmic reticulum as well as a lengthy prosegment on the NH_2-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) *Eco*RI rat cathepsin B cDNA fragment isolated from λrcB3 (16), and washed under reduced-stringency conditions (17). After autoradiography, selected positive clones were plaque-purified and phage DNA was isolated, digested with *Eco*RI 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-

Abbreviation: bp, base pair(s).

*Present address: Department of Genetica Molecular, Centro de Investigation y Desarrollo de Barcelona, Barcelona, Spain.

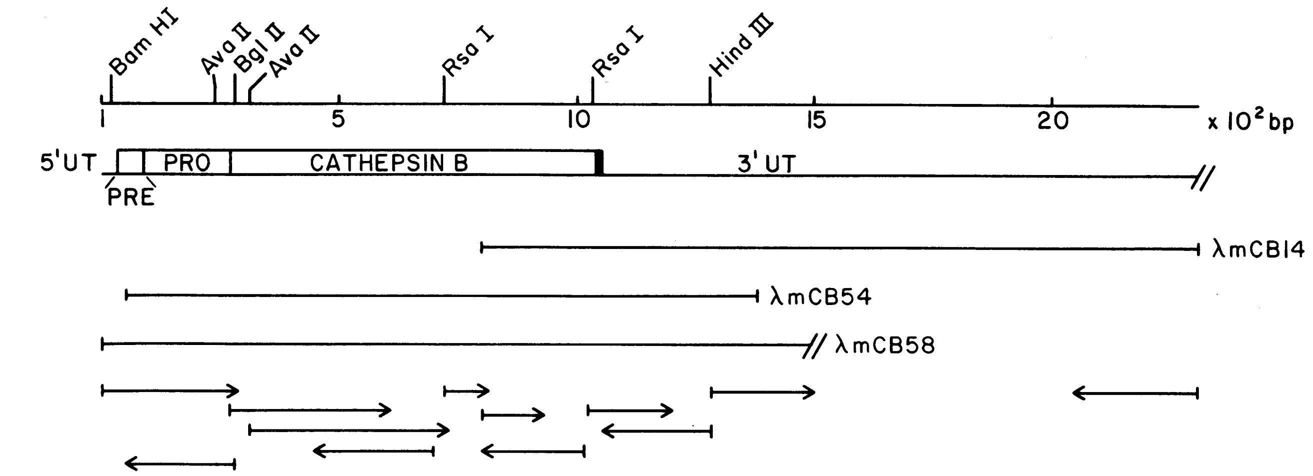


FIG. 1. Restriction map and sequencing strategy for mouse preprocathepsin B cDNA. The map was constructed from overlapping clones λmCB14, λmCB54, and λmCB58 as shown. Arrows indicate 5' to 3' direction and length of each sequenced fragment. UT, untranslated.

plied 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 *Eco*RI, 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 dC to dG 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

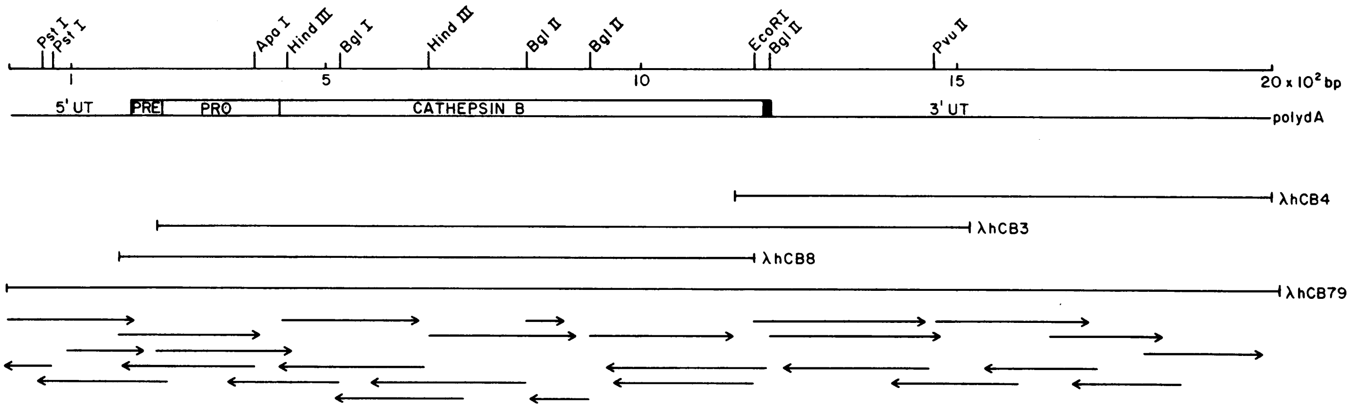


FIG. 2. Restriction map and sequencing strategy for human preprocathepsin B cDNA. The map was constructed from hepatoma cDNA clones λhCB3, λhCB4, and λhCB8, and from human kidney cDNA clone λhCB79, spanning the regions indicated. Arrows indicate 5' to 3' direction and length of each sequenced fragment. UT, untranslated.

```
Human                                                AATTCCGCGGCAACCGCTCCGGCAACGCCAACCGCTCCGCTGCGCGCAGGCTGGGCTGCAGGCTCTCGGCTGCAG -120

Human    CGCTGGGCTGGTGTGCAGTGGTGCGACCACGGCTCACGGCAGCCTCAGCCACCCAGATGTAAGCGATCTGGTTCCCACCTCAGCCTTCCGAGTAGTGGATCTAGGATCTGGCTTCCAAC -1
Mouse                                                                                         CTGCGCTCGGTGAGTGCAGGATCCAGCATCCAGG
Rat                                                                                                                      T

         1                                                                  ↓        20
         Met Trp Gln Leu Trp Ala Ser Leu Cys Cys Leu Leu Val Leu Ala Asn Ala Arg Ser Arg Pro Ser Phe His Pro Val Ser Asp Glu Leu
Human    ATG TGG CAG CTC TGG GCC TCC CTC TGC TGC CTG CTG GTG TTG GCC AAT GCC CGG AGC AGG CCC TCT TTC CAT CCC GTG TCG GAT GAG CTG  90
Mouse    ATG TGG TGG TCC TTG ATC CTT CTT TCT TGC CTG CTG GCA CTG ACC AGT GCC CAT GAC AAG CCT TCC TTC CAC CCG CTG TCG GAT GAC CTG
Rat                                       C    C                                                           T       A           C     A
         Met Trp Trp Ser Leu Ile Leu Leu Ser Cys Leu Leu Ala Leu Thr Ser Ala His Asp Lys Pro Ser Phe His Pro Leu Ser Asp Asp Leu
                                 Pro                                                                                         Met

                                             40                                                                              60
         Val Asn Tyr Val Asn Lys Arg Asn Thr Thr Trp Gln Ala Gly His Asn Phe Tyr Asn Val Asp Met Ser Tyr Leu Lys Arg Leu Cys Gly
Human    GTC AAC TAT GTC AAC AAA CGG AAT ACC ACG TGG CAG GCC GGG CAC AAC TTC TAC AAC GTG GAC ATG AGC TAC TTG AAG AGG CTA TGT GGT  180
Mouse    ATT AAC TAT ATC AAC AAA CAG AAT ACA ACA TGG CAG GCT GGA CGC AAC TTC TAC AAT GTT GAC ATA AGC TAT CTG AAG AAG CTG TGT GGC
Rat                                                                                                                       C        A
         Ile Asn Tyr Ile Asn Lys Gln Asn Thr Thr Trp Gln Ala Gly Arg Asn Phe Tyr Asn Val Asp Ile Ser Tyr Leu Lys Lys Leu Cys Gly
                                                                                                                     Pro

                                                                                    ↓80
         Thr Phe Leu Gly Gly Pro Lys Pro Pro Gln Arg Val Met Phe Thr Glu Asp Leu Lys Leu Pro Ala Ser Phe Asp Ala Arg Glu Gln Trp
Human    ACC TTC CTG GGT GGG CCC AAG CCA CCC CAG AGA GTT ATG TTT ACC GAG GAC CTG AAG CTG CCT GCA AGC TTC GAT GCA CGG GAA CAA TGG  270
Mouse    ACT GTC CTG GGT GGA CCC AAA CTG CCA GGA AGG GTT GCG TTC GGT GAG GAC ATA GAT CTA CCT GAA ACC TTT GAT GCA CGG GAA CAA TGG
Rat                                         G       G  A           G      A C             A               T                         G
         Thr Val Leu Gly Gly Pro Lys Leu Pro Gly Arg Val Ala Phe Gly Glu Asp Ile Asp Leu Pro Glu Thr Phe Asp Ala Arg Glu Gln Trp
                                             Glu         Gly     Ser             Asn             Ser

                                             100                                                                             120
         Pro Gln Cys Pro Thr Ile Lys Glu Ile Arg Asp Gln Gly Ser Cys Gly Ser Cys Trp Ala Phe Gly Ala Val Glu Ala Ile Ser Asp Arg
Human    CCA CAG TGT CCC ACC ATC AAA GAG ATC AGA GAC CAG GGC TCC TGT GGC TCC TGC TGG GCC TTC GGG GCT GTG GAA GCC ATC TCT GAC CGC  360
Mouse    TCC AAC TGC CCG ACC ATT GGA CAG ATT AGA GAC CAG GGC TCC TGC GGC TCT TGT TGG GCA TTT GGG GCA GTG GAA GCC ATT TCT GAC CGA
Rat            T               C  C        C               G       T                                               G
         Ser Asn Cys Pro Thr Ile Gly Gln Ile Arg Asp Gln Gly Ser Cys Gly Ser Cys Trp Ala Phe Gly Ala Val Glu Ala Ile Ser Asp Arg
                                 Ala                                                                             Met

                                ↓       ↓                                            140
         Ile Cys Ile His Thr Asn Ala His Val Ser Val Glu Val Ser Ala Glu Asp Leu Leu Thr Cys Cys Gly Ser Met Cys Gly Asp Gly Cys
Human    ATC TGC ATC CAC ACC AAT GCG CAC GTC AGC GTG GAG GTG TCG GCG GAG GAC CTG CTC ACC TGC TGT GGC AGC ATG TGT GGG GAC GGC TGT  450
Mouse    ACC TGC ATT CAC ACC AAT GGC CGA GTC AAC GTG GAG GTG TCT GCT GAA GAC CTG CTT ACT TGC TGT GGC ATC CAG TGT GGG GAC GGC TGT
Rat       T                                    T                       G               C           T                   T
         Thr Cys Ile His Thr Asn Gly Arg Val Asn Val Glu Val Ser Ala Glu Asp Leu Leu Thr Cys Cys Gly Ile Gln Cys Gly Asp Gly Cys
         Ile

                                             160                                                                             180
         Asn Gly Gly Tyr Pro Ala Glu Ala Trp Asn Phe Trp Thr Arg Lys Gly Leu Val Ser Gly Gly Leu Tyr Glu Ser His Val Gly Cys Arg
Human    AAT GGT GGC TAT CCT GCT GAA GCT TGG AAC TTC TGG ACA AGA AAA GGC CTG GTT TCT GGT GGC CTC TAT GAA TCC CAT GTA GGG TGC AGA  540
Mouse    AAT GGT GGC TAT CCC TCT GGA GCA TGG AAC TTC TGG ACA AAA AAA GGC CTG GTT TCA GGT GGA GTC TAC GAT TCT CAT ATA GGC TGC TTA
Rat                                                                T  G          T         T           A     A     A
         Asn Gly Gly Tyr Pro Ser Gly Ala Trp Asn Phe Trp Thr Lys Lys Gly Leu Val Ser Gly Gly Val Tyr Asp Ser His Ile Gly Cys Leu
                                                             Arg                                     Asn

                                                                                     200
         Pro Tyr Ser Ile Pro Pro Cys Glu His His Val Asn Gly Ser Arg Pro Pro Cys Thr Gly Glu Gly Asp Thr Pro Lys Cys Ser Lys Ile
Human    CCG TAC TCC ATC CCT CCC TGT GAG CAC CAC GTC AAC GGC TCC CGG CCC CCA TGC ACG GGG GAG GGA GAT ACC CCC AAG TGT AGC AAG ATC  630
Mouse    CCG TAC ACC ATC CCT CCC TGC GAG CAC CAT GTC AAT GGC TCC CGT CCC CCG TGT ACT GGG GAG GGG GAT ACT CCA AGG TGC AAC AAG AGC
Rat        C                       T   A   T                               A   C       A       A           C  A               TG
         Pro Tyr Thr Ile Pro Pro Cys Glu His His Val Asn Gly Ser Arg Pro Pro Cys Thr Gly Glu Gly Asp Thr Pro Arg Cys Asn Lys Ser
                                                                                                             Lys             Met

                                             220                                                                             240
         Cys Glu Pro Gly Tyr Ser Pro Thr Tyr Lys Gln Asp Lys His Tyr Gly Tyr Asn Ser Tyr Ser Val Ser Asn Ser Glu Lys Asp Ile Met
Human    TGT GAG CCT GGC TAC AGC CCG ACC TAC AAA CAG GAC AAG CAC TAC GGA TAC AAT TCC TAC AGC GTC TCC AAT AGC GAG AAG GAC ATC ATG  720
Mouse    TGT GAA GCT GGC TAC TCC CCA TCC TAC AAA GAG GAT AAG CAC TTT GGG TAC ACT TCC TAC AGC GTG TCT AAC AGT GTG AAG GAG ATC ATG
Rat            G                 A             G   A              A                        T         G     C  A
         Cys Glu Ala Gly Tyr Ser Pro Ser Tyr Lys Glu Asp Lys His Phe Gly Tyr Thr Ser Tyr Ser Val Ser Asn Ser Val Lys Glu Ile Met
                                 Thr                             Tyr                             Asp     Glu

                                                                                     260
         Ala Glu Ile Tyr Lys Asn Gly Pro Val Glu Gly Ala Phe Ser Val Tyr Ser Asp Phe Leu Leu Tyr Lys Ser Gly Val Tyr Gln His Val
Human    GCC GAG ATC TAC AAA AAC GGC CCC GTG GAG GGA GCT TTC TCT GTG TAT TCG GAC TTC CTG CTC TAC AAG TCA GGA GTG TAC CAA CAC GTC  810
Mouse    GCA GAA ATC TAC AAA AAT GGC CCA GTG GAG GGT GCC TTC ACT GTG TTT TCT GAC TTC TTG ACT TAC AAA TCA GGA GTA TAC AAG CAT GAA
Rat        G                                           T   T                                               C
         Ala Glu Ile Tyr Lys Asn Gly Pro Val Glu Gly Ala Phe Thr Val Phe Ser Asp Phe Leu Thr Tyr Lys Ser Gly Val Tyr Lys His Glu

                                             280                                                                             300
         Thr Gly Glu Met Met Gly Gly His Ala Ile Arg Ile Leu Gly Trp Gly Val Glu Asn Gly Thr Pro Tyr Trp Leu Val Ala Asn Ser Trp
Human    ACC GGA GAG ATG ATG GGT GGC CAT GCC ATC CGC ATC CTG GGC TGG GGA GTG GAG AAT GGC ACA CCC TAC TGG CTG GTT GCC AAC TCC TGG  900
Mouse    GCC GGT GAT ATG ATG GGT GGC CAC GCC ATC CGC ATC CTG GTC TGG GGA GTA GAG AAT GGA GTT CCC TAC TGG CTG GCA GCC AAC TCT TGG
Rat                  G         A       T               T      G          A                 A                  T    A       C
         Ala Gly Asp Met Met Gly Gly His Ala Ile Arg Ile Leu Val Trp Gly Val Glu Asn Gly Val Pro Tyr Trp Leu Ala Ala Asn Ser Trp
                     Val                                     Gly         Ile                                 Val

                                                                                     320
         Asn Thr Asp Trp Gly Asp Asn Gly Phe Phe Lys Ile Leu Arg Gly Gln Asp His Cys Gly Ile Glu Ser Glu Val Val Ala Gly Ile Pro
Human    AAC ACT GAC TGG GGT GAC AAT GGC TTC TTT AAA ATA CTC AGA GGA CAG GAT CAC TGC GGA ATC GAA TCA GAA GTG GTG GCT GGA ATT CCA  990
Mouse    AAC CTT GAC TGG GGT GAT AAT GGC TTC TTT AAA ATC CTC AGA GGA GAA AAC CAC TGT GGC ATT GAA TCA GAA ATT GTG GCT GGA ATC CCA
Rat          G                         T                               G               A                   C
         Asn Leu Asp Trp Gly Asp Asn Gly Phe Phe Lys Ile Leu Arg Gly Glu Asn His Cys Gly Ile Glu Ser Glu Ile Val Ala Gly Ile Pro
             Val

                    ↓                    339
         Arg Thr Asp Gln Tyr Trp Glu Lys Ile
Human    CGC ACC GAT CAG TAC TGG GAA AAG ATC TAATCTGCCGTGGGCCTGTCGTGCCAGTCCTGGGGGCGAGATCGGGGTAGAAAGTCATTTTATTCTTTAAGTTCACGTAAGAT 1100
Mouse    CGC ACT GAC CAG TAC TGG GGA AGA TTC TAATCTGCTTTGACTTCATTGTCCAGTCCTTAGGGGCTTTTTCCAAAATTTAGCGGCCTTGGCAGAGAATGAGGTAGACAGGG
Rat              C G                                  GG     G     AT          A AC                AATGCA     G  G     C T GGA T
         Arg Thr Asp Gln Tyr Trp Gly Arg Phe
                 Gln

Human    ACAAGTTTCAGGCAGGGTCTGAAGGACTGGATTGGCCAAAGTCCTCCAAGGAGACCAAGTCCTGGCTACATCCCAGCCTGTGGTTACAGTGCAGACAGGCCATGTGAGCCACCGCTGCC 1219
Mouse    GGATCTTTGATTC
Rat      CTT GA  CT

Human    AGCACAGAGCGTCCTTCCCCCTGTAGACTAGTGCCGTGGGAGTACCTGCTGCCCAGCTGCTGTGGCCCCCTCCGTGATCCATCCATCTCCAGGGAGCAAGACAGAGACGCAGGATGGAA 1338

Human    AGCGGAGTTCCTAACAGGATGAAAGTTCCCCCATCAGTTCCCCCAGTACCTCCAAGCAAGTAGCTTTCCACATTTGTCACAGAAATCAGAGGAGAGATGGTGTTGGGAGCCCTTTGGAG 1457

Human    AACGCCAGTCTCCAGGTCCCCCTGCATCTATCGAGTTTGCAATGTCACAACCTCTCTGATCTTGTGCTCAGCATGATTCTTTAATAGAAGTTTTATTTTTCGTGCACTCTGCTAATCAT 1576

Human    GTGGGTGAGCCAGTGGAACAGCGGGAGCCTGTGCTGGTTTGCAGATTGCCTCCTAATGACGCGGCTCAAAAGGAAACCAAGTGGTCAGGAGTTGTTTCTGACCCACTGATCTCTACTAC 1695

Human    CACAAGGAAAATAGTTTAGGAGAAACCAGCTTTTACTGTTTTTGAAAAATTACAGCTTCACCCTGTCAAGTTAACAAGGAATGCCTGTGCCAATAAAAGGTTTCTCCAACTTG-polyA 1814
```

FIG. 3. Nucleotide and predicted amino acid sequences of human and mouse preprocathepsin B cDNAs. The composite human and mouse sequences were constructed from sequenced DNA fragments from the overlapping clones as 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 791 nucleotides in the 3′ untranslated region, including a canonical hexanucleotide polyadenylylation signal, AATAAA (24), located 16 bp upstream from a stretch of poly(dA). 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 λmCB24, λmCB54, and λmCB58 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 polyadenylylation 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 islets and to the secreted form from 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.

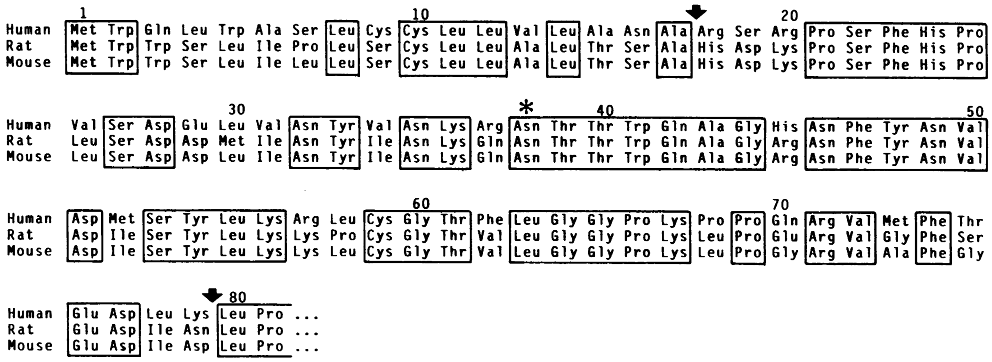
One conserved residue in the propeptide that may be of interest is the cysteine at position 59. It is possible that this additional thiol amino acid plays a role in regulating the enzymatic activity of procathepsin B by forming a disulfide bond with the active-site cysteine-152, although the oxidation 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 preprocathepsin 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 1. Homology between different regions of human, rat, and mouse preprocathepsin B

	Amino acid (% homology)		Nucleotide (% homology)	
Prepeptide				
Human/rat	8/17	(47%)	35/51	(69%)
Human/mouse	8/17	(47%)	33/51	(65%)
Rat/mouse	16/17	(94%)	49/51	(96%)
Proregion				
Human/rat	42/62	(68%)	138/186	(74%)
Human/mouse	44/62	(71%)	137/186	(74%)
Rat/mouse	56/62	(90%)	173/186	(93%)
Cathepsin B				
Human/rat	213/254	(84%)	629/762	(83%)
Human/mouse	210/254	(83%)	626/762	(82%)
Rat/mouse	237/254	(93%)	699/762	(92%)
COOH-terminal peptide				
Human/rat	3/6	(50%)	14/18	(78%)
Human/mouse	3/6	(50%)	14/18	(78%)
Rat/mouse	6/6	(100%)	18/18	(100%)

Cathepsin B sequences compared are the single-chain forms.



```
        1                                     10                                   ↓            20
Human  Met Trp Gln Leu Trp Ala Ser Leu Cys Cys Leu Leu Val Leu Ala Asn Ala Arg Ser Arg Pro Ser Phe His Pro
Rat    Met Trp Trp Ser Leu Ile Pro Leu Ser Cys Leu Leu Ala Leu Thr Ser Ala His Asp Lys Pro Ser Phe His Pro
Mouse  Met Trp Trp Ser Leu Ile Leu Leu Ser Cys Leu Leu Ala Leu Thr Ser Ala His Asp Lys Pro Ser Phe His Pro

                        30                          *   40                                              50
Human  Val Ser Asp Glu Leu Val Asn Tyr Val Asn Lys Arg Asn Thr Thr Trp Gln Ala Gly His Asn Phe Tyr Asn Val
Rat    Leu Ser Asp Asp Met Ile Asn Tyr Ile Asn Lys Gln Asn Thr Thr Trp Gln Ala Gly Arg Asn Phe Tyr Asn Val
Mouse  Leu Ser Asp Asp Leu Ile Asn Tyr Ile Asn Lys Gln Asn Thr Thr Trp Gln Ala Gly Arg Asn Phe Tyr Asn Val

                                            60                                      70
Human  Asp Met Ser Tyr Leu Lys Arg Leu Cys Gly Thr Phe Leu Gly Gly Pro Lys Pro Pro Gln Arg Val Met Phe Thr
Rat    Asp Ile Ser Tyr Leu Lys Lys Pro Cys Gly Thr Val Leu Gly Gly Pro Lys Leu Pro Glu Arg Val Gly Phe Ser
Mouse  Asp Ile Ser Tyr Leu Lys Lys Leu Cys Gly Thr Val Leu Gly Gly Pro Lys Leu Pro Gly Arg Val Ala Phe Gly

                     ↓ 80
Human  Glu Asp Leu Lys Leu Pro ...
Rat    Glu Asp Ile Asn Leu Pro ...
Mouse  Glu Asp Ile Asp Leu Pro ...
```

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 potential cleavage sites.

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.

1. Takio, K., Towatari, T., Katunuma, N., Teller, D. C. & Titani, K. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3666–3670.
2. Carne, A. & Moore, C. H. (1978) *Biochem. J.* **173**, 73–83.
3. Ohno, S., Emori, Y., Imajoh, S., Kwasaki, H., Kisaragi, M. & Suzuki, K. (1984) *Nature (London)* **312**, 566–570.
4. Barrett, A. J. & McDonald, J. K. (1980) *Mammalian Proteases* (Academic, New York), Vol. 1.
5. Steiner, D. F., Docherty, K. & Carroll, R. (1984) *J. Cell. Biochem.* **24**, 121–130.
6. Docherty, K., Hutton, J. C. & Steiner, D. F. (1984) *J. Biol. Chem.* **259**, 6041–6044.
7. Docherty, K., Carroll, R. & Steiner, D. F. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3245–3249.
8. Bienkowski, R. S. (1983) *Biochem. J.* **214**, 1–10.
9. Mort, J. S., Ledu, M. S. & Recklies, A. D. (1983) *Biochim. Biophys. Acta* **755**, 369–375.
10. Dufek, V., Matous, B., Kral, V. & Bures, L. (1984) *Neoplasma* **31**, 581–590.
11. Olstein, A. D. & Liener, I. E. (1983) *J. Biol. Chem.* **258**, 11049–11056.
12. Sloane, B. F., Dunn, J. R. & Honn, K. V. (1981) *Science* **212**, 1151–1153.
13. Poole, A. R., Tiltman, K. J., Recklies, A. D. & Stoker, T. A. M. (1978) *Nature (London)* **273**, 545–547.
14. DeWet, J. R., Fukushima, J., Dewji, N. N., Wilcox, E., O'Brien, J. S. & Helinski, D. R. (1984) *DNA* **3**, 437–447.
15. Huynh, T. V., Young, R. A. & Davis, R. W. (1985) in *DNA Cloning*, ed. Glover, D. M. (IRL, Arlington), Vol. I, pp. 49–78.
16. San Segundo, B., Chan, S. J. & Steiner, D. F. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 2320–2324.
17. Chan, S. J., Episkopou, V., Zeitlin, S., Karathanasis, S. K., MacKrell, A., Steiner, D. F. & Efstratiadis, A. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 5046–5050.
18. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
19. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. (1980) *J. Mol. Biol.* **143**, 161–178.
20. Sanchez-Pescador, R. & Urdea, M. S. (1984) *DNA* **3**, 339–343.
21. Ritonja, A., Popovic, T., Turk, V., Wiedemann, K. & Machleidt, W. (1981) *FEBS Lett.* **181**, 169–172.
22. Docherty, K. & Steiner, D. F. (1981) *Annu. Rev. Physiol.* **44**, 625–638.
23. Van Heijne, G. (1983) *Eur. J. Biochem.* **133**, 17–21.
24. Nevins, J. R. (1983) *Annu. Rev. Biochem.* **52**, 441–446.
25. Sly, W. S. & Fischer, H. D. (1982) *J. Cell. Biochem.* **18**, 67–85.
26. Takahashi, T., Dehdarani, A. H., Schmidt, P. G. & Tang, J. (1984) *J. Biol. Chem.* **259**, 9874–9882.
27. San Segundo, B., Chan, S. J. & Steiner, D. F. (1986) *FEBS Lett.* **201**, 251–256.
28. Graf, M., Baici, A. & Strauli, P. (1981) *Lab Invest.* **45**, 587–596.
29. Lawrence, J. B. & Singer, R. H. (1985) *Nucleic Acids Res.* **13**, 1777–1799.
30. Fong, D., Calhoun, D. H., Hsieh, W.-T., Lee, B. & Wells, R. D. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 2909–2913.