Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 4910–4914, August 1985
Biochemistry

Cloning and sequence analysis of cDNA for human cathepsin D

(lysosomal enzyme/aspartyl proteases)

PHYLLIS L. FAUST*, STUART KORNFELD*, AND JOHN M. CHIRGWIN†

*Departments of Medicine and Biological Chemistry and †Department of Anatomy and Neurobiology, Washington University School of Medicine, St. Louis, MO 63110

Contributed by Stuart Kornfeld, April 3, 1985

ABSTRACT **An 1110-base-pair cDNA clone for human cathepsin D was obtained by screening a λgt10 human hepatoma G2 cDNA library with a human renin exon 3 genomic fragment. Poly(A)$^+$ RNA blot analysis with this cathepsin D clone demonstrated a message length of about 2.2 kilobases. The partial clone was used to screen a size-selected human kidney cDNA library, from which two cathepsin D recombinant plasmids with inserts of about 2200 and 2150 base pairs were obtained. The nucleotide sequences of these clones and of the λgt10 clone were determined. The amino acid sequence predicted from the cDNA sequence shows that human cathepsin D consists of 412 amino acids with 20 and 44 amino acids in a pre- and a prosegment, respectively. The mature protein region shows 87% amino acid identity with porcine cathepsin D but differs in having nine additional amino acids. Two of these are at the COOH terminus; the other seven are positioned between the previously determined junction for the light and heavy chains of porcine cathepsin D. A high degree of sequence homology was observed between human cathepsin D and other aspartyl proteases, suggesting a conservation of three-dimensional structure in this family of proteins.**

Cathepsin D is a lysosomal endoprotease that is present in all mammalian cells (for review, see ref. 1). It is a member of the aspartyl protease family, among which are the well-studied secretory enzymes renin, pepsin, and chymosin (2). Cathepsin D is the only aspartyl protease known to be lysosomal rather than secretory. Recently, the complete amino acid sequence for the mature protein of porcine cathepsin D has been determined (3). Alignment of this sequence with that of renin showed an overall 48.9% homology (3). In addition, the region surrounding the first active-site aspartyl (residue 32 in the pepsin numbering convention) was even more highly conserved.

The entire human renin gene has now been cloned and its organization has been determined (4, 5). Of the 10 exons present in the gene, exon 3 codes for amino acids surrounding the first active-site aspartic acid residue. Comparison of the amino acids encoded by this exon with the corresponding region of porcine cathepsin D revealed that 76% of the residues were identical. Given the evolutionary relationship of the proteins (2), one would expect the nucleotide homology to be very similar to the protein homology. In fact, the nucleotide homology between renin and human cathepsin D in this region was determined to be 75%.

Since human renin is known to be expressed at high levels only in the juxtaglomerular cells of the kidney, whereas cathepsin D is ubiquitous, it seemed reasonable to clone cathepsin D by choosing a non-renal tissue source for construction of a cDNA library and screening with a renin probe under conditions of reduced hybridization stringency. We report here the isolation and characterization of a cDNA clone containing the complete protein-encoding region for human cathepsin D.

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.

MATERIALS AND METHODS

Materials. Restriction endonucleases, T4 polynucleotide kinase, and T4 ligase were from New England Biolabs; terminal deoxynucleotidyltransferase, from P-L Biochemicals; *Escherichia coli* DNA polymerase I and Klenow fragment, *E. coli* RNase H, and calf alkaline phosphatase, from Bethesda Research Laboratories; avian myeloblastosis virus reverse transcriptase, from Boehringer Mannheim; and [γ-^{32}P]ATP (5000 Ci/mmol; 1 Ci = 37 GBq) and [α-^{32}P]dATP and -dCTP (800 Ci/mmol), from New England Nuclear.

Library Construction. A conventionally constructed cDNA library in λgt10 (6), made from human hepatoma cell line G2 (HepG2) mRNA, was kindly provided by R. Moore (Monsanto) and D. B. Wilson (Dept. of Medicine, Washington University). From this library a partial cDNA clone for cathepsin D was obtained.

A human kidney cDNA library was screened for a full-length clone. Total RNA was isolated by the guanidinium thiocyanate method (7) from a portion of surgically removed, perfused kidney and twice selected by chromatography on an oligo(dT)-cellulose column (8). Poly(A)$^+$ RNA (29 μg) was reverse-transcribed under the conditions of Retzell *et al.* (9). The first-strand heteroduplex was converted to fully double-stranded cDNA by replacement synthesis (10). The double-stranded cDNA was then resolved by preparative electrophoresis in an agarose gel and cDNAs 1.8–2.5 kilobases (kb) long were cut out and isolated (11). These cDNAs were then oligo(dC)-tailed (12) and annealed with plasmid pUC19 (13) that had been oligo(dG)-tailed at the *Sac* I site. *E. coli* K-12 strain DH-1 (14) was transformed and filter replicas of ampicillin-resistant colonies were prepared by the method of Hanahan and Meselson (15).

Library Screening. Phage-plaque replicas from the HepG2 λgt10 library were prepared by the method of Woo (16) and screened with a human renin exon 3 genomic subclone, pH10-BE0.6 (4). Nitrocellulose filter (Schleicher & Schuell) replicas were prehybridized in 2× NaCl/Cit/0.1% $NaDodSO_4$/5× Denhardt's solution/0.01% sonicated salmon sperm DNA at 50°C overnight. (NaCl/Cit is 0.15 M NaCl/15 mM sodium citrate; Denhardt's solution is 0.02% polyvinylpyrrolidone/0.02% Ficoll/0.02% bovine serum albumin.) Hybridization was at 50°C for 24 hr in 2× NaCl/Cit/0.1% $NaDodSO_4$/1× Denhardt's solution/0.01% sonicated salmon sperm DNA/10% (wt/vol) dextran sulfate with the human renin probe [labeled to $>10^8$ cpm/μg of DNA by nick-translation (17) in the presence of [α-^{32}P]dATP and -dCTP]. Filters were washed at 45°C in 0.1× NaCl/Cit/0.1% $NaDodSO_4$ prior to autoradiography. Positive phage were plaque-purified (11) at least three times.

Abbreviations: kb, kilobase(s); bp, base pair(s).

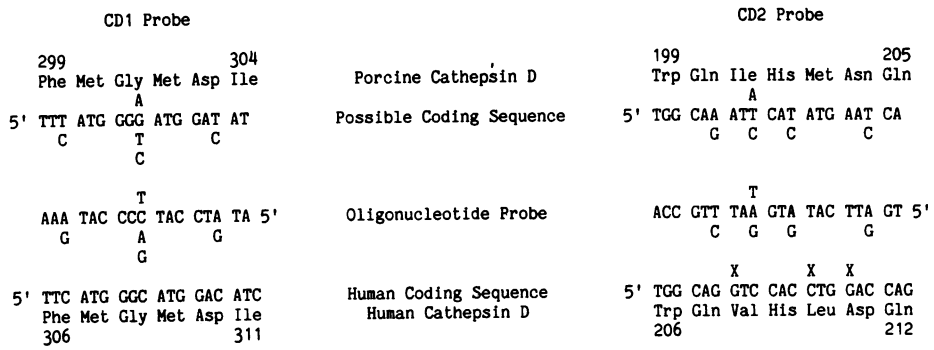

FIG. 1. Oligonucleotide probes used for identifying human cathepsin D cDNA. The rationale for choosing oligonucleotide sequences is illustrated. X, mismatch of the oligonucleotide probe relative to the determined human cathepsin D sequence.

The human kidney library was screened with the partial cDNA for cathepsin D, obtained from the HepG2 library. All screening was performed as described above except that the prehybridization and hybridization temperature was raised to 55°C and stringent washing was done at 50°C.

Oligonucleotide Probes. Two mixed-oligonucleotide probes, based on the porcine sequence of cathepsin D, were synthesized with an Applied Biosystems solid-phase synthesizer. These probes were kindly provided by Gerald Galluppi (Monsanto).

Blot Hybridization Analysis. Southern blots (18) of the λgt10 positive clones were probed with renin exon 3 by hybridization in 20% (vol/vol) formamide/5× NaCl/Cit/0.02 M sodium phosphate, pH 6.5/0.01% sonicated salmon sperm DNA/1× Denhardt's solution/10% dextran sulfate for 24 hr at 42°C. The nitrocellulose filter was washed in 0.1× NaCl/Cit/0.1% $NaDodSO_4$ at 37°C prior to autoradiography. Hybridization of these blots with ^{32}P-end-labeled oligonucleotide probes (19) was in 6× NaCl/Cit/10× Denhardt's solution/torula tRNA (100 μg/ml) at 40°C for probe CD1 and 48°C for probe CD2. Filters were washed at 5°C below hybridization temperature in 6× NaCl/Cit.

Blot analysis of poly(A)$^+$ RNAs from various human, rat, mouse, and bovine tissues was done with the partial cathepsin D clone. Analysis was as described by Thomas (20) except that for the non-human RNAs the hybridization formamide concentration was lowered to 20% and stringent washing was at 42°C in 0.2× NaCl/Cit/0.1% $NaDodSO_4$.

DNA Sequencing Methods. All positive HepG2 λgt10 clones were inserted into the *Eco*RI sites of pUC19 and M13mp18 (13) for mapping and sequence analyses, respectively. Positive clones of the kidney library were inserted into the *Sac* I site of M13mp18. DNA sequencing, using shotgun subcloning as well as deletion subcloning (21), was performed by the dideoxy chain-termination method of Sanger, with recent modifications (22). Sequence through the homopolymer tails was obtained by the method of Maxam and Gilbert (19).

RESULTS

Isolation of a Partial cDNA Clone for Human Cathepsin D. A human hepatoma (HepG2) library in λgt10 was screened with a human renin genomic subclone, pH10BE0.6 (4), which contains the 124 base pairs (bp) of exon 3 coding plus flanking intron sequences. From the ≈150,000 plaques screened, 4 putative positive clones were obtained. These were plaque-purified and their DNA was isolated, digested with *Eco*RI, electrophoresed in an agarose gel, and then transferred to a nitrocellulose filter. Southern blot analysis under reduced stringency (see *Materials and Methods*) showed inserts of 800–1400 bp that hybridized with the renin probe (data not shown).

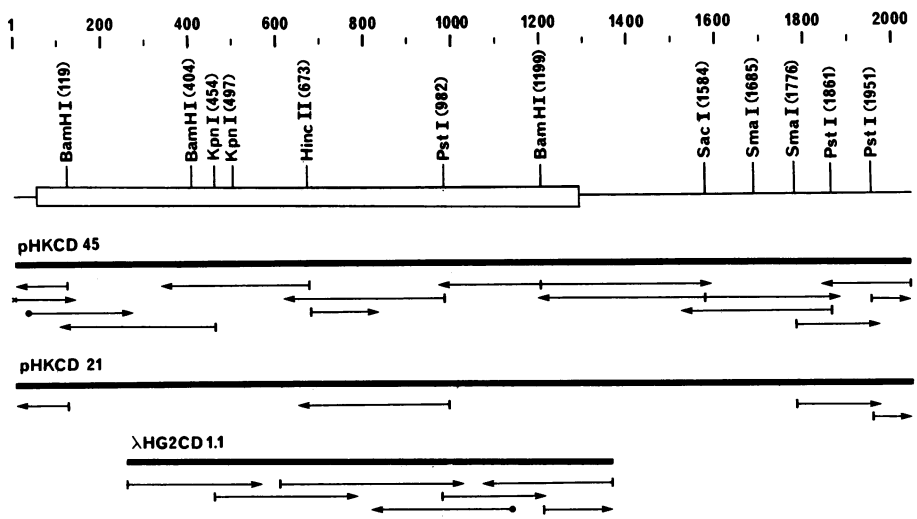

FIG. 2. Sequence analysis strategy of the human cathepsin D cDNA clones. Scale at top is in bp. The sequence corresponding to the coding region is indicated by the open box. Only the positions of relevant restriction sites are indicated. The poly(dA)·poly(dT) and poly(dG)·poly(dC) tails are not included in the map. Horizontal arrows indicate the direction and extent of sequence determination. Arrows originating at short vertical lines indicate Sanger (dideoxynucleotide chain-termination) sequencing with the M13 universal primer; those originating at solid circles represent Sanger sequencing primed with oligonucleotides based on cathepsin D; the arrow with an x at its origin indicates sequence determined from a 5'-end-labeled *Xba* I site (in pUC19 polylinker) by the chemical method of Maxam and Gilbert.

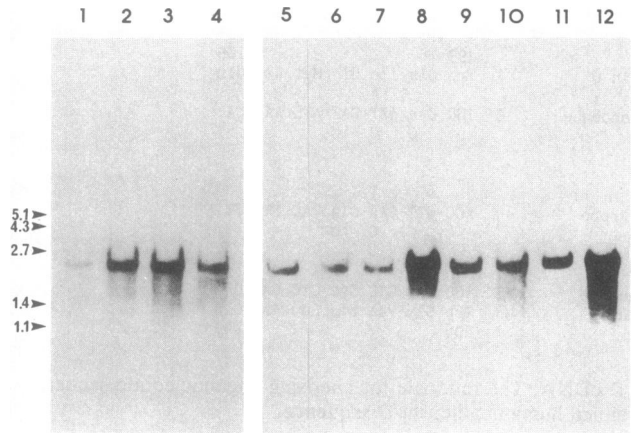

FIG. 3. Blot hybridization analysis of poly(A)$^+$ RNAs. Lanes: 1, human SWEi B lymphocytes (3 μg of RNA); 2, human JY fibroblasts (5 μg); 3, human kidney (2.5 μg); 4, HepG2 (4.5 μg); 5, juvenile bovine liver (5 μg); 6, rat brain (5 μg); 7, rat anterior pituitary (2.5 μg); 8, rat adrenal (3.2 μg); 9, rat liver (5 μg); 10, rat kidney (5 μg); 11, mouse pancreas (10 μg); 12, mouse submaxillary gland (4 μg). The positions and sizes (in kb) of denatured double-stranded DNA markers are indicated.

To verify that the clones contained cathepsin D sequences, we screened them with two mixed oligonucleotide probes corresponding to amino acids 299–304 (CD1) and 199–205 (CD2) of the porcine cathepsin D sequence (Fig. 1). On Southern analysis with ^{32}P-end-labeled probes, all four clones were negative with the CD2 oligonucleotide. However, one clone, λHG2CD1.1, showed a strong hybridization with the CD1 oligonucleotide. The 1100-bp insert of this isolate was then subcloned in the single-stranded phage M13mp18 and partially sequenced. Positive identification as cathepsin D cDNA was based on near identity with the porcine sequence. The lack of hybridization of this clone to the CD2 oligonucleotide was due to unexpected differences in the human amino acid sequence which caused three nucleotide mismatches with the 21-mer probe (Fig. 1).

Isolation of a Complete Human Cathepsin D cDNA Clone. The partial sequence analysis of λHG2CD1.1 showed that the clone did not contain the entire protein coding region (Fig. 2). Indeed, poly(A)$^+$ RNA blot analysis using this partial cDNA as a probe (Fig. 3) demonstrated a message length of about 2.2 kb in human tissues, as well as in many other species. This analysis also indicated that human kidney would be a good source of cathepsin D mRNA. To isolate longer cDNAs, a human kidney library was constructed, in the plasmid vector pUC19, only with cDNAs in the size range 1.8–2.5 kb. Screening at high stringency with the partial cathepsin cDNA clone and this size selection excluded the possibility of obtaining a renin clone from this library, since the renin message is about 1.6 kb long (23). From the 30,000 recombinants screened, 6 clones positive for hybridization were identified as cathepsin D cDNA by restriction mapping of their inserts. Two of these recombinant plasmids, pHKCD45 and pHKCD21, had inserts of 2200 and 2150 bp, respectively. These two isolates and the λHG2CD1.1 clone were subjected to further sequence analysis.

```
-64                                                                  M  Q  P  S  S  L  L  P  L  A  L  C  L  L  A  A  P   -48
  1  GGCTATAAGCGCACGGCCTCGGCGACCCTCTCCGACCCGGCCGCCGCCGCCATGCAGCCCTCCAGCCTTCTGCCGCTCGCCCTCTGCCTGCTGGCTGCACCC   102

-47  A  S  A▼L  V  R  I  P  L  H  K  F  T  S  I  R  R  T  M  S  E  V  G  G  S  V  E  D  L  I  A  K  G  P   -14
103  GCCTCCGCGCTCGTCAGGATCCCGCTGCACAAGTTCACGTCCATCCGCCGGACCATGTCGGAGGTTGGGGGCTCTGTGGAGGACCTGATTGCCAAAGGCCCC   204

-13  V  S  K  Y  S  Q  A  V  P  A  V  T  E▼▼G  P  I  P  E  V  L  K  N  Y  M  D  A  Q  V  Y  G  E  I  G  I    21
205  GTCTCAAAGTACTCCCAGGCGGTGCCAGCCGTGACCGAGGGGCCCATTCCCGAGGTGCTCAAGAACTACATGGACGCCCAGTACTACGGGGAGATTGGCATC   306
                                  •
 22  G  T  P  P  Q  C  F  T  V  V  F  D  T  G  S  S  N  L  W  V  P  S  I  H  C  K  L  L  D  I  A  C  W  I    55
307  GGGACGCCCCCCCAGTGCTTCACAGTCGTCTTCGACACGGGCTCCTCCAACCTGTGGGTCCCCTCCATCCACTGCAAACTGCTGGACATCGCTTGCTGGATC   408
                                                 ▼
 56  H  H  K  Y  N  S  D  K  S  S  T  Y  V  V  K  N  G  T  S  F  D  I  H  Y  G  S  G  S  L  S  G  Y  L  S  Q    89
409  CACCACAAGTACAACAGCGACAAGTCCAGCACCTACGTGAAGAATGGTACCTCGTTTGACATCCACTATGGCTCGGGCAGCCTCTCCGGGTACCTGAGCCAG   510

 90  D  T  V  S  V  P  C  Q  S  A  S  S  A  S  A  L  G  G  V  K  V  E  R  Q  V  F  G  E  A  T  K  Q  P  G   123
511  GACACTGTGTCGGTGCCCTGCCAGTCAGCGTCGTCAGCCTCTGCCCTGGGCGGTGTCAAAGTGGAGAGGCAGGTCTTTGGGGAGGCCACCAAGCAGCCAGGC   612

124  I  T  F  I  A  A  K  F  D  G  I  L  G  M  A  Y  P  R  I  S  V  N  N  V  L  P  V  F  D  N  L  M  Q  Q   157
613  ATCACCTTCATCGCAGCCAAGTTCGATGGCATCCTGGGCATGGCCTACCCCCGCATCTCCGTCAACAACGTGCTGCCCGTCTTCGACAACCTGATGCAGCAG   714
                                                                  (R)
158  K  L  V  D  Q  N  I  F  S  F  Y  L  S  R  D  P  D  A  Q  P  G  G  E  L  M  L  G  G  T  D  S  K  Y  Y   191
715  AAGCTGGTGGACCAGAACATCTTCTCCTTCTACCTGAGCAGGGACCCAGATGCGCAGCCTGGGGGTGAGCTGATGCTGGGTGGCACAGACTCCAAGTATTAC   816
                                                                  A
192  K  G  S  L  S  Y  L  N  V  T  R  K  A  Y  W  Q  V  H  L  D  Q  V  E  V  A  S  G  L  T  L  C  K  E  G   225
817  AAGGGTTCTCTGTCCTACCTGAATGTCACCCGCAAGGCCTACTGGCAGGTCCACCTGGACCAGGTGGAGGTGGCCAGCGGGCTGACCCTGTGCAAGGAGGGC   918
                •
226  C  E  A  I  V  D  T  G  T  S  L  M  V  G  P  V  D  E  V  R  E  L  Q  K  A  I  G  A  V  P  L  I  Q  G   259
919  TGTGAGGCCATTGTGGACACAGGCACTTCCCTCATGGTGGGCCCGGTGGATGAGGTGCGCGAGCTGCAGAAGGCCATCGGGGCCGTGCCGCTGATTCAGGGC  1020

260  E  Y  M  I  P  C  E  K  V  S  T  L  P  A  I  T  L  K  L  G  G  K  G  Y  K  L  S  P  E  D  Y  T  L  K   293
1021 GAGTACATGATCCCCTGTGAGAAGGTGTCCACCCTGCCCGCGATCACACTGAAGCTGGGAGGCAAAGGCTACAAGCTGTCCCCAGAGGACTACACGCTCAAG  1122

294  V  S  Q  A  G  K  T  L  C  L  S  G  F  M  G  M  D  I  P  P  P  S  G  P  L  W  I  L  G  D  V  F  I  G   327
1123 GTGTCGCAGGCCGGGAAGACCCTCTGCCTGAGCGGCTTCATGGGCATGGACATCCCGCCACCCAGCGGGCCACTCTGGATCCTGGGCGACGTCTTCATCGGC  1224

328  R  Y  Y  T  V  F  D  R  D  N  N  R  V  G  F  A  E  A  A  R  L                                          348
1225 CGCTACTACACTGTGTTTGACCGTGACAACAACAGGGTGGGCTTCGCCGAGGCTGCCCGCCTCTAGTTCCCAAGGCGTCCGCGCGCCAGCACAGAAACAGAG  1326
1327 GAGAGTCCCAGAGCAGGAGGCCCCTGGCCCAGCGGCCCCTCCCACACACACCCACACACTCGCCCGCCCACTGTCCTGGGCGCCCTGGAAGCCGGCGGCCCA  1428
1429 AGCCCGACTTGCTGTTTTGTTCTGTGGTTTTCCCCTCCCTGGGTTCAGAAATGCTGCCTGCCTGTCTGTCTCTCCATCTGTTTGGTGGGGGTAGAGCTGATC  1530
1531 CAGAGCACAGATCTGTTTCGTGCATTGGAAGACCCCACCCAAGCTTGGCAGCCGAGCTCGTGTATCCTGGGGCTCCCTTCATCTCCAGGGAGTCCCCTCCCC  1632
1633 GGCCCTACCAGCGCCCGCTGGGCTGAGCCCCTACCCCACACCAGGCCGTCCTCCCGGGCCCTCCCTTGGAAACCTGCCCTGCCTGAGGGCCCCTCTGCCCAG  1734
1735 CTTGGGCCCAGCTGGGCTCTGCCACCCTACCTGTTCAGTGTCCCGGGCCCGTTGAGGATGAGGCCGCTAGAGGCCTGAGGATGAGCTGGAAGGAGTGAGAGG  1836
1837 GGACAAAACCCACCTTGTTGGAGCCTGCAGGGTGGTGCTGGGACTGAGCCAGTCCCAGGGGCATGTATTGGCCTGGAGGTGGGGTTGGGATTGGGGGCTGGT  1938
1939 GCCAGCCTTCCTCTGCAGCTGACCTCTGTTGTCCTCCCCTTGGGCGGCTGAGAGCCCCAGCTGACATGGAAATACAGTTGTTGGCCTCCGGCCTCCCCTC  2038
```

FIG. 4. Nucleotide and corresponding amino acid sequence of pHKCD45 encoding human preprocathepsin D. The deduced amino acid residues (standard one letter code) are indicated above the nucleotide sequence. Amino acid no. 1 is assigned to the first residue of the mature protein. For the λHG2CD1.1 clone, the single nucleotide difference at position 775 is indicated below the line and the resulting amino acid encoded is indicated in parentheses above the line. The single and double arrowheads indicate the ends of signal sequence and prosequence, respectively. Asterisks indicate active-site aspartyl residues (nos. 33 and 231). Arrows point to N-glycosylation sites (asparagine-70 and -199). The presumed polyadenylylation signal within the 3′ untranslated region is underlined.

Nucleotide Sequence Analysis. Fig. 2 shows the restriction map and sequencing strategy for the cloned cDNAs. The complete nucleotide sequence of pHKCD45 is presented in Fig. 4. Sequence was determined on both the message and complementary strands. pHKCD45 and pHKCD21 turned out to be identical in 5′ and 3′ extent except for the length of poly(A) tail in the clones. There was, however, a single base pair difference between the human kidney clones and the HepG2 clone. This nonconservative difference occurs at nucleotide 775 in the pHKCD45 sequence and changes the predicted amino acid sequence (Fig. 4). Whether this represents a true human polymorphism, is merely specific to HepG2 cells, or is a cloning artifact can be investigated, as the restriction endonuclease *Avr* II recognizes the sequence only as it is present in the HepG2 clone.

The expected 5′ AATAAA 3′ poly(A) addition signal (24) does not occur in either full-length clone obtained from the kidney library, although both clones were polyadenylylated. There is a hexanucleotide 5′ AATACA 3′ which begins 30 nucleotides upstream from the poly(A) tail. This noncanonical sequence has recently been shown to be a functional polyadenylylation signal (25).

Predicted Amino Acid Sequence of Human Cathepsin D. The amino acid sequence predicted by nucleotide analysis is shown in Fig. 4. Biosynthetic studies of cathepsin D have shown it to be synthesized with pre- and propeptide regions (26, 27). The NH_2-terminal region presented here consists of many hydrophobic amino acids characteristic of a signal peptide. The assignment of residue −43 as the beginning of the propiece is based on work by Erickson *et al.* (26), where the prepiece was shown to be 20 amino acids long. Similarities in the NH_2-terminal residues determined for the propiece of porcine cathepsin D (26) with those presented here confirm this determination (Fig. 5). The beginning of the mature protein at residue 1 is also a well-conserved sequence (28). The open reading frame continues until a termination codon at nucleotide 1288 delimits a polypeptide chain of 412 amino acids. There are then 20 amino acids in the prepiece, 44 amino acids in the propiece, and 348 amino acids in the mature protein. The molecular weight for the mature unglycosylated protein calculated from the predicted sequence is 37,800. The sequence predicts N-glycosylation sites at positions 70 and 199, both of which have been found to be glycosylated (29, 30).

DISCUSSION

A human renin genomic fragment has been used in the isolation of a cDNA clone coding for human cathepsin D. The full-length clones eventually obtained were found by sequencing to be 2038 nucleotides long, with 51 nucleotides in the 5′ noncoding region, 1235 nucleotides in the coding region, and 752 nucleotides in the 3′ untranslated region preceding the poly(A) tail. The cathepsin D message length was determined to be about 2.2 kb by blot analysis of electrophoretically fractionated poly(A)$^+$ RNA (Fig. 3). This was much larger than expected for a precursor protein of M_r 53,000 (27) and also much larger than the 1.5- to 1.6-kb message lengths for renin, pepsinogen (31), and chymosin (32). The larger size of the cathepsin D message has been shown here to be due to a long 3′ untranslated region. It is interesting that a cDNA clone for rat cathepsin B, another lysosomal enzyme with a precursor protein of M_r 43,000, has been found to have a message 2.3 kb long, with the increase in length also a consequence of an extended 3′ noncoding region (33). Cathepsin B is a member of the thiol protease gene family and structurally unrelated to cathepsin D. Although of unknown significance, perhaps this long 3′ untranslated region is a common feature of lysosomal protease mRNAs.

The amino acid sequences of human and porcine cathepsin Ds, human renin (23), human pepsinogen (31), and bovine chymosin (32) are compared in Fig. 5. The alignment between the two cathepsin D sequences reveals that there are nine more amino acids in the human protein than in the porcine sequence. Two of these are at the COOH terminus. The other seven are at residues 98–104 in the human mature protein and are thus positioned between the previously determined junction of the light chain and heavy chain of porcine cathepsin D (3). Since the porcine amino acid sequence was determined from isolated light chain and heavy chain, the existence of seven additional amino acids at this position suggests that they are present in the single-chain enzyme but lost during proteolytic conversion to the two-chain form of the enzyme;

```
H-CD  -64  MQPSSLLPLALCLLAAPASA----------LVRIPLHKFTSIRRTMSEVGGSVEDLIAK--GPVSKYSQA  -7
H-RN  -66  .DGWRRM.RWGL..LLWGSC TFGLPTDTTTFK..F.KRMP...ESLK.R.VDMAR.GPEWSQ.MKR----  -1
H-PG  -62  .KW-----.L.LG.V.LSEC ---------IMYKV..IRKK.L...L..R.LLKDF.KKHNLN.AR..FPQ  -7
B-CH  -58  .RC----LVV.LAVF.LSQG --------AEIT....Y.GK.L.KALK.H.LLEDF.QKQQY.IS....GF  -1
P-CD        . .  ..L. . .. .              .I ...    .  .    P  .. .   I .

H-CD   -6  VPAVTEGPIPEVLKNYMDAQYYGEIGIGTPPQCFTVVFDTGSSNLWVPSIHCKLLDIACWIHHKYNSDKSST  66
H-RN    1  -LTLGNTTSSVI.T....T.............T.K...|.....V....SK.SR.YT..VY.KLFDASD..S  71
H-PG   -6  WE.P.L-VDEQP.E..L.ME.F.T......A.D.....|..........VY.SS.--..TN.NRF.PED...  63
B-CH    1  -----GEVASVP.T..L.S..F.K.YL.....E...L..|.....DF.....Y..SN--..KN.QRFDPR....  65
P-CD    1        ..............................|.............................G....  66

H-CD   67  YVKNGTSFDIHYGSGSLSGYLSQDTVSVPCQSASSASALGGVKVERQVFGEATKQPGITFIAAKFDGILGMA  138
H-RN   72  .KH...ELTLR.ST.TV..F....IIT.------------..IT.T-.M.....EM.ALP.ML.E...VV..G  131
H-PG   64  .QSTSETVS.T..T..MT.I.GY...Q.------------..ISDTN.I..LSETE..SFLYY.P......L.  124
B-CH   66  FQNL.KPLS....T..MQ.I.GY...T.-----------SNIVDIQ.TV.LS.QE..DV.TY.E.......  126
P-CD   67  .....T.A.....................SN-------V..I.....T.........L.............  131

H-CD  139  YPRISVNNVLPVFDNLMQQKLVDQNIFSFYLSRD-PDAQP-GGELMLGGTDSKYYKGSLSYLNVTRKAYWQV  208
H-RN  132  FIEQRIGR.T.I...IIS.GVLKEDV....YN..SENS.SL..QIV...S.PQH.E.NFH.I.LIKTGV..I  203
H-PG  125  ..S..SSGAT.....IWN.G..SQDL..V...A.---D.S-.SVVIF..I..S..T...NWVP..VEG...I  192
B-CH  127  ..SLASEYSI.....M.NRH..A.DL..V.MD.N---G.E--SM.T..AI.PS..T...HWVP..VQQ...F  193
P-CD  132  .........V.............KD......N..-.G...-........I.........D.H.....I  201

H-CD  209  HLDQVEVASGLTLCKEGCEAIVDTGTSLMVGPVDEVRELQKAIGAVPLIQGEYMIPCEKVSTLPAITLKLGG  280
H-RN  204  QMKG.S.G.STL..ED..L.L.|..A.YIS.STSSIEK.ME.L..K-KRLFD.VVK.NEGP...D.SFH...  274
H-PG  193  TV.SITMNGEAIA.A...Q...|.....LT..TSPIANI.SD...SENSD.DMVVS.SAI.S..D.VFTIN.  264
B-CH  194  TV.S.TISGVVVA.EG..Q..L|.....KL...SSDILNI.Q....TQNQY..FD.D.DNL.YM.TVVFEIN.  265
P-CD  202  .MN..A.G.S.....G......|.....I..QPE.....G.....................PS..DV.VT...  273

H-CD  281  KGYKLSPEDYTLKVSQAGKTLCLSGFMGMDIPPPSGPLWILGDVFIGRYYTVFDRDNNRVGFAEAARL  348
H-RN  275  .E.T.TSA..VFQE.YSS.K..TLAIHA......T..T.A..AT..RKF..E...R...I...L.R--  340
H-PG  265  VQ.PVP.SA.I.QSEGS-----.I...Q..NL.TE..E.........RQ.F.....A..Q..L.PV.--  326
B-CH  266  .M.P.T.SA..SQDQGF-----.T...QSEN---H.-QK........RE..S....A..L..L.K.I--  323
P-CD  274  .K....S.N..........Q.I.............G.....................L....L.....--  339
```

FIG. 5. Alignment of amino acid sequences of human cathepsin D (H-CD), human renin (H-RN), human pepsinogen (H-PG), bovine chymosin (B-CH), and porcine cathepsin D (P-CD). Dots indicate insertion of gaps at positions at which there are no homologous residues. Blanks in the porcine sequence indicate amino acids not determined. The active-site aspartyl residues are boxed. Potential glycosylation signals are underlined. The vertical arrows indicate possible sites of proteolytic cleavage of single-chain cathepsin D. The large arrowhead denotes the end of the signal sequence for all of the proteins. Locations of exon splice junctions for human renin and pepsin genes are indicated by the small arrowheads; the arrowhead with an asterisk points to a junction unique to the renin gene.

this raises some uncertainties as to the actual position and nature of the proteolytic cleavage.

Erickson and Blobel (34) have shown that very late in the biosynthesis of cathepsin D in porcine kidney cells, the mass of the 30-kDa heavy chain decreases by about 1 kDa. The loss of seven amino acids from the NH_2 terminus of the heavy chain plus two from its COOH terminus would correspond to about 1 kDa. This suggests that the initial cleavage of the single-chain enzyme is between residues 97 and 98 in the human sequence, followed by trimming of both ends of the heavy chain.

The significance of these cleavages is unknown. Both single-chain and two-chain cathepsin D are proteolytically active (35). The rat cathepsin B cDNA clone has been found to encode six amino acids COOH-terminal to the previously determined end of the mature protein as well as two additional amino acids at the junction site of its light and heavy chains (33). Other lysosomal enzymes, such as β-glucuronidase (34) and α-L-iduronidase (36), show decreases in molecular mass at similar times in their biosynthesis.

Cathepsin D has been isolated from a wide variety of mammalian tissues (1). That the protein is highly conserved is evidenced by the 87% amino acid identity between human kidney and porcine spleen cathepsin D (Fig. 5). The results of the RNA blot hybridization (Fig. 3) are also consistent with the protein data: cathepsin D mRNA seems to be present in a variety of tissues and highly conserved among four mammalian species (man, cow, rat, and mouse).

Fig. 5 demonstrates the marked degree of conservation in the aspartyl protease family. In the mature-protein regions, cathepsin D shows 46%/58% identical amino acid/nucleotide residues with renin, 49%/58% with pepsin, and 47%/57% with chymosin. Renin and pepsin have 39%/53% amino acid/nucleotide identities. Thus renin and pepsin are somewhat more similar to cathepsin D than to each other. Sequence alignment in the propeptide region also shows residues conserved in the four aspartyl proteases. A comparison of propeptide sequences for human, bovine (37), porcine (38), and chicken (39) pepsinogens reveals that an average of 64% of the residues are conserved. The propeptide sequence for bovine chymosin has an average 52% homology with these pepsinogen propeptides. Similar analysis for cathepsin D and renin shows average homologies of 43% and 34%, respectively. Further divergence of the renin prosequence, not taken into account in this comparison, is evidenced by the need to insert gaps in order to reveal the identities with pepsinogen (Fig. 5). These comparisons suggest common structural features for the pepsin, chymosin, and cathepsin D prosegments, perhaps relating to the acid-activation of these proteins (40, 41).

The homology observed throughout the length of the aspartyl proteases indicates that their three-dimensional structures are very similar (2). Cathepsin D, however, is the only one of the proteins that is targeted to lysosomes. Renin, pepsin, and chymosin are all targeted to secretory granules. This family of proteins then offers a unique system for examining structural differences that may represent intracellular protein-targeting signals.

We thank Dr. Peter S. Rotwein, Dr. John C. Rogers, and Ms. Ida M. Schaefer for their help and valuable discussions. This investigation was supported by grants R01 CA08759 and 5732 GM07200 from the U.S. Public Health Service and by grant 83-1271 from the American Heart Association.

1. Barret, A. J. (1977) in *Proteinases in Mammalian Cells and Tissues*, ed. Barret, A. J. (North Holland, New York), pp. 209–248.
2. Tang, J. (1979) *Mol. Cell. Biochem.* **26**, 93–109.
3. Shewale, J. G. & Tang, J. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 3703–3707.
4. Hobart, P. M., Fogliano, M., O'Connor, B. A., Schaefer, I. M. & Chirgwin, J. M. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 5026–5030.
5. Miyazaki, H., Fukamizu, A., Hirose, S., Hayashi, T., Hori, H., Ohkubo, H., Nakanishi, S. & Murakami, K. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 5999–6003.
6. Kemp, D. J., Coppel, R. L., Cowman, A. F., Saint, R. B., Brown, G. V. & Anders, R. F. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3787–3791.
7. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **24**, 5294–5299.
8. Aviv, H. & Leder, P. (1972) *Proc. Natl. Acad. Sci. USA* **69**, 1408–1412.
9. Retzell, E. F., Collett, M. S. & Faras, A. J. (1980) *Biochemistry* **19**, 513–518.
10. Gubler, U. & Hoffman, B. J. (1983) *Gene* **25**, 263–269.
11. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
12. Deng, G. R. & Wu, R. (1981) *Nucleic Acids Res.* **9**, 4173–4188.
13. Norrander, J., Kempe, T. & Messing, J. (1983) *Gene* **26**, 101–106.
14. Hanahan, D. (1983) *J. Mol. Biol.* **166**, 557–580.
15. Hanahan, D. & Meselson, M. (1983) *Methods Enzymol.* **100**, 333–342.
16. Woo, S. L. C. (1979) *Methods Enzymol.* **68**, 389–395.
17. Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) *J. Mol. Biol.* **113**, 237–251.
18. Southern, E. M. (1975) *J. Mol. Biol.* **98**, 503–517.
19. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
20. Thomas, P. S. (1983) *Methods Enzymol.* **100**, 255–266.
21. Hong, G. F. (1982) *J. Mol. Biol.* **158**, 539–549.
22. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3963–3965.
23. Imai, T., Miyazaki, H., Hirose, S., Hori, H. Hayashi, T., Ryoichiro, K., Ohkubo, H., Shigetada, N. & Murakami, K. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 7405–7409.
24. Nevins, J. R. (1983) *Annu. Rev. Biochem.* **52**, 441–446.
25. Mason, P. J., Jones, M. B., Elkington, J. A. & Williams, J. G. (1985) *EMBO J.* **4**, 205–211.
26. Erickson, A. H., Conner, G. E. & Blobel, G. (1981) *J. Biol. Chem.* **256**, 11224–11231.
27. Gieselmann, V., Pohlmann, R., Hasilik, A. & von Figura, K. (1983) *J. Cell Biol.* **97**, 1–5.
28. Takahashi, T. & Tang, J. (1981) *Methods Enzymol.* **80**, 565–581.
29. Takahashi, T., Schmidt, P. & Tang, J. (1983) *J. Biol. Chem.* **258**, 2819–2830.
30. Hasilik, A. & von Figura, K. (1981) *Eur. J. Biochem.* **121**, 125–129.
31. Sogawa, K., Fujii-Kuriyama, Y., Mizukami, Y., Ichihara, Y. & Takahashi, K. (1983) *J. Biol. Chem.* **258**, 5306–5311.
32. Harris, T. J. R., Lower, P. A., Thomas, P. G., Eaton, M. A. W., Millican, T. A., Patel, T. P., Bose, C. C., Carey, N. H. & Doel, M. T. (1982) *Nucleic Acids Res.* **10**, 2177–2187.
33. San Segundo, B., Chan, S. J. & Steiner, D. F. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 2320–2324.
34. Erickson, A. H. & Blobel, G. (1983) *Biochemistry* **22**, 5201–5205.
35. Huang, J. S., Huang, S. S. & Tang, J. (1979) *J. Biol. Chem.* **254**, 11405–11417.
36. Myerowitz, R. & Neufeld, E. (1981) *J. Biol. Chem.* **256**, 3044–3048.
37. Harboe, M., Andersen, P. M., Foltmann, B., Kay, J. & Kassell, B. (1974) *J. Biol. Chem.* **249**, 4487–4494.
38. Stepanov, V. M., Baratova, L. A., Pugacheva, I. B., Belyanova, L. P., Revina, L. P. & Timokhina, E. A. (1973) *Biochem. Biophys. Res. Commun.* **54**, 1164–1170.
39. Baudys, M. & Kostka, V. (1983) *Eur. J. Biochem.* **136**, 89–99.
40. Hartsuck, J. A., Marciniszyn, J., Jr., Huang, J. S. & Tang, J. (1977) in *Acid Proteases: Structure, Function and Biology*, ed. Tang, J. (Plenum, New York), pp. 85–102.
41. Hasilik, A., von Figura, K., Conzelmann, E., Nehrkorn, H. & Sandhoff, K. (1982) *Eur. J. Biochem.* **125**, 317–321.