FEBS Letters 369 (1995) 326–330 FEBS 15849

Molecular cloning and sequence analysis of human preprocathepsin C

Alenka Pariš*, Borut Štrukelj, Jože Pungerčar, Metka Renko, Iztok Dolenc, Vito Turk

Department of Biochemistry and Molecular Biology, Jožef Stefan Institute, Jamova 39, 61000 Ljubljana, Slovenia

Received 29 June 1995

Abstract **A cDNA clone (C1) coding for human preprocathepsin C was isolated from a human ileum cDNA library using a rat kidney-derived RT-PCR probe and its complete nucleotide sequence determined. The full-length 1857 bp sequence codes for a protein of 463 amino acid residues with a calculated molecular mass of 51848 Da. Comparison of the deduced amino acid sequence with that of rat preprocathepsin C indicates an 87.5% identity. A multiple alignment of the deduced cathepsin C sequence of 233 residues which, by analogy to other cysteine proteinases, corresponds to the mature protein, confirms that human cathepsin C belongs to the papain superfamily.**

Key words: Preprocathepsin C; Ileum (human); cDNA cloning; Cysteine proteinase

1. Introduction

Cysteine proteinases are remarkably widespread and are present in almost all life forms. They catalyze the hydrolysis of many proteins with different specificities and are considered to play an important role in intracellular protein degradation and turnover [1]. The primary structures of human cathepsins B, L, H, S, K and O have been reported as protein sequences data or as deduced from their cDNAs. Since their sequences share a high degree of similarity and are similar to that of papain, they all belong to the papain superfamily [2].

Cathepsin C or dipeptidyl aminopeptidase I (EC 3.4.14.1) is a lysosomal proteinase, capable of removing dipeptides sequentially from the amino terminus of peptide and protein substrates [3]. It has been reported that for the dipeptidyl aminopeptidase activity, cathepsin C requires halide ion as well as reducing agents to achieve maximal hydrolytic activity [4]. Cathepsin C is, in addition to lysosomal proteolysis, involved in the functions of the alimentary tract [5], cell growth [6], neuraminidase activation [7] and proliferation of basal cell carcinomas [8]. Thiele and Lipsky reported that cathepsin C activity is present at higher levels in cytotoxic lymphocytes and myeloid cells, indicating its involvement in the induction, development or differentiation of cytolytic effector cells [9].

Unlike cathepsins B, H, L and S, which are small monomeric proteins with molecular masses of 20–30 kDa, cathepsin C is a high molecular mass oligomeric protein of about 200 kDa, as estimated from gel filtration analysis [10,11]. Recently, a cDNA clone coding for rat cathepsin C was isolated from a rat kidney cDNA library [5]. The deduced amino acid sequence has shown that rat cathepsin C has an extremely long propeptide and apparently belongs to the papain superfamily.

In this paper, we present for the first time the entire amino acid sequence of the human preprocathepsin C deduced from its nucleotide sequence. Additionally, in order to determine the copy number of the human cathepsin C gene, Southern blot analysis was performed.

*Corresponding author. Fax: (386) (61) 273 594.

2. Materials and methods

2.1. Materials

Restriction and DNA modifying enzymes were purchased from Boehringer-Mannheim (Germany) or Pharmacia (Sweden). All other chemicals of analytical grade were from Sigma (USA) and Serva (Germany) unless otherwise stated. Oligonucleotides were synthesized on an Applied Biosystems 381A DNA synthesizer (USA) and purified by polyacrylamide gel electrophoresis. [α-^{35}S]dGTP and [α-^{35}S]dATP used for hybridization and nucleotide sequencing, and nylon membranes (Hybond-N) were obtained from Amersham (UK). The bacterial strain *E. coli* Y1090 *hsd*R, used as a host for bacteriophage λgt11, was from Amersham, while *E. coli* DH5a was obtained from Gibco-BRL (USA). Plasmids pGEM-11Zf(−) and pUC19 were purchased from Promega (USA) and Pharmacia, respectively.

2.2. cDNA library construction

Total RNA was isolated from human ileum by the guanidinium thiocyanate/cesium trifluoroacetate method [12,13]. Poly(A)$^+$ RNA was purified by affinity chromatography on oligo(dT)-cellulose [14]. cDNA was synthesized using a cDNA Synthesis System Plus of Amersham. After fractionation by gel chromatography, a size-enriched cDNA of more than 400 bp was used for the construction of a cDNA library in λgt11 using a cloning kit from the same manufacturer.

2.3. Preparation of a probe for screening

In order to obtain the hybridization probe, total RNA was extracted from a rat kidney by the guanidinium thiocyanate/cesium trifluoroacetate method [12,13], and poly(A)$^+$ RNA was separated by affinity chromatography on oligo(dT)-cellulose [14]. The first strand of cDNA was performed with reverse transcriptase using an oligo(dT)$_{16}$ primer.

The single-stranded DNA was amplified using a PCR kit (Perkin-Elmer Cetus) with two oligonucleotide primers 5′-CCGAATTCGAATGACTACAAGTGG-3′ and 5′-CAGGGATCCGCCACGGACGTTTCT-3′ deduced from the nucleotide sequence of rat cathepsin C [5]. The amplified product of about 450 bp was cloned into pUC19, sequenced and ^{35}S-labeled using a Random Primed DNA Labeling kit (Boehringer Mannheim).

2.4. Screening of cDNA library

Recombinant plaques were transferred onto nylon membranes, fixed with ultraviolet light, and hybridized as described in [15] with ^{35}S-labeled probe (2×10^9 cpm/μg) at 42°C for 24 h. The replica filters were subsequently washed in $1 \times$ SSC, 0.1% SDS at 25°C for 10 min, and $0.1 \times$ SSC, 0.1% SDS at 50°C for 1 min, dried and exposed to Kodak X-Omat S film. Positive plaques were subjected to a second screening under the same conditions.

2.5. Nucleotide sequencing and sequence analysis

Phage DNA containing cDNA for human cathepsin C was isolated from plate lysates using Wizard Lambda Preps DNA Purification System (Promega). The cDNA was excised with *Xho*I, ligated into pGEM-11Zf(−) and sequenced by the dideoxy chain termination method (16) using a T7 sequencing kit (Pharmacia). The complete nucleotide sequence of both strands was determined by use of internal primers constructed on the basis of the previously determined sequence. Nucle-

0014-5793/95/$9.50 © 1995 Federation of European Biochemical Societies. All rights reserved.
SSDI 0014-5793(95)00777-6

otide and protein sequences were analyzed on a computer by DNASIS (Pharmacia) and PC/GENE (IntelliGenetics, USA) programmes.

2.6. Southern blot analysis

20 μg of genomic DNA isolated from human spleen was digested with *Eco*RI, *Hin*dIII and *Pst*I. The restriction fragments were separated by gel electrophoresis on a 0.6% agarose, blotted by capillary transfer (20 × SSC, 20 h) onto Hybond-N nylon membrane and UV-crosslinked. A cDNA fragment of the first 510 bp, corresponding to the 5′-end of the C1 clone, was labeled by random priming with [^{32}P]dGTP (110 TBq/mmol, Amersham) using a Random Primed DNA Labeling Kit (Boehringer Mannheim). The Southern blot was hybridized at 42°C for 24 h in 6 × SSC, 5 × Denhardt's solution, 50% deionized formamide, 0.1% SDS and 50 mg/ml of sonicated salmon sperm DNA. Washes were

```
   1  AAT TCT TCA CCT CTT TTC TCA GCT CCC TGC AGC ATG GGT GCT GGG CCC TCC TTG CTG CTC GCC GCC    66
   1                                              Met Gly Ala Gly Pro Ser Leu Leu Leu Ala Ala    11

  67  CTC CTG CTG CTT CTC TCC GGC GAC GGC GCC GTG CGC TGC GAC ACA CCT GCC AAC TGC ACC TAT CTT   132
  12  Leu Leu Leu Leu Leu Ser Gly Asp Gly Ala Val Arg Cys Asp Thr Pro Ala Asn Cys Thr Tyr Leu    33

 133  GAC CTG CTG GGC ACC TGG GTC TTC CAG GTG GGC TCC AGC GGT TCC CAG CGC GAT GTC AAC TGC TCG   198
  34  Asp Leu Leu Gly Thr Trp Val Phe Gln Val Gly Ser Ser Gly Ser Gln Arg Asp Val Asn Cys Ser    55

 199  GTT ATG GGA CCA CAA GAA AAA AAA GTA GTG GTG TAC CTT CAG AAG CTG GAT ACA GCA TAT GAT GAC   264
  56  Val Met Gly Pro Gln Glu Lys Lys Val Val Val Tyr Leu Gln Lys Leu Asp Thr Ala Tyr Asp Asp    77

 265  CTT GGC AAT TCT GGC CAT TTC ACC ATC ATT TAC AAC CAA GGC TTT GAG ATT GTG TTG AAT GAC TAC   330
  78  Leu Gly Asn Ser Gly His Phe Thr Ile Ile Tyr Asn Gln Gly Phe Glu Ile Val Leu Asn Asp Tyr    99

 331  AAG TGG TTT GCC TTT TTT AAG TAT AAA GAA GAG GGC AGC AAG GTG ACC ACT TAC TGC AAC GAG ACA   396
 100  Lys Trp Phe Ala Phe Phe Lys Tyr Lys Glu Glu Gly Ser Lys Val Thr Thr Tyr Cys Asn Glu Thr   121

 397  ATG ACT GGG TGG GTG CAT GAT GTG TTG GGC CGG AAC TGG GCT TGT TTC ACC GGA AAG AAG GTG GGA   462
 122  Met Thr Gly Trp Val His Asp Val Leu Gly Arg Asn Trp Ala Cys Phe Thr Gly Lys Lys Val Gly   143

 463  ACT GCC TCT GAG AAT GTG TAT GTC AAC ACA GCA CAC CTT AAG AAT TCT CAG GAA AAG TAT TCT AAT   528
 144  Thr Ala Ser Glu Asn Val Tyr Val Asn Thr Ala His Leu Lys Asn Ser Gln Glu Lys Tyr Ser Asn   165

 529  AGG CTC TAC AAG TAT GAT CAC AAC TTT GTG AAA GCT ATC AAT GCC ATT CAG AAG TCT TGG ACT GCA   594
 166  Arg Leu Tyr Lys Tyr Asp His Asn Phe Val Lys Ala Ile Asn Ala Ile Gln Lys Ser Trp Thr Ala   187

 595  ACT ACA TAC ATG GAA TAT GAG ACT CTT ACC CTG GGA GAT ATG ATT AGG AGA AGT GGT GGC CAC AGT   660
 188  Thr Thr Tyr Met Glu Tyr Glu Thr Leu Thr Leu Gly Asp Met Ile Arg Arg Ser Gly Gly His Ser   209

 661  CGA AAA ATC CCA AGG CCC AAA CCT GCA CCA CTG ACT GCT GAA ATA CAG CAA AAG ATT TTG CAT TTG   726
 210  Arg Lys Ile Pro Arg Pro Lys Pro Ala Pro Leu Thr Ala Glu Ile Gln Gln Lys Ile Leu His Leu   231

 727  CCA ACA TCT TGG GAC TGG AGA AAT GTT CAT GGT ATC AAT TTT GTC AGT CCT GTT CGA AAC CAA GCA   792
 232  Pro Thr Ser Trp Asp Trp Arg Asn Val His Gly Ile Asn Phe Val Ser Pro Val Arg Asn Gln Ala   253

 793  TCC TGT GGC AGC TGC TAC TCA TTT GCT TCT ATG GGT ATG CTA GAA GCG AGA ATC CGT ATA CTA ACC   858
 254  Ser Cys Gly Ser Cys Tyr Ser Phe Ala Ser Met Gly Met Leu Glu Ala Arg Ile Arg Ile Leu Thr   275
                      ▲
 859  AAC AAT TCT CAG ACC CCA ATC CTA AGC CCT CAG GAG GTT GTG TCT TGT AGC CAG TAT GCT CAA GGC   924
 276  Asn Asn Ser Gln Thr Pro Ile Leu Ser Pro Gln Glu Val Val Ser Cys Ser Gln Tyr Ala Gln Gly   297

 925  TGT GAA GGC GGC TTC CCA TAC CTT ATT GCA GGA AAG TAC GCC CAA GAT TTT GGG CTG GTG GAA GAA   990
 298  Cys Glu Gly Gly Phe Pro Tyr Leu Ile Ala Gly Lys Tyr Ala Gln Asp Phe Gly Leu Val Glu Glu   319

 991  GCT TGC TTC CCC TAC ACA GGC ACT GAT TCT CCA TGC AAA ATG AAG GAA GAC TGC TTT CGT TAT TAC  1056
 320  Ala Cys Phe Pro Tyr Thr Gly Thr Asp Ser Pro Cys Lys Met Lys Glu Asp Cys Phe Arg Tyr Tyr   341

1057  TCC TCT GAG TAC CAC TAT GTA GGA GGT TTC TAT GGA GGC TGC AAT GAA GCC CTG ATG AAG CTT GAG  1122
 342  Ser Ser Glu Tyr His Tyr Val Gly Gly Phe Tyr Gly Gly Cys Asn Glu Ala Leu Met Lys Leu Glu   363

1123  TTG GTC CAT CAT GGG CCC ATG GCA GTT GCT TTT GAA GTA TAT GAT GAC TTC CTC CAC TAC AAA AAG  1188
 364  Leu Val His His Gly Pro Met Ala Val Ala Phe Glu Val Tyr Asp Asp Phe Leu His Tyr Lys Lys   385

1189  GGG ATC TAC CAC CAC ACT GGT CTA AGA GAC CCT TTC AAC CCC TTT GAG CTG ACT AAT CAT GCT GTT  1254
 386  Gly Ile Tyr His His Thr Gly Leu Arg Asp Pro Phe Asn Pro Phe Glu Leu Thr Asn His Ala Val   407
                                                                              ▲
1255  CTG CTT GTG GGC TAT GGC ACT GAC TCA GCC TCT GGG ATG GAT TAC TGG ATT GTT AAA AAC AGC TGG  1320
 408  Leu Leu Val Gly Tyr Gly Thr Asp Ser Ala Ser Gly Met Asp Tyr Trp Ile Val Lys Asn Ser Trp   429

1321  GGC ACC GGC TGG GGT GAG AAT GGC TAC TTC CGG ATC CGC AGA GGA ACT GAT GAG TGT GCA ATT GAG  1386
 430  Gly Thr Gly Trp Gly Glu Asn Gly Tyr Phe Arg Ile Arg Arg Gly Thr Asp Glu Cys Ala Ile Glu   451

1387  AGC ATA GCA GTG GCA GCC ACA CCA ATT CCT AAA TTG TAG GGT ATG CCT TCC AGT ATT TCA TAA TGA  1452
 452  Ser Ile Ala Val Ala Ala Thr Pro Ile Pro Lys Leu ***                                       464

1453  TCT GCA TCA GTT GTA AAG GGG AAT TGG TAT ATT CAC AGA CTG TAG ACT TTC AGC AGC AAT CTC AGA  1518
1519  AGC TTA CAA ATA GAT TTC CAT GAA GAT ATT TGT CTT CAG AAT TAA AAC TGC CCT TAA TTT TAA TAT  1584
1585  ACC TTT CAA TCG GCC ACT GGC CAT TTT TTT CTA AGT ATT CAA TTA AGT GGG AAT TTT CTG GAA GAT  1650
1651  GGT CAG CTA TGA AGT AAT AGA GTT TGC TTA ATC ATT TGT AAT TCA AAC ATG CTA TAT TTT TTA AAA  1716
1717  TCA ATG TGA AAA CAT AGA CTT ATT TTT AAA TTG TAC CAA TCA CAA GAA AAT AAT GGC AAT AAT TAT  1782
1783  CAA AAC TTT TAA AAT AGA TGC TCA TAT TTT TAA AAT AAA GTT TTA AAA ATA ACT GCA AAA AAA AAA  1848
1849  AAA AAA AAA                                                                              1857
```

Fig. 1. The nucleotide and deduced amino acid sequences of preprocathepsin C. Polyadenilation signal is underlined. The protein sequence region is numbered starting from the putative initiation methionine and ending by the termination codon. Two amino acid residues (Cys^{258} and His^{405}) important for the catalytical activity are designated with arrowheads. These data have been submitted to the GenBank and have been assigned accession number X87212.

performed at 42°C, 55°C and 65°C in 2 × SSC, 0.1% SDS; 1 × SSC, 0.1% SDS and 0.1 × SSC, 0.1% SDS. The washed membrane was exposed to X-ray film.

3. Results and discussion

The aim of this study was to isolate clones coding for human cathepsin C and to determine its primary structure. In order to obtain a probe for screening, a DNA isolated from human ileum cDNA library was first used as a template. However, PCR analysis with the primers corresponding to the rat cathepsin cDNA [5] did not give any positive bands, therefore we decided to isolate total RNA from rat kidney. After poly(A)$^+$ RNA enrichment, the sample was subjected to reverse transcription and subsequent PCR amplification. The nucleotide sequence of the PCR product was confirmed and the probe was radioactively-labeled. Screening of about 2×10^5 independent clones from a human ileum cDNA library with the rat cDNA probe resulted in the isolation of four independent clones. The C1 cDNA clone with an approximative length of 2 kb was finally isolated and, after subcloning, completely sequenced.

The nucleotide sequence and the primary structure of the C1 clone deduced from its cDNA sequence are shown in Fig. 1. The 1857 bp cDNA contains an open reading frame of 1389 nucleotides, corresponding to an encoded preproprotein of 463 amino acid residues with a calculated molecular mass of 51848 Da. The initiation codon was assigned to the first in-frame ATG at position 34 which almost perfectly matches to the consensus Kozak sequence [17]. The coding sequence ends with a TAG stop codon at position 1323–1325. The 3′-untranslated region consisting of 413 nucleotides is followed by a short poly(A) tail. A putative polyadenylation signal (AATAAA), which is common to eukaryotic mRNAs, is located 23 nucleotides upstream of the poly(A) addition site. The first 24 amino acid residues of the protein show a typical hydrophobic character, similar to other eukaryotic signal sequences.

The N-terminal amino acid sequences of three polypeptide chains obtained by partial protein sequencing of cathepsin C isolated from human spleen [18] were confirmed in the deduced amino acid sequence of the C1 clone. The deduced protein sequences D^{25}TPANCTYLD^{34} and L^{231}PTSWDWRN^{239} correspond, by analogy with other cysteine proteinases, to the start of the proregion and of the mature region of cathepsin C, respectively. The third deduced sequence D^{395}PFNPFELTN^{404} probably represents the N-terminal end of the light chain of the mature protein. It was previously shown that an approximately 17 kDa polypeptide, with N-terminal amino acid sequence corresponding to the proregion of cathepsin C, was present in purified mature enzyme, indicating that a substantial part of the proregion still remains bound in the mature cathepsin C [10,18].

McGuire and co-workers [11] reported on the purification and characterization of dipeptidyl aminopeptidase I (cathepsin C) from human spleen. After isolation of cathepsin C, the purified protein was partially digested with trypsin and resulted fragments were analyzed by Edman degradation. Altogether 141 sequenced amino acid residues, compartmented into nine fragments, represent approximately 45% of mature cathepsin C. Comparison of the deduced amino acid sequence of C1 clone with the sequence of all nine fragments reveals 6 different amino acid residues. Rather unexpectedly, two cysteine residues (Cys^{321} and Cys^{355}) which are found in the deduced sequence of C1 clone and are preserved in the sequence of rat cathepsin C [5] are missing in the analyzed fragments [11]. The differences in the amino acid composition might be due to the presence of two related genes in human genome. Therefore, in order to determine the copy number of the human cathepsin C gene, a Southern blot of genomic DNA digested with *Eco*RI, *Pst*I and *Hin*dIII, respectively, was probed using a fragment corresponding to the first 510 bp of the C1 clone (Fig. 2). By analogy to the structures of the mouse cathepsin B gene [19], human cathepsin S gene [20] and human cathepsin L gene [21], at least one intron may be present within a selected region covering by the probe. Additionally, an internal *Pst*I restriction site is located within the sequence of the probe. Digestion of genomic DNA with *Hin*dIII gave two strong bands whereas *Eco*RI and *Pst*I digestion resulted in the presence of three bands. These results suggest that restriction enzymes may cut within the presumed intron region or that two copies of cathepsin C are present in human genome. The gene organization of other cathepsins [19–21] suggests that all three restriction sites are located in the intron domain.

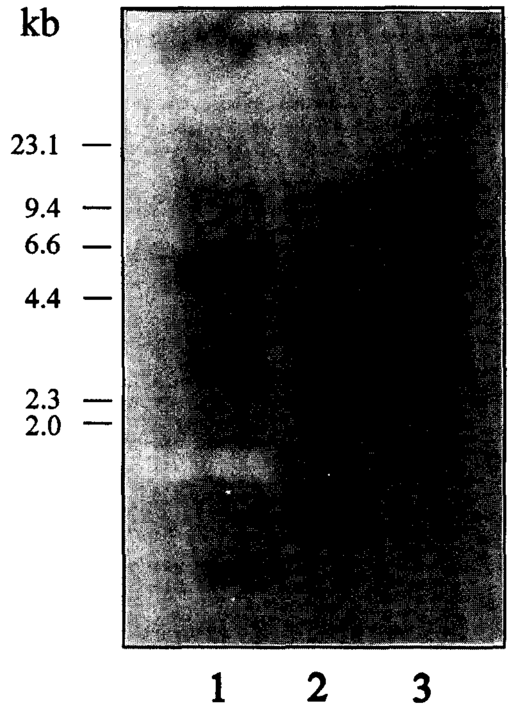

Fig. 2. Southern blot analysis of human genomic DNA using a 510 bp fragment of human preprocathepsin cDNA. Lanes 1–3 contain genomic DNA digested with *Eco*RI, *Pst*I and *Hin*dIII, respectively.

Search for protein modification sites in the predicted preprocathepsin C sequence with the PC/GENE programme revealed that there are four potential N-glycosylation sites at residues 29, 53, 119 and 276. This is in agreement with the results reported for rat liver cathepsin C [10] where it was shown that the monomeric form of cathepsin C is composed of two glycoprotein subunits with at least two glycosylation sites. Additionally, two potential tyrosine sulfatation sites were found at residues 75 and 340, four potential protein kinase C phosphorylation sites at residues 48, 138, 164 and 209, and four potential myristylation sites at residues 2, 44, 111, and 256.

```
        1............10..........20........30........40..........50...........
pap     IPEYVD----WRQ-KG--AVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNE--YSEQELLDCD-R-R
cat B   LPASFDAREQWPQ-CP--TIKEIRDQGSCGSCWAFGAVEAISDRICIHTNAHVSVEVSAEDLLTCC-GSM
cat H   YPPSVD----WRK-KG-NFVSPVKNQGACGSCWTFSTTGALESAIAIATGKMLS--LAEQQLVDCA-QDF
cat L   APRSVD----WRE-KG--YVTPVKNQGQCGSCWAFSATGALEGQMFRKTGRLIS--LSEQNLVDCS-GPQ
cat S   LPDSVD----WRE-KG--CVTEVKYQGSCGACWAFSAVGALEAQLKLKTGKLVT--LSAQNLVDCSTEKY
cat O   LPLRFD----WRD-KQ--VVTQVRNQQMCGGCWAFSVVGAVESAYAIK-GKPLED-LSVQQVIDCS---Y
cat K   APDSVD----YRK-KG--YVTPVKNQGQCGSCWAFSSVGALEGQLKKKTGKLLN--LSPQNLVDCVSE--
cat C   LPTSWD----WRNVHGINFVSPVRNQASCGSCYSFASMGMLEARIRILTNNSQTPILSPQEVVSCS-Q-Y
         *   *                   *  ** *  *                             *

        60........70..........80................90.................100..........
pap     -SYGCNGGYPWS-ALQLVAQY-GIHYRNTY-------PYEGVQ---------RYCRSREKG---------
cat B   CGDGCNGGYPAE-AWNFWTR-KGLVSGGLYESHVGCRPYSIPPCEHHVNGSRPPCTGEGDTPKCSKICEP
cat H   NNYGCQGGLPSQ-AFEYILYNKGIMGEDTY-------PYQGKD---------GYCKFQPGK---------
cat L   GNEGCNGGLMDY-AFQYVQDNGGLDSEESY-------PYEATE---------ESCKYNPKY---------
cat S   GNKGCNGGFMTT-AFQYIIDNKGIDSDASY-------PYKAMD---------QKCQYDSKY---------
cat O   NNYGCNGGSTLN-ALNWLNKMQVKLVKDSEY------PFKAQN---------GLCHYFSGS---------
cat K   -NDGCGGGYMTN-AFQYVQKNRGIDSEDAY-------PYVGQE---------ESCMYNPTG---------
cat C   -AQGCEGGFPYLIAGKY-AQDFGLV-EEACF------PYTGTD---------SPCKMKEDC---------
           ** **     *                       *                *

        ...........110........120........130.......140.......150..............
pap     PYAAKTD-----GVRQVQPYN-EGALLYSIAN-QPVSVVLEAAGKDFQLYRGGIFVGP-CG---NKV---
cat B   GYSPTYKQDKHYGYNSYSVSNSEKDIMAEIYKNGPVEGAFS-VYSDFLLYKSGVYQH--VT-GEMMG---
cat H   AIGFVK------DVANITIY-DEEAMVEAVALYNPVSFAFE-VTQDFMMYRTGIYSSTSCHKTPDKV---
cat L   SVANDT------GFVDIPKQ--EKALMKAVATVGPISVAIDAGHESFLFYKEGIYFEPDCSS-EDM----
cat S   RAATCS------KYTELPYG-REDVLKEAVANKGPVSVGVDARHPSFFLYRSGVYYEPSCTQ--NV----
cat O   HSGFSIK-----GYSAYDFSDQEDEMAKALLTFGPLVVIVDAV--SWQDYLGGIIQHH-CSSGEA-----
cat K   KAAKCR------GYREIPEGN-EKALKRAVARVGPVSVAIDASLTSFQFYSKGVVYYDESCNS-DNL----
cat C   FRYYSS---EYHYVGGFYGGCNEALMKLELVHHGPMAVAFE-VYDDFLHYKKGIYHHTGLRDPFNPFELT
                              *           *              *  *

        160.............170.......180.......190.......200.........210........
pap     DHAVAAVGYGPN------YILIKNSWGTGWGENGYIRIKRGTGNSYGVCGLYT--SSFYPVKN------
cat B   GHAIRILGWGVENG--TPYWLVANSWNTDWGDNGFFKILRGQDH----CGIESEVVAGIPRTDQYWEKI
cat H   NHAVLAVGYGEKNG--IPYWIVKNSWGPQWGMNGYFLIERGK-NM---CGLAACAS--YPIPLV-----
cat L   DHGVLVVGYGFESTNNK-YWLVKNSWGEEWGMGGYVKMAKDRRNH---CGIASAAS--YPTV-------
cat S   NHGVLVVGYGDLNG--KEYWLVKNSWGHNFGEEGYIRMARNKGNH---CGIASFPS--YPEI-------
cat O   NHAVLITG-FDKTGS-TPYWIVRNSWGSSWGVDGYAHVKMGS-NV---CGIADSVSS-IFV--------
cat K   NHAVLAVGYGIQKGN--KHWIIKNSWGENWGNKGYILMARNKNNA---CGIANLASF--PKM-------
cat C   NHAVLLVGYGTDSASGMDYWIVKNSWGTGWGENGYFRIRRGTDE----CAIESIAVAATPIPKL ----
         *                     ***    *  *              *
```

Fig. 3. Comparison of the deduced amino acid sequence of the mature region of human cathepsin C with the papain [24] and human cathepsins B [25], H [26], L [27], S [28], O [29] and K [30]. Numbering is according to papain. Gaps introduced to optimize the alignment are denoted by a dash. An asterisk (*) indicates the conserved amino acid residue in all compared sequences.

Comparison of the deduced amino acid sequences of human preprocathepsin C with its rat counterpart [5] reveals that 87.5% of amino acid residues are identical (data not shown).

Fig. 3 shows a multiple alignment of the deduced amino acid sequence of the mature region of human cathepsin C with cysteine proteinases of the papain superfamily. According to the alignment, the deduced amino acid sequence of human cathepsin C shows 33%, 32%, 37%, 33.3%, 32%, 28.5% and 33% identity with papain and human cathepsins B, H, L, S, O and K, respectively. The amino acid sequence of human cathepsin H shares the highest degree of identity with human cathepsin C. Both N- and C-terminal parts of these cysteine proteinases are conserved more than the central regions. The amino acid residues Cys^{25}, His^{159} (papain numbering), which are proposed to be important for the catalytic activity, as well as Gln^{19}, Asn^{175} and Trp^{177} [2,22], are conserved in human cathepsin C and all the sequences compared. As in rat cathepsin C, a tyrosine residue (Tyr^{26}) next to the cysteine in the active site in the human cathepsin C sequence is substituted for tryptophan residue, which is otherwise conserved among other cysteine proteinases. This substitution might affect substrate specificity. Secondary structures of some cysteine proteinases were predicted by hydropathy analysis [23] (Fig. 4). Although there is a relatively small degree of sequence identity within the amino acid sequences of cathepsins H, L, S and C, there is a striking similarity in the hydrophobicity patterns which reflects structural and functional similarities in these proteins.

The availability of a cDNA for human preprocathepsin C will enable further studies on tissue-distribution and expression of cathepsin C in normal and pathological conditions which will, at least partially, clarify the role of cathepsin C in different physiological stages and disorders.

Acknowledgements: This work was supported by a grant from the Ministry of Science and Technology of Slovenia. The authors wish to thank Prof. Roger H. Pain for his critical remarks.

References

[1] Agarwal, S.K. (1990) Biochem. Educ. 18, 67–71.
[2] Berti, P.J. and Storer, A.C. (1995) J. Mol. Biol. 246, 273–283.

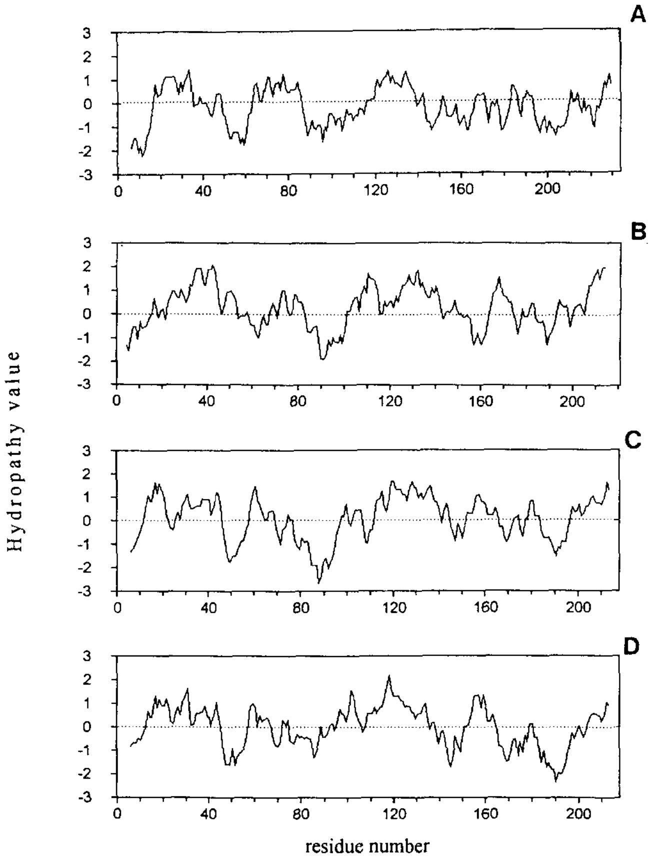

Fig. 4. Comparison of the hydropathy profiles of the deduced human cathepsin C (**A**) with the mature forms of human cathepsins H (**B**), L (**C**) and S (**D**). Hydrophobicity values were calculated using a window size of 11 amino acids and are plotted against amino acid position according to the method of Kyle and Doolittle [23].

[3] Izumya, N. and Fruton, J.S. (1956) J. Biol. Chem. 218, 59–76.
[4] McDonald, J.K., Ellis, S. and Reilly, T.J. (1966) J. Biol. Chem. 241, 1494–1501.
[5] Isidoh, K., Muno, D., Sato, N. and Kominami, E. (1991) J. Biol. Chem. 266, 16312–16317.
[6] Doughty, M.J. and Gruenstein, E.I. (1987) Biochem. Cell Biol. 65, 617–625.
[7] D'Agrosa, R.M. and Callahan, J.W. (1988) Biochem. Biophys. Res. Commun. 157, 770–775.
[8] Schlagenhauff, B., Klessen, C., Teichmann-Dorr, S., Breuninger, H. and Rassner, G. (1992) Cancer 70, 1133–1140.
[9] Thiele, D.L. and Lipsky, P.E. (1990) Proc. Natl. Acad. Sci. USA 87, 83–87.
[10] Nikawa, T., Towatari, T. and Katunuma, N. (1992) Eur. J. Biochem. 204, 381–393.
[11] McGuire, M.J. and Lipsky, P.E. and Thiele, D.L. (1992) Arch. Biochem. Biophys. 295, 280–288.
[12] Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979) Biochemistry 18, 5294–5299.
[13] Mirkes, P.E. (1985) Anal. Biochem. 148, 376–383.
[14] Aviv, H. and Leder, P. (1979) Proc. Natl. Acad. Sci. USA 69, 1409–1412.
[15] Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd Edn., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
[16] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463–5467.
[17] Kozak, M. (1987) J. Mol. Biol. 196, 947–950.
[18] Dolenc, I., Turk, B., Pungerčič, G., Ritonja, A. and Turk, V. (manuscript submitted).
[19] Qian, F., Frankfater, A., Shan, S.J. and Steiner, D.F. (1991) DNA Cell Biol. 10, 159–168.
[20] Shi, P-G., Webb, A.C., Foster, K.E, Knoll, J.H.M., Lemere, C.A., Munger, J.S. and Chapman, H.A. (1994) J. Biol. Chem. 269, 11530–11536.
[21] Shanhan, S.S., Popescu, N.C., Ray, D., Fleischmann, R., Gottesman, M.M. and Troen, B.R. (1993) J. Biol. Chem. 268, 1039–1045.
[22] Kominami, E., Isidoh, K., Muno, D. and Sato, N. (1992) Biol. Chem. Hoppe-Seyler 373, 367–373.
[23] Kyle, J. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105–132.
[24] Cohen, L.W., Coghalan, V.M. and Dihel, L.C. (1986) Gene 48, 219–227.
[25] Ritonja, A., Popovič, T., Turk, V., Wiedenmann, K. and Machleidt, W. (1985) FEBS Lett. 181, 169–172.
[26] Ritonja, A., Popovič, T., Kotnik, M., Machleidt, W. and Turk, V. (1988) FEBS Lett. 228, 341–345.
[27] Fuchs, R., Machleidt, W. and Gassen, H.G. (1988) Biol. Chem. Hoppe-Seyler 369, 469–475.
[28] Shi, G.-P., Munger, J.S., Meara, J.P., Rich, D.H., Chapman, H.A. (1992) J. Biol. Chem. 267, 7258–7262.
[29] Velasco, G., Ferrando, A.A., Puente, X.S., Sanchez, L.M. and Lopez-Otin, C. (1994) J. Biol. Chem. 269, 27136–27142.
[30] Inaoka, T., Bilbe, G., Ishibashi, O., Tezuka, K., Kumegawa, M. and Kokubo, T. (1995) Biochem. Biophys. Res. Commun. 206, 89–96.