Structure and expression of the human cystatin C gene

Magnus ABRAHAMSON,*§ Isleifur OLAFSSON,* Astridur PALSDOTTIR,† Magnus ULVSBÄCK,‡ Åke LUNDWALL,‡ Olafur JENSSON† and Anders GRUBB*

*Department of Clinical Chemistry, University Hospital, S-221 85 Lund, Sweden, †The Blood Bank, P.O. Box 1408, IS-121 Reykjavík, Iceland, and ‡Department of Clinical Chemistry, University of Lund, Malmö General Hospital, S-214 01 Malmö, Sweden

The structural organization of the gene for the human cysteine-proteinase inhibitor cystatin C was studied. Restriction-endonuclease digests of human genomic DNA hybridized with human cystatin C cDNA and genomic probes produced patterns consistent with a single cystatin C gene and, also, the presence of six closely related sequences in the human genome. A 30 kb restriction map covering the genomic region of the cystatin C gene was constructed. The positions of three polymorphic restriction sites, found at examination of digests of genomic DNA from 79 subjects, were localized in the flanking regions of the gene. The gene was cloned and the nucleotide sequence of a 7.3 kb genomic segment was determined, containing the three exons of the cystatin C structural gene as well as 1.0 kb of 5′-flanking and 2.0 kb of 3′-flanking sequences. Northern-blot experiments revealed that the cystatin C gene is expressed in every human tissue examined, including kidney, liver, pancreas, intestine, stomach, antrum, lung and placenta. The highest cystatin C expression was seen in seminal vesicles. The apparently non-tissue-specific expression of this cysteine-proteinase inhibitor gene is discussed with respect to the structure of its 5′-flanking region, which shares several features with those of housekeeping genes.

INTRODUCTION

Cystatin C is a cysteine-proteinase inhibitor with widespread distribution in human biological fluids (Abrahamson *et al.*, 1986). It is a low-molecular-mass protein consisting of 120 amino acid residues in a single polypeptide chain (Grubb & Löfberg, 1982) and belongs to Family 2 of the cystatin protein superfamily (Barrett *et al.*, 1986*a*). Cystatin C is synthesized as a preprotein with a signal peptide, indicating that the functions of the inhibitor are mainly extracellular (Abrahamson *et al.*, 1987).

The amino acid sequences for the polypeptide chains of the three other human Family 2 cystatins known, namely cystatins S, SN and SA (Isemura *et al.*, 1984, 1986, 1987), show that they are evolutionarily very closely related, with approx. 90 % identical residues in pairwise comparisons. They also display immunochemical cross-reactivity. Their 121-amino-acid-residue sequences demonstrate about 50 % identity with that of cystatin C. Cystatin S-immunoreactive cystatins are mainly found in secretions like saliva, tears and semen, and their extracellular distribution is thus more restricted than that of cystatin C (Abrahamson *et al.*, 1986).

The primary structures of four additional human cystatins are known. Cystatins A and B (Turk *et al.*, 1983; Ritonja *et al.*, 1985), both consisting of polypeptide chains of 98 amino acid residues and with a mainly intracellular distribution, are members of Family 1. The high-molecular-mass glycoproteins L- and H-kininogen (Ohkubo *et al.*, 1984; Kitamura *et al.*, 1985) constitute Family 3 of the cystatin superfamily. In human body fluids, the kininogens are mainly found in blood plasma and synovial fluid (Abrahamson *et al.*, 1986). The eight members of the cystatin superfamily in humans so far characterized are all tight-binding inhibitors of papain-like cysteine proteinases (reviewed by Barrett *et al.*, 1986*b*).

Cystatin C has a key role in the development of brain haemorrhage in patients suffering from hereditary cystatin C amyloid angiopathy (HCCAA), in which it is deposited as amyloid in the walls of the rupturing cerebral arteries (Cohen *et al.*, 1983; Grubb *et al.*, 1984). We recently described the cloning of a cDNA for human cystatin C (Abrahamson *et al.*, 1987). The cDNA detects a restriction-endonuclease-*Alu*I restriction-fragment-length polymorphism (RFLP) in DNA from HCCAA patients, but not from unaffected relatives or normal control subjects, which shows that a mutated cystatin C gene causes the disease (Palsdottir *et al.*, 1988). Using this cDNA in Southern-blot analysis of DNA from human–rodent cell hybrids, we have assigned the human cystatin C gene to chromosome 20 (Abrahamson *et al.*, 1989). In the present study we have used the cDNA as a probe in order to characterize the human cystatin C gene and study its expression in human tissues.

EXPERIMENTAL

Reagents

Oligonucleotides were synthesized with an Applied Biosystems Nucleotide Synthesizer. They were purified by PAGE in 7 M-urea, followed by phenol/chloroform extractions and ethanol precipitation. [γ-^{32}P]ATP (3000 Ci/mmol), [α-^{32}P]dCTP (3000 Ci/mmol) and [α-^{35}S]dATP (600 Ci/mmol) were obtained from Amersham International. Oligonucleotides were radiolabelled with [γ-^{32}P]ATP and phage-T$_4$ polynucleotide kinase (Boehringer) to a specific radioactivity greater than 10^8 c.p.m./μg. Double-stranded DNA probes were labelled to high specific radioactivity ($> 10^9$ c.p.m./μg) with [α-^{32}P]dCTP by using a commercial random-priming kit (Multiprime; Amersham International). Restriction endonucleases and DNA-modifying enzymes were purchased from International Biotechnologies, New Haven, CT, U.S.A., unless otherwise stated.

Abbreviations used: nt, nucleotide(s); HCCAA, hereditary cystatin C amyloid angiopathy; RFLP, restriction-fragment-length polymorphism; p.f.u., plaque-forming units.

§ To whom correspondence and reprint requests should be sent.

These sequence data will appear in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases under the accession number X52255.

Genomic Southern blotting and hybridization

Genomic DNA was isolated from human blood cells (Bell *et al.*, 1981). Digests of genomic DNA (8 μg) were separated by electrophoresis in agarose gels and transferred to nylon filters (Hybond-N; Amersham International). Hybridization to radiolabelled human cystatin C cDNA or genomic probes was accomplished as previously described (Palsdottir *et al.*, 1988). The cystatin C cDNA probe used (designated C6a) consists of 771 bp flanked by linker sequences with *Eco*RI sites. The cDNA contains coding sequence for the entire mature protein and a putative signal peptide, 74 bp of 5′ non-coding sequence and 259 bp of 3′ non-coding sequence, including a polyadenylation signal 11 bp from the 3′ end (Abrahamson *et al.*, 1987).

Screening of a human genomic library

A human genomic library in λ Charon 4A phage vector was kindly given by Dr. R. Wydro, Integrated Genetics, Framingham, MA, U.S.A. Approx. 5×10^5 plaque-forming units (p.f.u.) of the genomic library (with *Escherichia coli* 803 as host) were screened for clones carrying the cystatin C gene by plaque hybridization with the radiolabelled full-length human cystatin C cDNA as probe. Hybridizing plaque were purified, and phage DNA was isolated from plate lysates by standard procedures (Maniatis *et al.*, 1982). Digests of isolated DNA from the clones were analysed with respect to hybridization to radiolabelled cDNA and to oligonucleotides oC1 and oC2 (see Fig. 2 below), corresponding to the 5′ and 3′ ends of the cDNA respectively. Restriction fragments of phage clone inserts were subcloned in pUC18 and amplified using *E. coli* JM109 as host.

Polymerase chain reaction

A sample of extracted λ-phages from a plate lysate of clone λC3 (10^5 p.f.u./μl) was dialysed overnight against distilled water. A 50 μl portion of the dialysed sample was directly used as source of template DNA for enzymic amplification in a 100 μl reaction mixture containing 2.5 units of *Taq* polymerase (Cetus Corp.) and oligonucleotides Cg5 and CIV (see Fig. 2 below), both at 1 μM final concentration, as primers. In all, 30 cycles of amplification (each cycle: denaturation, 92 °C for 1 min; primer annealing, 55 °C for 1 min; extension, 72 °C for 1 min) were accomplished by using a programmable heating block (Hybaid). The major amplification product was isolated by preparative agarose-gel electrophoresis, phosphorylated using phage-T_4 polynucleotide kinase (Boehringer) and ligated in the *Sma*I site of M13mp18.

Nucleotide sequence analysis

The 'shotgun' approach (Bankier & Barrell, 1983) was used to generate a random selection of templates for sequencing of large gene fragments. Nucleotide sequences were determined by dideoxy sequencing with [α-^{35}S]dATP as labelled deoxynucleotide (Biggin *et al.*, 1983) on subclones in M13mp18 or M13mp19 vectors (Pharmacia), using the M13 universal primer and the Klenow fragment of *E. coli* DNA polymerase I (Boehringer) or a modified phage-T_7 DNA polymerase (Sequenase; United States Biochemical Corp.). The sequence of GC-rich regions resulting in compressed gel bands were solved by substituting ITP for GTP in the sequencing reactions, as described by the manufacturer of the Sequenase kit.

For a given large DNA fragment, sequence data from randomly generated templates was collected until approx. 85% of the nucleotide sequence of the fragment was determined on both strands. In order to obtain complete double-stranded sequence for the fragment, two different strategies were followed. In the first of these, 17-mer oligonucleotides corresponding to sequences approx. 50 bp upstream of the sequences to be determined, for subsequent use as primers in sequencing reactions with appropriate random subclones in M13, selected from previous sequence data, were synthesized as templates. Alternatively, inserts of previously sequenced M13mp18 subclones were cleaved out with *Eco*RI/*Hin*dIII, isolated and reintroduced in M13mp19, thereby allowing sequencing of the other strands of the inserts.

Nucleotide sequences were aligned and analysed by computer programs described by Staden (1982*a*,*b*).

Northern-blot analysis

A human placenta was obtained immediately after delivery. Other human tissue specimens were collected at autopsy, within 15 h after death. Total cellular RNA from the tissue samples was isolated according to Chomeczynski & Sacchi (1987). RNA samples of approx. 10 μg were electrophoresed in 1%-agarose/formamide denaturing gels and subsequently blotted on to Hybond-N nylon filters (Amersham International). Prehybridization and hybridization to radiolabelled cDNA (5×10^6 c.p.m./ml) in 50% formamide was performed at 42 °C. Filters were washed at a final high stringency in $0.2 \times$SSC ($1 \times$SSC = 150 mM-NaCl/15 mM-sodium citrate, pH 7.0)/0.5% SDS at 68 °C for 30 min.

RESULTS AND DISCUSSION

Mapping and cloning of the human cystatin C gene

Single and double digests of genomic DNA with 12 restriction enzymes were analysed by hybridization after Southern blotting, at first with cystatin C cDNA and, after cloning of the cystatin C gene (below), with genomic fragments C3A-C (see the legend to Fig. 1) as probes. The restriction-digest data were consistent with a single cystatin C gene and allowed the construction of a 30 kb map of the genomic region covering the gene (Fig. 1*a*). Using the same hybridization probes, we have previously found three polymorphic restriction sites on examination of DNA from 79 (34 unrelated) individuals (Palsdottir *et al.*, 1988). Two of these, for *Pst*I and *Sac*I, were localized to positions within the 30 kb map, both downstream from the structural gene (Fig. 1). The third polymorphic restriction site, for *Eco*RI, was mapped to a position 18 kb downstream from the central *Eco*RI site in Fig. 1(*a*).

Five out of approx. 5×10^5 λ-phage clones in a human genomic library exhibited strong hybridization signals to a full-length human cystatin C cDNA probe (Abrahamson *et al.*, 1987). One of the clones contained restriction fragments that hybridized both to an oligonucleotide probe corresponding to a sequence in the 5′ end of the cystatin C cDNA and one in the 3′ end (Fig. 2). The insert of this clone, designated λC3, was cut out with *Eco*RI. Due to an internal *Eco*RI site this resulted in two fragments of 5.0 kb (C3E1) and 10.5 kb (C3E2). C3E1 and C3E2 were subcloned in pUC18 and analysed by restriction-endonuclease cleavage and Southern-blot analyses. The partial maps of the clones were in full agreement with the genomic map of the cystatin C gene (Fig. 1). A 3.3 kb *Xba*I–*Eco*RI fragment of C3E1 and a 4.0 kb *Eco*RI–*Bam*HI fragment of C3E2 were sequenced (Fig. 1*b*). A DNA fragment overlapping the fragments C3E1 and C3E2 was obtained by oligonucleotide-directed enzymic amplification using olignucleotides corresponding to sequences in the 3′ end of C3E1 and in the 5′ end of C3E2 (respectively) as primers (Fig. 2) and DNA from phage clone λC3 as template. Both strands of the 421 bp amplification product were sequenced, showing that no additional *Eco*RI fragment was present between the gene fragments C3E1 and C3E2. Thereby was the entire nucleotide sequence of a 7.3 kb genomic segment containing the

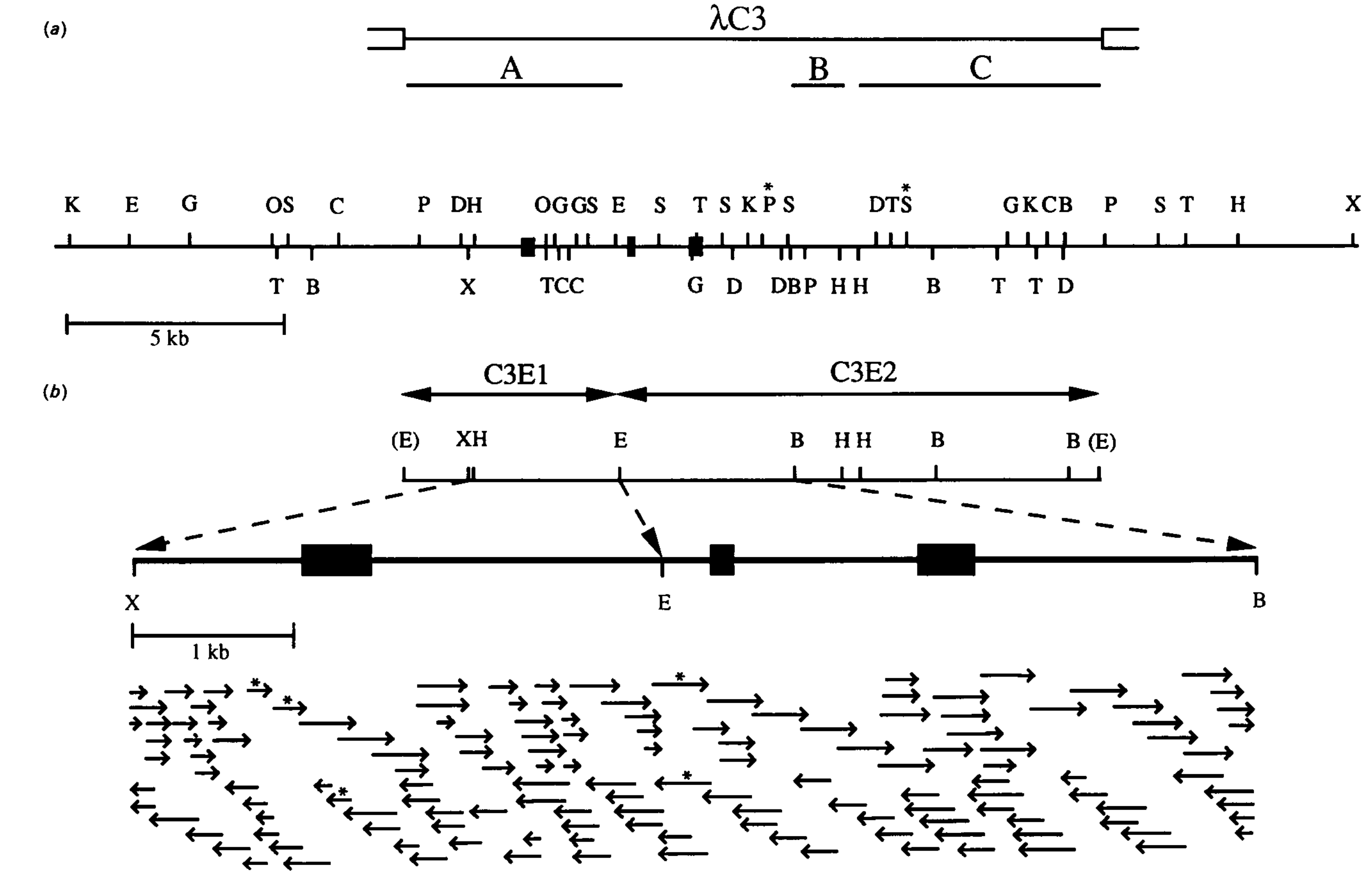
(a)
λC3
A
B
C
K E G OS C P DH OGGS E S T S K P S DTS GKCB P S T H X
*
*
T B X TCC G D DBP HH B T T D
5 kb
(b)
C3E1
C3E2
(E) XH E B HH B B (E)
X
E
B
1 kb
*

Fig. 1. Structure of the human cystatin C gene

Exons are depicted by solid boxes. (*a*) Restriction map of a 30 kb region around the cystatin C gene, mapped using genomic DNA and 12 restriction enzymes. The probes used to construct the map were the human cystatin C cDNA, C6a (Abrahamson *et al.*, 1987) and three genomic fragments from the genomic library clone λC3: A, 5 kb *Eco*RI fragment, B, 1.2 kb *Hin*dIII–*Bam*HI fragment and C, 6 kb *Hin*dIII–*Eco*RI fragment. Restriction enzymes: K, *Kpn*I; E, *Eco*RI; G, *Bgl*II; O, *Xho*I; T, *Taq*I; S, *Sac*I; B, *Bam*HI; C, *Hin*cII; P, *Pst*I; D, *Dra*I; X, *Xba*I; H, *Hin*dIII. *Pst*I sites are only shown in the flanking DNA, the number of sites within the gene being too many to depict. Two polymorphic restriction sites within the mapped region are indicated (*). (*b*) Sequencing strategy. A partial map of the insert of the genomic clone, λC3, is shown. The insert was cut out with *Eco*RI and the resulting two fragments of 5.0 kb (C3E1) and 10.5 kb (C3E2) were subcloned in pUC18. A 3.3 kb *Xba*I–*Eco*RI fragment of C3E1 and a 4.0 kb *Eco*RI–*Bam*HI fragment of C3E2 were sequenced. The arrows indicate direction and length of gel readings from different M13 subclones. The arrows marked with an asterisk represent sequences determined using oligonucleotides corresponding to parts of the cystatin C gene sequence as primers, either in sequencing reactions or for enzymic amplification.

cystatin C gene determined on both strands, with each nucleotide determined on average 5.5 times.

Structural analysis of the human cystatin C gene

Comparison of the cystatin C gene sequence with that of the human cystatin C cDNA (Abrahamson *et al.*, 1987) showed that the human cystatin C gene is composed of three exons. The gene has two intron sequences of 2252 and 1254 bp respectively, resulting in an overall gene size of approx. 4.3 kb (Fig. 2). The introns are positioned between the nucleotide triplets encoding amino acid residues 55–56 and 93–94 of the mature protein (Grubb & Löfberg, 1982) respectively. All intron–exon junctions are close matches to the consensus sequences proposed for the donor and acceptor sites of introns (Mount, 1982). The positions of the introns are exactly identical with the ones in the cystatin SN and cystatin SA genes (Saitoh *et al.*, 1987). However, the first intron in the cystatin C gene is considerably larger than the first introns in the cystatin SN and cystatin SA genes (1511 bp for the cystatin SN gene and approx. 1200 bp for the cystatin SA gene). The introns in the three sequenced Family 2 cystatin genes occupy positions similar to those occupied by the introns within the three repeats in the upstream part of the human kininogen gene, which encode the three Family 2 cystatin-like *N*-terminal domains of L- and H-kininogen (Kitamura *et al.*, 1985).

The only deviation seen when the exon parts of the cystatin C gene is compared with the cDNA is a guanosine residue at position 1072 of the gene sequence (Fig. 2), which is not found in the cDNA. We re-examined the corresponding part of the cDNA sequence after carrying out sequencing reactions on both strands with ITP substituted for GTP, which normally solves sequences in GC-rich regions, but failed to demonstrate an additional guanosine residue at this position. The observed deviation is therefore probably due to an artefact at cDNA synthesis. Alternatively, it could represent a genetic variation. During the final preparation of this manuscript, a report describing the sequences of the exon and exon flanking parts of the human cystatin C gene was published (Levy *et al.*, 1989). Examination of the sequences of those cystatin C gene segments revealed 12 smaller differences compared with the sequence given here (Table 1). Levy *et al.* (1989) give no details as to the amount of sequencing supporting their published sequences, but in all these regions our sequence has been unambiguously determined

```
TCTAGACCTTGACCAAATGCTCACGGATACTCAGGACCTCCGAAGACTCCGGAAGGCCGAGACCTCTATCTTTCCCTTTTCTGAGGAAATGAAGCTAGGG  100
Xba I
GGTCGTATGCTTTTAGGGCCCAGGGGGCCTTGACACCTCTCTGGAGGGCAGGGTGACCTAAGCTTGGGAGGAGGGACAGGTCTTTGGGAAGAGATGGAGC  200
                                     EC
TGAGGGGGTGGAGTGGGTAGAAGGGGGGCATAGGTGTGGAAGGGGTGGGGAAGAGGGGCCAGGAGTGGGGTGAGAAGAGGGGAAAGAAGAGGATAGGAG  300
                                                 GC box
GGACAGGGACAGGGAGAGGCCAGGAATGGGTAGGAAGGGAAGACAGAGGAGGGGCGGACGAGGGAGGGGCGATGAAGAGGGGCCGAGCTGGGGTGGGGG  400

CGAAGTGGGAGCAAGGACAGGAGTGGAGGAGGGAGATGAGGGGCATGGAGGCGGTGTCTGGGGGTTGAAAGGGGAAAGGGACAGAGGAGAAGGAGCCTG  500
                                                                              ┬------2------┐
AAGAGTGGCGGCATGGAGGGGCCCGAAGTGGGGAGGTGGGAATGGTCGGATGGATGGGGAGGAATGGGGAGGATAGGAGAGGCAGGGGAGGCTGGAAGAC  600
                                                                           └-------1-------┘
AGTAGGAGAGGAGGGGGAGGGGAGGGGAGGGGATGGATGGGAATGGATAGGGAAGGCCAGGGAGACTGGGGGGCCAGGGGAGGAGGGGAGGGAGAGGGGA  700
                                                                               ┌------2------┐
GGCGGGAAGGGAGAGGAGAGGCCGGGAGGGGCTGGGAGGGGAGGGAAGGGGATGGATGGGGAAGGACAGGGAAGCCTGGAGGGGCAGGGGAGGCTGGGAT  800  71% (2/6)
                             ┌====3a=======---b---┬====3a====┬---b---=====3a====┐
GGGAGGAGACTGATAGGGAGGGACCTGGAGGCGGGGAGAGGCAGGGGAGGCGGGGAGAGGCAGGGGAGGCGGGGAGGGAGAAGGAGGTGGAGGGACGA  900  72% (4/5)
                                   └-------1-------┘ └-------1-------┘
GGCGTTCCGGAAAAGGGAGTGCAGGCCGCGGTGGGGTGGGGCGGCGAAGGCCGGAAGGGATAAAACCGCAGTCGCCGGCCTCGCGGGGCTCACGGCCTCG  1000  73% (14/14)
                                        GC box
      |-->cDNA                                                                    -26
      | »»»»»»»»oC1»»»»»»»»                                                      MetAlaGlyProLeuA
CCTCGGTATCGCAGCGGGTCCTCTCTATCTAGCTCCAGCCTCTCGCCTGCGCCCCACTCCCCGCGTCCCGCGTCCTAGCCGACCATGGCCGGCCCCTGC  1100  72% (12/13)
                                -10                          1                          10
rgAlaProLeuLeuLeuLeuAlaIleLeuAlaValAlaLeuAlaValSerProAlaAlaGlySerSerProGlyLysProProArgLeuValGlyGlyPr
GCGCCCCGCTGCTCCTGCTGGCCATCCTGGCCGTGGCCCTGGCCGTGAGCCCCGCGGCCGGCTCCAGTCCCGGCAAGCCGCCGCGCCTGGTGGAGGCCC  1200  81% (11/19)
                     20                            30                            40
oMetAspAlaSerValGluGluGluGlyValArgArgAlaLeuAspPheAlaValGlyGluTyrAsnLysAlaSerAsnAspMetTyrHisSerArgAla
CATGGACGCCAGCGTGGAGGAGGAGGGTGTGCGGCGTGCACTGGACTTTGCCGTCGGCGAGTACAACAAAGCCAGCAACGACATGTACCACAGCCGCGCG  1300  64% (11/13)
         50                55
LeuGlnValValArgAlaArgLysGln
CTGCAGGTGGTGCGCGCCCGCAAGCAGGTGCGTGCCGCCCCCCGCAGGGTCCGAAGCCCCGGCCCCGCCGTCCCAGCCTCCCCCCGCGCTGCTCCCGGAC  1400  81% (13/17)

CCCGTGCTGCTCCTCTCCGGCGCCTGGGCTTCTCACCCGGACCCTGTTCCCGGTCTTCGCTGTCACCCCCGAGCCCTTGGCGGGCGTGTCCGGGAATGCC  1500  72% (10/10)

CTGAGTCTGGCCTGGCCTGGAACCGCACGGACACGTCAGGTCCGCGCCCGGCGCCTAAGATCAGCTTCCAGGAGCCAGCTCGAGCTGCGCCGCAGCGCGG  1600  71% (12/16)

GGGCAGGCACAGGACGTCCCACACAGCTCTGTCCCGACCTCCTGGGACGCTGGGCTCCGGGTGCGCTGCTGAATACGCAGGAGAAGGAAAAACAGAGCCC  1700

CTCATCTGGCTGCCTCCTTGTTCTTCCACCAAATAATTTTTAAGCACTTTTGTTTCTTTTAACTTTTTATTCTCCCGTAATTTCAGACTTACAGAGCAGC  1800

TGCAAAATAGCACAGGGCATCCCTGTAGATCTTTCACCCAGATCCCTCAGATGTTAGCACAGATGTTAACATTTTGCCCTGTTTGTTCCCCCCAGCTCAC  1900

TTTCCGAATGTGAATATGGGTGTGTGTATATGTACTGTATATGCATTTTTTGCCTGAATCTTTTGAAAGTAAACACTTCAGTGTGTTTTTCCTAAAAACA  2000

AGGGCAATATCGCACATAATCACTATTCCATTTTGAAAATCAGCACTGACACATTCATGATCTAATGCTCAGACTCATTCAGCTTTCCCTGTCTGTTGAC  2100

GGCCTTCGTGCCCTAGGACGCTCCTGGGCCATGAGTGCATGAGTTCAGCCCTGTCCTGTCCCCTCGGCCTCTTTTAACCTGCAGCAGCAGCTGGGTCTGT  2200

GCCACGACCGTGCCATTTCCCAGGGTCTAGCAGGTGTGGATGGAGACTATGCTGACTCTGGGTGGGCTTGATGCTGCTCAAGATGAGATCTAGGCCATGC  2300

GGCTCATCCTCCTCCCAGAATACTCTCCGCAGGGGCCACACGTGAGCCTGGCACCTTGTCCTGCAGAGCCCTGCTTCCCTCCCCAAGTCACACCCCTGGG  2400

CACAGTCCCCTAGGACTAGCGGCCTTCACCCTCAGGCCGGCTGACCACCCCCTACATCCCAGGGCAGCTGAGTCCCTGCTGGGGTGGAGCATGCCTGACC  2500

CTGCCTCTATCAGCTGATGCAGAGTTAGACCTCAGCCAGATGAAGACACCCAGCAGACCGGAGTAGGGGGTCGGATCGGGAGGGAGCTTCAGTAGGGCTA  2600

CCAGGCCCAGCTTGACCTGCATCCCATGGCAGAGCAGCGAACAGTGACACAGACTTTAGAGCTCCTCCACCTTCTCTTGGAAATTCAGAGGAGTCCAGAC  2700

CAGCCGCGTTTCTCCCGCAGCCGTCAGCTGGGGCCCTCTCCCTGCACAGGAGGGCCATTCCCTGGTGTAGGGTCCCCGCTGGCCTGCACCTCTCTCTCAC  2800

TGGGAGTGAAGCATGGGGCATTGTAGATCGTTGGGCCCTGGAGCCTATTTTACAGAGGAGCAGACTGACACCAGAGGGATCACAGGCCTTGCTTGTGCTT  2900

CTACAGGACTTGTGTGTGGTTCCACAGGGCAAGGTCTAGCACCCTGGTCCCAGGGTCCCTCATCCCATGCTTCTCCACAGTTCTGACAAGTCATGTTTTG  3000

GGGCGGCACTGTGCAGGGAAAGCATTCAGTTCTCTTCTGAAGTTGCAACCCTAAGACATGCAGGTGTGTGACTCACTTTAGAAATATTGCCTTGAAAATC  3100

ACACCTGGAATGGAAGCACGTGGGAAGCAATGTTTATTGGCCTAAAACATCAATGTATGTGAGCATCTCATCTCCTAGTGAGAAATGAGGAAAAATACCT  3200
                                   »»»»»»»»Cg5»»»»»»»»
CTGGGTTAAATGGCAGGAATGAGATGCTCTGTGGACTGAATGCCAGGAACTGGAAGTTTGCCGAAATTTCATCATCACATGAGAACCTTCCTAGAATAGA  3300

TCCAGTGTCCCTGCCCCCTGGGTCATAGGTAGCGGAATTCAGTTAATCCTTGGCATTGGCATAGAGAAACAAGTTACTGGGGAGGCCTGGGGCATGGCAT  3400
                                  EcoR I
CCCTGCCAGCTGGCAGGAGGAGGTGGCTTGTGTGCCTTGCAGGTGACAATGTGGGCAGCTCATGAAGGTAGGCTTGAAGCCCCAGGCAAGCCCAGTGACC  3500
```

```
                                                                               56          60
         .         .         .         .         .         .         .         IleValAlaGlyValAsnTyr
CGGTCACAGTGAAGTGCCTGTGTGTGTAAGAAACTGACAGAACGTGCTGTCCCTGCCTCCTGCTCTTTCACATGTGTAGATCGTAGCTGGGGTGAACTAC  3600
                     70                            80                            90       93
PheLeuAspValGluLeuGlyArgThrThrCysThrLysThrGlnProAsnLeuAspAsnCysProPheHisAspGlnProHisLeuLysArg      .
TTCTTGGACGTGGAGCTGGGCCGAACCACGTGTACCAAGACCCAGCCCAACTTGGACAACTGCCCCTTCCATGACCAGCCACATCTGAAAAGGGTATGTG  3700
                                  «««««««««CIV«««««««««

         .         .         .         .         .         .         .         .         .         .
CCTTATATGGGTCCAGGGCCAGTCATACACTGCAGAGGGGTGTGTGTGTGTGTGTGTGTGTGATGCACATGTTCTGCAGGGTACGTGTGCATGTGCCT  3800
         .         .         .         .         .         .         .         .         .         .
GAGTGTGTGTACACGTGGAGATGCATATGTGTTTCCAAGTATAGGTTTGTGTGGGGAAGTGCACGTGAGTGTGTGCAAGTGGGTGTGTGGATGTGTTGGG  3900
         .         .         .         .         .         .         .         .         .         .
GAAGTGTGGGGGTTTATGCATGGATAGGTGTGGGTGTGTGAGTATATGTGTGTGCACATATGTGTGGGGGTGTCTGTGCACATACATTCACATACATGGG  4000
         .         .         .         .         .         .         .         .         .         .
GGTTTCTGTGAACATACATGCTGTATTATGAGGAATTACCTGCCCTGTGTGTGTGTGTGCATGTGGGTAGATAGATGTATGAGGGAGTATGTGGATTCAT  4100
         .         .         .         .         .         .         .         .         .         .
GCATAGATCCGTATGGGTGAAGGTTAGGGTGAGTTCATGTAGATGCCTATGTGTGTGCATGTAGACGGGGTGGTGTGGAGAGGAGTGATGAATTTGATTT  4200
         .         .         .         .         .         .         .         .         .         .
GCTAAGAGGGCTTTAGCTTGGGATTGGGGTACTGGGAGCTCCACCCTATGTGCTTTGGAGTGTTGCCTACTGGACTGCAGGAAGCAGCTGCAGGGCTGGG  4300
         .         .         .         .         .         .         .         .         .         .
TGCTGGGCAGGGAGAAAGGGCTCTGTCTAATCCCAGCCTTAGGCACCTGCCCACAGCCACGGCCATGCTGAGCAGATTAGAGGGTAGTAGAGGCCTGTTA  4400
         .         .         .         .         .         .         .         .         .         .
GCAGCCAAACCCTCAGACCTGCCCCAGCTCACCCAACACCAGCTTCTCCAAGGACCCAGGATTTCTTGTGACGTCTCTGCTCAGGGGAGAGCCACACTCT  4500
         .         .         .         .         .         .         .         .         .         .
CCTTGTCATCTCCTCACCACCCCATGTGCTATGACTTGGAATTTCCAGTTCCCTGGGTTCTTCCCTCTCGCCCTTCATAGTGCTGGCCTGAGGGCTGGAG  4600
         .         .         .         .         .         .         .         .         .         .
GTGGAAGGAGCTGGGGGCAGTGAGCTGCCTCCCCCTGCCCTGTACCCTTAGGGCTCCCGAGGCCTTGCACAGGCTGCTCCTCACAGGGCTGTGCTGGGGC  4700
         .         .         .         .         .         .         .         .         .         .
AGGAATCCTGCAGGCTGGGGTGGGGGCCCAGTGCCACCAGGTAGAGTTGGAGCCCTGGGAGAGGGATGGAGCAGTCACTGCCCAGTCCAGCGGTGCTCTG  4800
         .         .         .         .         .         .         .         .         .         .
GGATCAGCGGGGGTGGGTGGAGGGGTAGATAAGGGCCCAGTGTTTCACCTCCATCCTTCTTCACTCCCACTCTGATGTCCCGTGCCTGGGCATCTTCTGC  4900
                                               94                100                           110
         .         .         .         .       LysAlaPheCysSerPheGlnIleTyrAlaValProTrpGlnGlyThrMetTh
TTTGAACTGTAACCCACACTGATTGTCCCCTCTGTTTCCTTTCACAGAAAGCATTCTGCTCTTTCCAGATCTACGCTGTGCCTTGGCAGGGCACAATGAC  5000
                         120
rLeuSerLysSerThrCysGlnAspAla***        .         .         .         .         .         .         .
CTTGTCGAAATCCACCTGTCAGGACGCCTAGGGGTCTGTACCGGGCTGGCCTGTGCCTATCACCTCTTATGCACACCTCCCACCCCCTGTATTCCCACCC  5100
         .         .         .         .         .         .         .         .         .         .
CTGGACTGGTGGCCCCTGCCTTGGGGAAGGTCTCCCCATGTGCCTGCACCAGGAGACAGACAGAGAAGGCAGCAGGCGGCCTTTGTTGCTCAGCAAGGGG  5200
         .         .         .         .      »»»»»»»oC2»»»»»»»      .        cDNA—>|    .         .
CTCTGCCCTCCCTCCTTCCTTCTTGCTTCTCATAGCCCCGGTGTGCGGTGCATACACCCCCACCTCCTGCAATAAAATAGTAGCATCGGCTCCCTCTGAG  5300
         .         .         .         .         .         .         .         .         .         .
TTCTTGGCTGTCTGGGGATGTGCACACAGGCAGGGTTTCCGCAGTTCCTTTATGAAGCCTCCTTGTCCTGCTGGTGTGAAGATGAGAGGAGTACCTGGGA  5400
         .         .         .         .         .         .         .         .         .         .
GCTGACGCGGCCACAGCAAGGCCATCAGGCAAGCTGCTGCTATAGGAGTCCTGAGTCTCAGAGCAGGGAGAGCAGCCAGGGGCTGGGAAACAAGGCGTTG  5500
         .         .         .         .         .         .         .         .         .         .
AGCCACAGCAGCCCCTGGTGAGGGGGTCAGGGAGAGGAGGGGCCCAAGCTGCTGCCCCCAGAAGCTGGGCTGGTGATTTGGTCTGAGCTGCTGGTCAGAC  5600
         .         .         .         .         .         .         .         .         .         .
CAGGAGGAGGGAGCTGGCTGTGTCCATAGACAGGGGCCAGGCCTCGGTGGAGCTCGCCAGGTCATAGATTCCATCTGTGCTTGCAGAGTTGAGCAGACCC  5700
         .         .         .         .         .         .         .         .         .         .
CAGGGACTGAGCTGCTCTCATCAAATCCCCTAGACACTAAATAGTGATCAATGCGTGCTTCTTTCCAAGAGACTTAGGCCCCTACGTGGAAACAAAGTCC  5800
         .         .         .         .         .         .         .         .         .         .
AAAAAGCCTGTCTGTGTAACTGAGCAGGAGAAATGCCAGCCAGACCAGAGACCAGCAAGATCACAGAGAAAGTGGCACCTGCTGTCATCCATCCTGGGAG  5900
         .         .         .         .         .         .         .         .         .         .
GCTGTGTCCTCCCAGAGTTTTACAGGCTCATAGCCACCAGCCCTTTCTCTGGACTTGGATTTAAAACAAGTTGTTAGTCCAGCTTCATGGTAGGGATGAA  6000
         .         .         .         .         .         .         .         .         .         .
GTTAAAGCACAAATGAGACCGACAATGCAGTGGACTCTTCTTTTTCCAAGGAGGATTAAGAGAGAACATTCCCTTCATGTCAGCATTATCCTATAGAATA  6100
         .         .         .         .         .         .         .         .         .         .
GTGGATCTTTGTGCCACAAGCAGGAGGCTACAGTCCTAAGACACAGCCCACCCTGCACAGTCCTTCCCTTTGCCATTTAGGACAGGGGTGCAGCCCTTA  6200
         .         .         .         .         .         .         .         .         .         .
CAGCCAGTGCTGCTGGGGCAGGCCTCCCAGGTACCTTCCAACCTCGGAGCTGCACTGACCCCCTCTTTTGGCTGCTGCCATGAACACTGAAGGTGAAATG  6300
         .         .         .         .         .         .         .         .         .         .
ATGAAGGAAAATCCACCCCAGTAGCGCCAAGGCCAGCACCCCACATGCTGTCCCTGCAAAGTCAGAGCAGGGGACACGGCAGGGAGAGTGGGGAGCGTCT  6400
         .         .         .         .         .         .         .         .         .         .
GCTTCCTTTTCCTCTCCACCCTGACTTGCTTTAACAGGGTCTTTGGGGTGAGGGAAGTGTAGCACCCCCTAGGATGCCATGAGCTGAGGAGATTTGAGTC  6500
         .         .         .         .         .         .         .         .         .         .
TGGGCCTCACCCTGCTGGGGGTGGGTGGTCTTAGTCACTCAGTTTTCTGTGGGGGGTTGGGGGATAGAAGTGGTGGTTGTGGTCAGTGGTAAAGGGTCCC  6600
         .         .         .         .         .         .         .         .         .         .
TGCAGGGAGGTGCTCCTGGTTTTCTACCTCGGTGAAGGATCTTTGACACAGGCAACCTCCATCCAGGCACCCAGTGGAAACGGACAGCTGACCATATAGC  6700
         .         .         .         .         .         .         .         .         .         .
CCAGTGCCTTTTCTGACACACTTCTTCCTTGGGGAAGTTGAGGTTGGGGGCGGGCTGTGCACAGGGCTCCCCAGCCTCCCTGTGTGTCTGCTTCCTTCCC  6800
         .         .         .         .         .         .         .         .         .         .
TGCCCTCCGCCCGGTGGCCCCATGGCCCACAGCCCACAGGGAAGCAAGTTGGTGCCACCCCTGCCTCCCCTCCGCAGCCGGCGGGCCTGTAAATGCCTCC  6900
         .         .         .         .         .         .         .         .         .         .
CTTGGGCCCCCATGAGGCCAGTGAGCCAGAAGCCCCTCTGACCCAGAGGTCACTGTGACCAGCACATCCTGGAGAGGAGGCCTCGCCCTTGCACCTTTCA  7000
```

```
         .         .         .         .         .         .         .         .         .         .
TCTGCAAGAGTTACTTTTACATTTATTTTAAATGTTGATTTAAGTATATACTTCATCTACAGTTATCCATATAAATAGATTCATTACAGGGCCTGGCTCT  7100
         .         .         .         .         .         .         .         .         .         .
GAATAAACACAGTAAGTTGGAGAGGCTTTTTCATTGATCTGGTTCTACACTGGTGGCTTTATAACATAGTGCTTTGACACCTCTTGCATTGTGTTGCTGG  7200
         .         .         .         .         .         .         .         .         .         .
TGACCATTGGGGTGGATCAGGAAGGGCCCAGTAGCTTCTGAGCTCCCTAGCATTCTTGTCCTGAAGGAAGCCAGCCTTTGCAGAAGGATCC          7292
                                                                                      BamH I
```

Fig. 2. Complete nucleotide sequence of the human cystatin C structural gene and flanking regions

The deduced amino acid sequence is shown above the coding parts of the three exons, numbered according to the amino acid sequence of native cystatin C (Grubb & Löfberg, 1982), with negative numbers for residues in the putative signal peptide. The restriction sites at the ends of the two sequenced segments of the gene and the 5′ and 3′ ends of the cystatin C cDNA (Abrahamson *et al.*, 1987) are marked. The positions and directions of oligonucleotides used as primers for enzymic amplification (Cg5 and CIV) or as hybridization probes (oC1 and oC2) are indicated. A putative TATA-box in the 5′-flanking region and a polyadenylation signal in the 3′-flanking region are underlined. The core sequences of two binding sites for transcription factor Sp1 (GC box) and a proposed enhancer core-like sequence (EC) are marked. Three long repeats in the 5′-flanking region (1, 2 and 3a-b-a) and an inverted repeat sequence in the first exon (arrows) are indicated. A 900 bp GC-rich region in the 5′ part of the gene is depicted by G+C percentage values and the ratio CpG/GpC dinucleotides (in parentheses) to the right of the actual sequence lines.

Table 1. Ambiguous sequence positions in the human cystatin C gene

Differences between the sequences of Fig. 2 and that published by Levy *et al.* (1989) are listed. The nucleotide numbering refers to Fig. 2.

Position	The present paper	Levy *et al.* (1989)	Comments on the Fig. 2 sequence*
728	G	–	DS, compr
928	G	–	DS, compr
1347–1349	GGG	GG–	DS, compr
1474–1476	CCC	CC–	DS, compr
1544–1546	GCG	–	DS, compr
3393	C	–	DS (4)
3416–3417	GG	G–	DS (4)
3462–3463	AT	TA	DS (4)
4713–4714	GG	G–	DS (5)
5141	T	TT	DS (8)
5206–5208	CCC	C– –	DS, compr, CCC in cDNA
5511	G	GA	DS, compr

* Abbreviations: DS, the sequence was determined on both strands (number of independent gel readings in parentheses); compr, the region is GC-rich, resulting in compressed gel bands; the sequence was solved by substituting ITP for GTP in the sequencing reactions.

on both strands. Another major difference is seen in the 3′ end of the segment covering exon 2 in Levy *et al.* (1989). The last 99 nt of this segment, which should correspond to nt 3736–3834 in our sequence, deviates totally from it. Examination of this part of the Levy *et al.* (1989) sequence revealed that it is very similar to a part of the phage M13 sequence and must represent an overlooked read-through at sequencing.

Flanking regions of the cystatin C gene

The first and last positions of the 771 bp sequence of the cystatin C cDNA are indicated in Fig. 2. The cystatin C gene sequence shows that the polyadenylation signal, AATAAA, present 11 nucleotides upstream of the 3′ end of the cDNA (Fig. 2, nt 5271–5276) is the only such signal present in the 2 kb sequence downstream of the structural gene. Addition of a normal-sized 100–200 bp polyadenylated tail at an mRNA position corresponding to the 3′ end of the cDNA would thus result in a cystatin C mRNA with a size of approx. 900 nt, provided that the cDNA is full-length. Northern-blot analysis of RNA from human placenta with the cDNA as probe revealed the presence of a single mRNA species of 800–900 nt (Fig. 3*a*), i.e. it confirms that the cDNA is essentially full-length and hence indicates that the 5′ end of the cDNA is at a position close to the start site for transcription of the cystatin C gene.

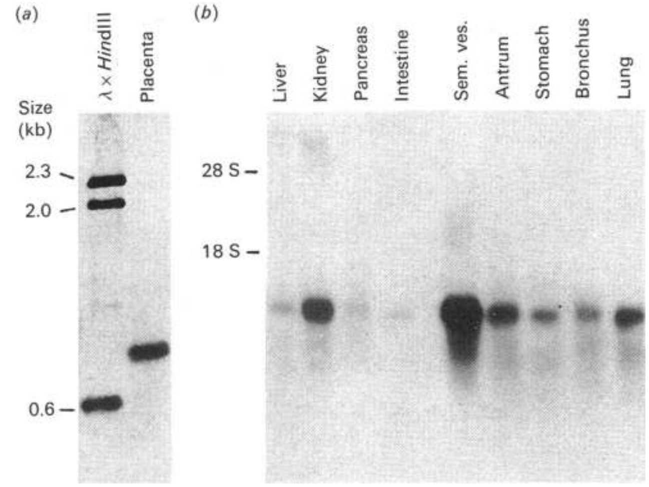

Fig. 3. Northern-blot analysis of RNA from human tissues

Samples of total RNA were hybridized to radiolabelled cystatin cDNA after electrophoretic separation and blotting on to nylon filters. (*a*) A 50 μg portion of RNA isolated from a placenta, taken immediately after delivery. *Hin*dIII-cleaved λ-phage DNA, end-labelled using [α-^{35}S]dATP and the Klenow fragment of *E. coli* DNA polymerase I, served as size marker ('λ × *Hin*dIII', left). (*b*) RNA samples (10 μg) from human post-mortem tissues (Sem. ves., seminal vesicle), isolated within 15 h after death. The positions of 28 S- and 18 S-rRNA bands are shown to the left.

The 5′-flanking region of the gene has a TATA-box-like pyrimidine sequence, ATAAAA, 45–50 nucleotides upstream from the position of the 5′ end of the cDNA, i.e. at the expected distance if the 5′ end of the cDNA indeed is close to the mRNA cap site (Breathnach & Chambon, 1981). ATAAA sequences are found at comparable positions in the cystatin SN and SA genes (Saitoh *et al.*, 1987). The suggested possible CAAT-box sequence present 34 nucleotides further upstream in the cystatin SN and SA 5′-flanking regions, CAT, however, is absent in the cystatin C gene flanking region.

The 5′-flanking region of the cystatin C gene has a notably high GC content, with > 70% G+C in the 400 bp sequence upstream from the translation start site (Fig. 2). This GC-rich 'island' also includes the coding part of exon 1 and the 5′ part of the first intron, totally giving a 900 bp region with a G+C content of 73%. A large number of CpG dinucleotides are found in this region, resulting in a ratio of CpG/GpC dinucleotides close to unity (Fig. 2). This region thus has the properties of an 'HTF island', where CpG dinucleotides are non-methylated in contrast with most of the CpG dinucleotides found, at much lower frequency, elsewhere in the genome (Bird, 1986). The predominant number of previously located HTF islands are found in promoter regions of housekeeping genes, i.e. genes that are constantly transcribed in most tissues (Bird, 1986).

Two sequences which agree with 'GC-boxes' with a core consensus GGGCGG, or its inverse, which has been shown to be present in several housekeeping genes and to bind transcription factor Sp1 (Dynan, 1986) are present in the 1 kb 5′-flanking region of the cystatin C gene (Fig. 2). An enhancer core-like sequence (Sassone-Corsi & Borrelli, 1986) is also found in this region (Fig. 2). Sequences normally associated with the promoter regions of tightly controlled genes, such as hormone-responsive elements (Berg, 1989), however, are absent.

The 1 kb 5′-flanking region of the cystatin C gene is highly repetitive. A striking pentanucleotide repeat, GGAGG, which previously has been observed to be present in ten copies in the 1.4 kb 5′-flanking region of the human leucocyte elastase gene (Takahashi *et al.*, 1988) is repeated 35 times. This pentanucleotide sequence is included in three longer direct repeats of 17 nt (three copies), 15 nt (two copies) and 31 nt (two overlapping copies). The copies of the three longer repeats are all found within 400 bp upstream of the proposed TATA box (Fig. 2) and could therefore represent elements of importance for transcription of the cystatin C gene. A 16 nt sequence starting 17 residues downstream of the proposed TATA box, i.e. around the expected start position for transcription of the cystatin C gene, is found in an inverted copy 57–72 nt downstream from the translational start site (Fig. 2). This inverted repeat would probably be of importance for the secondary structure of the cystatin C mRNA.

Expression of the human cystatin C gene

RNA was isolated from human autopsy tissue specimens in order to study the expression of the cystatin C gene. The cystatin C cDNA detected mRNA species in kidney, liver, pancreas, intestine, stomach, antrum, lung, seminal vesicles and placenta, i.e. in all tissues examined (Fig. 3*b*). The strongest signal was seen for seminal-vesicle RNA. The Northern-blot results thus correlate well with the results of a previous study, which has shown that cystatin C is present in all major human biological fluids and, also, that seminal plasma has the highest cystatin C content of these (Abrahamson *et al.*, 1986).

The apparently non-tissue specific expression of the cystatin C gene could be related to the structure of its 5′-flanking region, which shares some properties with the promoter region of 'housekeeping genes', i.e. lacks typical CAAT-box, is notably GC-rich and contains binding sites for transcription factor Sp1 (Dynan, 1986). The gene for the Family 3 cystatin, kininogen, is predominantly expressed in liver and has a promoter region with lower GC content, typical CAAT and TATA boxes and lack Sp1-binding sites (Kitamura *et al.*, 1985). The expression of the genes for Family 2 cystatins, cystatin SN and cystatin SA, seems to be more restricted than that of the cystatin C gene, with expression products mainly found in secretions like saliva, tears and semen (Abrahamson *et al.*, 1986). The 340 bp sequence upstream of the translation start site in the cystatin SN gene and the corresponding 135 bp upstream sequence of the cystatin SA gene have been determined (Saitoh *et al.*, 1987). These sequences are similar to the corresponding part of the 5′-flanking region of the cystatin C gene with respect to high GC content (61 and 64% respectively), but they differ in having a significantly lower CpG content (ratio CpG/GpC of 1/9 and 1/16 respectively) and in having a proposed CAAT box, which is not found in the 5′-flanking region of the cystatin C gene. Also, the proposed Sp1-binding 'GC box' sequence closest upstream from the mRNA cap site in the cystatin C gene is absent from the cystatin SN and SA genes.

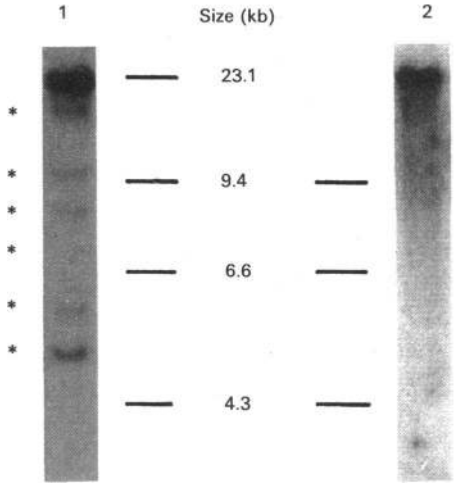

Fig. 4. Genomic sequences related to the sequence of the human cystatin C gene

*Sca*I digests of genomic DNA, hybridized to (1) cystatin C cDNA and (2) probe C3C (6 kb *Hin*dIII–*Eco*RI genomic fragment from clone λC3, see Fig. 1). Asterisks (*) show weak bands which did not hybridize to the 3′-flanking DNA probe, C3C. The positions of *Hin*dIII-cut λ-phage DNA size-marker bands are indicated. The blot was washed four times at 68 °C in 2 × SSC/0.1% SDS and finally for 10 min in 0.2 × SSC at 68 °C.

Size of the Family 2 cystatin gene family

Protein-sequence data for cystatins C, S, SN and SA have shown that they belong to the same protein family as defined by Dayhoff *et al.* (1979). The similarities in structure for the cystatin C, SN and SA genes, e.g. with respect to identical positions of intron sequences, reflects this grouping at the DNA level. At hybridization with the cystatin C cDNA to restriction-endonuclease digests of genomic DNA, we obtained results consistent with at least seven cross-reacting sequences in the genome. For example, the cystatin C cDNA hybridized weakly to six large *Sca*I restriction fragments in addition to the fragment harbouring the cystatin C gene. The cystatin C flanking gene probe, C3C, hybridized only to the cystatin C gene fragment, however (Fig. 4). Our results thus corroborate the estimate of approximately seven Family 2 cystatin genes (or pseudogenes) previously made after hybridizations of a cystatin SN genomic probe or a cystatin S cDNA probe to human DNA (Saitoh *et al.*, 1987; Al-Hashimi *et al.*, 1988). The entire nucleotide sequences of three of these are now known: the cystatin C gene, the cystatin SN gene, called CST1 and one cystatin pseudogene called CSTP1 (Saitoh *et al.*, 1987). The sequences of the exon parts of the cystatin SA gene, called CST2, are also known (Saitoh *et al.*, 1987). By analogy to the nomenclature used by Saitoh *et al.* (1987), we propose to call the cystatin C gene 'CST3'.

This work was supported by grants from Kungliga Fysiografiska Sällskapet i Lund, Konung Gustav V:s 80-årsfond, Direktör A. Påhlssons Stiftelse, A. Österlunds Stiftelse, G. och J. Kochs Stiftelse, M. Bergvalls Stiftelse, T. och E. Segerfalks Stiftelse, A. G. Crafoord's Foundation, Svenska Läkaresällskapet, the Medical Faculty, University of Lund, the Swedish Medical Research Council (Project nos. 05196 and 08660), the Icelandic Scientific Foundation, Heilavernd Society and the B. Magnusdottir and J. J. Bjarnason Foundation.

REFERENCES

Abrahamson, M., Barrett, A. J., Salvesen, G. & Grubb, A. (1986) J. Biol. Chem. **261**, 11282–11289

Abrahamson, M., Grubb, A., Olafsson, I. & Lundwall, Å. (1987) FEBS Lett. **216**, 229–233

Abrahamson, M., Islam, M. Q., Szpirer, J., Szpirer, C. & Levan, G. (1989) Hum. Genet. **82**, 223–226

Al-Hashimi, I., Dickinson, D. P. & Levine, M. J. (1988) J. Biol. Chem. **263**, 9381–9387

Bankier, A. T. & Barrell, B. G. (1983) in Techniques in Nucleic Acid Biochemistry (Flavell, R. A., ed.), vol. 83, pp. 1–73, Elsevier, Limerick

Barrett, A. J., Fritz, H., Grubb, A., Isemura, S., Järvinen, M., Katunuma, N., Machleidt, W., Müller-Esterl, W., Sasaki, M. & Turk, V. (1986*a*) Biochem. J. **236**, 312

Barrett, A. J., Rawlings, N. D., Davies, M. E., Machleidt, W., Salvesen, G. & Turk, V. (1986*b*) in Proteinase Inhibitors (Barrett, A. J. & Salvesen, G., eds.), pp. 515–569, Elsevier, Amsterdam

Bell, G. I., Karam, J. H. & Rutter, W. J. (1981) Proc. Natl. Acad. Sci. U.S.A. **78**, 5759–5763

Berg, J. M. (1989) Cell (Cambridge, Mass.) **57**, 1065–1068

Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 3963–3965

Bird, A. P. (1986) Nature (London) **321**, 209–213

Breathnach, R. & Chambon, P. (1981) Annu. Rev. Biochem. **50**, 349–383

Chomeczynski, P. & Sacchi, N. (1987) Anal. Biochem. **162**, 156–159

Cohen, D. H., Feiner, H., Jensson, O. & Frangione, B. (1983) J. Exp. Med. **158**, 623–628

Dayhoff, M. O., Barker, W. C. & Hunt, L. T. (1979) in Atlas of Protein Sequence and Structure (Dayhoff, M. O., ed.), vol. 5 (suppl. 3), pp. 9–20, National Biomedical Research Foundation, Washington

Dynan, W. S. (1986) Trends Genet. **2**, 209–213

Grubb, A. & Löfberg, H. (1982) Proc. Natl. Acad. Sci. U.S.A. **79**, 3024–3027

Grubb, A., Jensson, O., Gudmundsson, G., Arnason, A., Löfberg, H. & Malm, J. (1984) N. Engl. J. Med. **311**, 1547–1549

Isemura, S., Saitoh, E. & Sanada, K. (1984) J. Biochem. (Tokyo) **96**, 489–498

Isemura, S., Saitoh, & Sanada, K. (1986) FEBS Lett. **198**, 145–149

Isemura, S., Saitoh, E. & Sanada, K. (1987) J. Biochem. (Tokyo) **102**, 693–704

Kitamura, N., Kitagawa, H., Fukushima, D., Takagaki, Y., Miyata, T. & Nakanishi, S. (1985) J. Biol. Chem. **260**, 8610–8617

Levy, E., Lopez-Otin, C., Ghiso, J., Geltner, D. & Frangione, B. (1989) J. Exp. Med. **169**, 1771–1778

Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor

Mount, S. M. (1982) Nucleic Acids Res. **10**, 59–72

Ohkubo, I., Kurachi, K., Takasawa, T., Shiokawa, H. & Sasaki, M. (1984) Biochemistry **23**, 5691–5697

Palsdottir, A., Abrahamson, M., Thorsteinsson, L., Arnason, A., Olafsson, I., Grubb, A. & Jensson, O. (1988) Lancet **ii**, 603–604

Ritonja, A., Machleidt, W. & Barrett, A. J. (1985) Biochem. Biophys. Res. Commun. **131**, 1187–1192

Saitoh, E., Kim, H.-S., Smithies, O. & Maeda, N. (1987) Gene **61**, 329–338

Sassone-Corsi, P. & Borrelli, E. (1986) Trends Genet. **2**, 215–219

Staden, R. (1982*a*) Nucleic Acids Res. **10**, 2951–2961

Staden, R. (1982*b*) Nucleic Acids Res. **10**, 4731–4751

Takahashi, H., Nukiwa, T., Yoshimura, K., Quik, C. D., States, D. J., Holmes, M. D., Whang-Peng, J., Knutsen, T. & Crystal, R. G. (1988) J. Biol. Chem. **263**, 14739–14747

Turk, V., Brzin, J., Longer, M., Ritonja, A., Eropkin, M., Borchart, U. & Machleidt, W. (1983) Hoppe-Seylers Z. Physiol. Chem. **364**, 1487–1496

Received 11 September 1989/8 December 1989; accepted 5 January 1990