Biol. Chem., Vol. 385, pp. 363–372, May 2004

Comprehensive search for cysteine cathepsins in the human genome

Andrea Rossi[1], Quinn Deveraux[2], Boris Turk[3,*] and Andrej Sali[1,*]

[1] Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, University of California at San Francisco, San Francisco, CA 94143-2240, USA

[2] Cancer Biology Department, Genomics Institute of the Novartis Research Foundation (GNF), 10675 John Jay Hopkins Drive, San Diego, CA 92121, USA

[3] Department of Biochemistry and Molecular Biology, Jozef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia

* Corresponding authors
e-mail: sali@salilab.org; boris.turk@ijs.si

Abstract

Our study was aimed at examinating whether or not the human genome encodes for previously unreported cysteine cathepsins. To this end, we used analyses of the genome sequence and mRNA expression levels. The program TBLASTN was employed to scan the draft sequence of the human genome for the 11 known cysteine cathepsins. The cathepsin-like segments in the genome were inspected, filtered, and annotated. In addition to the known cysteine cathepsins, the scan identified three pseudogenes, closely related to cathepsin L, on chromosome 10, as well as two remote homologs, tubulointerstitial protein antigen and tubulointerstitial protein antigen-related protein. No new members of the family were identified. mRNA expression profiles for 10 known human cysteine cathepsins showed varying expression levels in 46 different human tissues and cell lines. No expression of any of the three cathepsin L-like pseudogenes was found. Based on these results, it is likely that to date all human cysteine cathepsins are known.

Keywords: Celera Discovery System (CDS); cysteine proteases; expression profile; human genome database; papain-like cathepsins; TBLASTN.

Introduction

Cysteine cathepsins are papain-like cysteine proteases normally localized in lysosomes. Eleven human cathepsins in the cysteine protease family have been sequenced so far: cathepsins B, H, L, S, C, K, O, F, V, X, and W (Table 1). Three-dimensional (3D) structures are available for 9 of the 11 sequences. Although these enzymes were believed to be involved primarily in intralysosomal protein degradation, the respective gene knockouts have revealed that some of them are also essential for a number of other important cellular processes (Turk et al., 2001b). These functions are often associated with a more restricted tissue localization. For example, cathepsin K is essential for normal bone remodeling (Chapman et al., 1997; Saftig et al., 1998); cathepsin S, and to a lesser extent cathepsins L, V and F, have important roles in the immune response (Nakagawa et al., 1998a,b; Shi et al., 1999); cathepsin C is essential for processing of the granule serine proteases granzyme A and B, neutrophil elastase and cathepsin G (Pham and Ley, 1999); and cathepsin L is involved in the terminal differentiation of keratinocytes (Roth et al., 2000). Cysteine cathepsins have also been implicated in various pathologies, such as cancer, osteoporosis, inflammation, and neurological disorders (Turk et al., 2000). These activities are often linked to their extralysosomal functions, thus making them potential drug targets, as illustrated by the crystallographic structures of cathepsins S, V, F, and K that were determined by researchers in the pharmaceutical industry (Table 1). A small inhibitor of cathepsin K is currently being tested in clinical trials as a drug against osteoporosis.

It is useful to have comprehensive knowledge about all members of a protein family of interest. For example, the diversity of their sequences may indicate their different functional roles. In the case of cathepsins, comprehensive knowledge about their sequences, structures, and functions would be advantageous because the discovery of a new member or subfamily with a unique tissue specificity could help in describing the evolution and biological roles of the entire family. To this end, we set out to identify all proteins in the human genome that have sequences similar to those of the known cysteine cathepsins.

Until the early 1990's, five of the cathepsins were discovered by traditional biochemical isolation and characterization. Over the last decade, however, the additional six cathepsins were identified through molecular biology techniques involving the detection of the corresponding genes. The stage is now set to conclude the discovery of all members of the family with the aid of a computational analysis of the complete human genomic sequence.

We start by identifying all potential cysteine cathepsins using a bioinformatic analysis of the draft sequence of the human genome. We then supplement the bioinformatics results with a comprehensive determination of the expression levels of cysteine cathepsins in various human tissues and cell lines.

Results

Scanning of the human genome

The mature sequences of the 11 known cysteine cathepsins were scanned against the NCBI and Celera versions

Table 1 Eleven cysteine cathepsins, two TIN-ag proteins, and three sequences with high similarity to CATL, encoded by the human genome.

Protein names in the literature	Short name	Chromosome and cyto-genetic band	PDB[a] code and structure references	Other references	Comments
Cathepsin B	CATB	8p22	1HUC (Musil et al., 1991) 1MIR* (Rat) (Cygler et al., 1996) 1PBH* (Turk et al., 1996) 2PBH*, 3PBH* (Podobnik et al., 1997)	(Chan et al., 1986)	Widely expressed
Cathepsin C (J, DDPI)	CATC	11q14.1	IJQP (RaT) (Olsen et al., 2001) 1K3B (Turk et al., 2001a)	(Paris et al., 1995)	Widely expressed; functional as a tetramer
Cathepsin F	CATF	11q13	1 M6D (on hold) (Somoza et al., 2002)	(Santamaria et al., 1999)	Widely expressed
Cathepsin H	CATH	15q24	8PCH (porcine) (Guncar et al., 1998)	(Fuchs and Gassen, 1989)	Widely expressed
Cathepsin K (O,O2)	CATK	1q21	1AYU, 1ATK (Zhao et al., 1997) 1MEM (McGrath et al., 1997) 7PCK* (Sivaraman et al., 1999) 1BY8* (Porcine) (LaLonde et al., 1999)	(Shi et al., 1995)	Expressed in osteoclasts and ovary
Cathepsin L	CATL	9q21	1ICF (Guncar et al., 1999) 1CJL* (Coulombe et al., 1996a,b)	(Gal and Gottesman, 1988)	Widely expressed
Cathepsin O	CATO	4q31	No structure available	(Velasco, et al., 1994)	Expressed in ovary, kidney and placenta
Cathepsin S	CATS	1q21	1GLO (Turkenburg et al., 2002)	(Shi et al., 1992)	High expression in cells of immune system
Cathepsin V (L2,U)	CATV	9q22	1FHO (Somoza et al., 2000)	(Santamaria et al., 1998a)	Expressed only in thymus and testis
Cathepsin W	CATW	11q13.1	No structure	(Brown et al., 1998)	Lymphopain; expressed only in T cells
Cathepsin X (Z,Y,P)	CATX	20q13	1EF7 (Guncar et al., 2000) 1DEU* (Sivaraman et al., 2000)	(Nagler and Menard, 1998; Santamaria et al., 1998b)	Widely expressed
Tubulointerstitial Nephritis Antigen	TIN-ag	6p11-12		(Zhou et al., 2000)	Expressed mostly in kidney and in the intestinal epithelium
Tubulointerstitial Nephritis Antigen Related Protein	TIN-ag-RP	1p34.3		(Wex et al., 2001)	Expressed in vascular smooth muscle cells
Cathepsin L-like 1	CATLL1	10q22.3		(11)	Pseudogene
Cathepsin L-like 2	CATLL2	10q22.3		(11)	Pseudogene
Cathepsin L-like 3	CATLL3	10q22.3		(11)	Pseudogene

[a]Protein Data Bank.
*Zymogen forms.

of the draft human genome using the program TBLASTN. For every queried sequence, TBLASTN identified approx. 30 segments of the genome that are similar to the query (cathepsin-like segments). The cathepsin-like segments may correspond to parts of genes that encode well-characterized proteins, parts of uncharacterized genes, pseudogenes or other non-coding regions, and parts of the genome that are sequenced or assembled incorrectly. Finally, some of the cathepsin-like matches, primarily short matches with large E-value scores, may occur by chance, without originating from a common ancestor. A search for novel cathepsin-like genes has to consider all of these possibilities.

As described in the following sections, we first associated some of the matches with known cathepsins. Next, we eliminated obvious artifacts. Finally, we interpreted the remaining matches to determine if any one of them could correspond to a novel cathepsin-like cysteine protease.

Identification of the 11 known cysteine cathepsins

Because each cathepsin-iike match generally covers only a small fraction of the query cathepsin sequence, the individual cathepsin-iike matches are insufficient for assigning the matched genomic region to a specific cathepsin. However, this task can be achieved by considering the clustering and the extent of the individual cathepsin-like matches along the genomic sequence (Figure 1). All the exons of each cathepsin query can be easily identified in both the public and Celera assemblies of the human genome.

Our predicted gene sequences for the 11 known cathepsins are the same as those in the NCBI and CDS

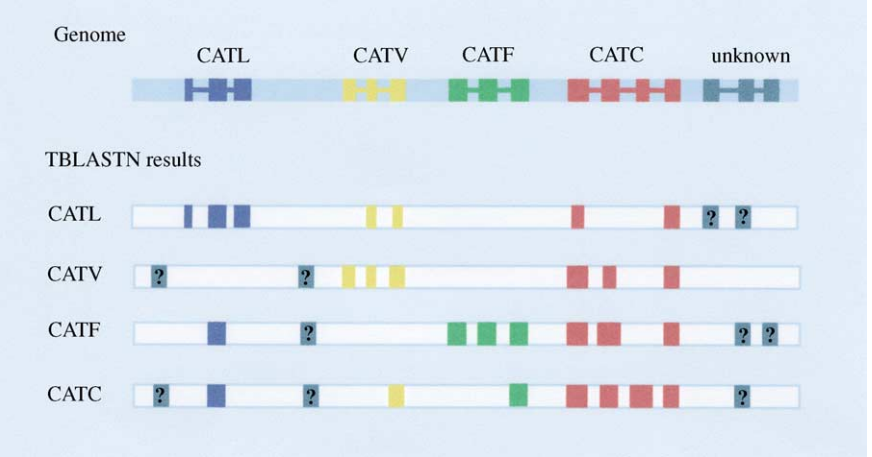

Figure 1 Scheme for discovering new cysteine cathepsins by querying the human genome with the known members of the family. Top; genes for the query cathepsins L, V, F, and C are already assigned in the genome. The exons of the new cathepsin, to be discovered by the TBLASTN searches, are indicated in dark grey. Bottom: multiple searches by TBLASTN highlight not only the exons corresponding to the query sequences, but also other parts of the genome. Discrimination between the matches that correspond to novel cathepsins and the artifacts involves careful manual inspection and use of protein sequence databases (see text).

annotations, except for small differences at the exon ends. These small differences occur because local sequence matching by TBLASTN does not consider the known signals for the start and end of gene exons. We also examined the chromosomes and the cytogenetic bands of the 11 cathepsins to verify that they were in agreement with the published data (Table 1). We did not find, at this level, any significant difference between the Human Genome Project and Celera CDS.

Every cysteine cathepsin was recovered by itself and by at least one other cathepsin in the family (Table 2). Therefore, if any one of the 11 known cysteine cathepsins was not yet documented, it would have been identified by our approach. This result validates our approach to identify previously unknown cysteine cathepsins.

Filtering of artifacts

With the TBLASTN matches corresponding to the 11 known cysteine cathepsins, we focused on the remaining cathepsin-like matches. Several matches for each cysteine cathepsin query were found in small contigs, typically 1–2 kbp long, that are still unmapped in the CDS database. There is no evidence that these matches correspond to real proteins; they may result from problems in the assembly of the genome and were ignored in further work.

Another example of an artifact was a 120-codon segment on chromosome 9 that matched three exons of cathepsin L and was located in the neighborhood of the cathepsin L gene. However, a stop codon corresponding to residue 35 of cathepsin L, in both the CDS and public genomic sequences, suggested that this cathepsin-like segment does not encode a functional gene, but is probably a pseudogene.

Identification of homologs of cysteine cathepsins

With the genomic sequences corresponding to the 11 known cysteine cathepsins and the obvious artifacts

Table 2 TBLASTN matches in the Celera human genome using the 11 known cysteine cathepsins as queries.

Query	Match															
	B	C	F	H	K	L	O	S	V	W	X	TIN-ag	TIN-ag-RP	L1	L2	L3
CATB	■	●		○	○	○		○	○		○					
CATC	○	■	○	○	○	○	○	○	○	○	○	○	○			
CATF	●	●	■	●	●	●	○	●	●	●						
CATH	○	●		■	●	●	●	●	●	○	○					
CATK		●		●	■	●	●	●	●		○			●	○	
CATL		●		○	●	■	○	●	●	○				●	●	●
CATO	○	●		○	●	●	■	●	●		○					
CATS	○	●		○	●		●	■	●		○				○	
CATV		●		○	●		●	●	■						●	
CATW		●		○			○			■						
CATX	○	●		○							■					

Identical results are obtained with the Human Genome Project version of the genome, except for the L.1, L2 and L3 (see results section). A full circle corresponds to at least one exon matched with an alignment significance E-value better than 10^{-5} (strong signal), and an open circle corresponds to at least one exon matched (weak signal). There is variability among cysteine cathepsins in terms of their ability to detect and be detected by other cathepsins. For example, cathepsin F as a query can detect cathepsins B, C, H, K, L, S, V, W and O, but only cahepsin C is able to detect the small exons of cathepsin F.

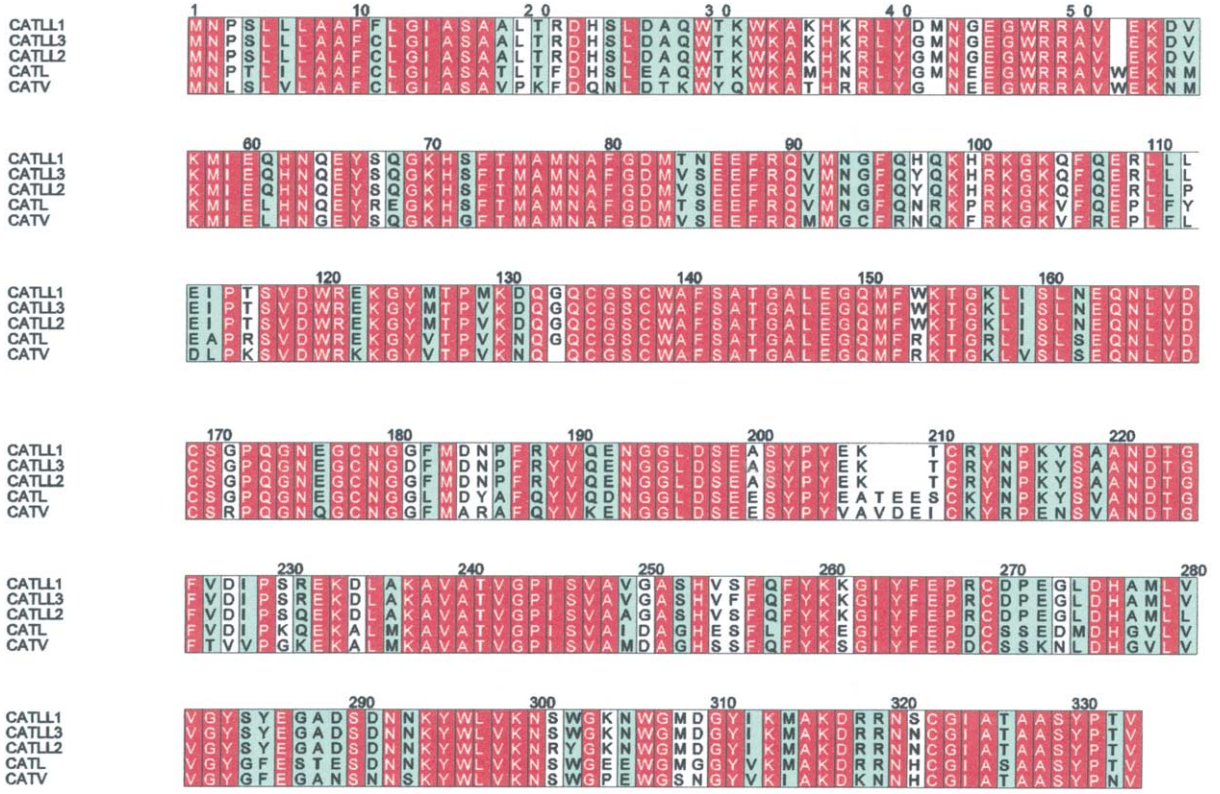

Figure 2 Multiple sequence alignment of cathepsins L and V, as well as putative cathepsins L1, L2, and L3 on chromosome 10. The blank positions in column 52 correspond to the stop codons in cathepsins L1, L2, and L3.

identified, the amino acid residue sequences of the remaining cathepsin-like segments were compared to the sequences in the non-redundant (NR) protein database at NCBI (Benson et al., 2002), using the program BLAST. Three of the cathepsin-like segments matched the amino acid residue sequence of cathepsin L at the 80% sequence identity level over the mature part of cathepsin L (Figure 2). In distinction to the public genome assembly, CDS contained only two of these three cathepsin-like segments, one of which is much shorter than the NCBI equivalent (3 vs. 5 exons).

The second group of matches of the cathepsin-like segments to the proteins in the NR database corresponded exactly to two known proteins that are similar to cathepsin C, but are not considered members of the cysteine cathepsin family. The first, on chromosome 6, is the tubulointerstitial protein antigen (TIN-ag) and the second, on chromosome 1, is TIN-ag-RP (TIN-ag-related protein; see Table 1).

mRNA expression profiles

Expression of cysteine cathepsins, based on their mRNA levels, was measured in 46 different tissues and cell lines (Figure 3). These samples included brain and neuronal tissues, tissues and organs involved in reproduction (placenta, ovary, uterus, testis), several glands (adrenal, pituitary, thyroid, salivary), lung, liver, heart, spleen and kidney, and a number of primarily lymphoma cell lines.

Some of the probes were unreliable because of the low signal-to-noise ratio. In general, a signal level of 200 corresponds to approximately 4 copies/cell, which is close to the lower limit of detection. Signals above 200 can thus be interpreted quantitatively, whereas signals between 0 and 200 imply only that the gene is expressed, but cannot be used to quantitatively track the expression levels or their trends.

Although there are several probes for some of the cysteine cathepsins on the chip (i.e., cathepsins K, X, O, V), only results for the most reliable probes are shown.

Cathepsin S had no reliable probes with a signal above 200. There were no probes for cathepsin B on the chip. In summary, the chip allowed us to obtain quantitative expression data for 9 out of 11 cathepsins (i.e., C, F, H, K, L, O, V, X and W) and qualitative estimates for cathepsin S.

The highest signals were obtained for cathepsins H and L in most of the samples examined, in agreement with previous data (Su et al., 2002). Cathepsin H is most abundant in kidney, lung and in several cancer cell lines. The highest expressions of cathepsin L were measured in fetal liver, liver, spleen, and placenta. Cathepsins C, F and O were also found to be expressed at relatively high levels, albeit not as high as cathepsins H and L in most of the samples. Cathepsins V, W and K were found in a limited number of samples, in agreement with previous observations (Brömme et al., 1999). Relatively low levels of cathepsin X expression were found only in lung, thymus and in the cancer cell line Hep3b. This observation is in apparent disagreement with the ubiquitous expression of cathepsin X determined by other methods, suggesting that the probe for cathepsin X may not be reliable.

None of the probes on the chip corresponded in their entirety to cathepsin L3. However, two probes for

cathepsins X and V with 100% sequence identity to cathepsin L3 over 8 residues did not reveal any signal, suggesting that cathepsin L3 is not expressed in any of the tested samples.

Discussion

With the study presented here we aimed to find out whether or not the human genome encodes for previously unreported cysteine cathepsins. Currently 11 human cysteine cathepsins are known (see Table 1 for references). The family is outlined in Figure 4, which displays a multiple sequence alignment, the corresponding clustering, and a superposition of crystallographic and modeled structures. All 11 known members are listed in the online MEROPS database (http://merops.sanger.ac.uk/; Rawlings et al., 2002). However, it is not clear whether or not these 11 cysteine cathepsins are all human cysteine cathepsins.

If the human genomic sequence were determined and assembled correctly, and all the genes were assigned reliably, the search for the detectable homologs of cysteine cathepsins would be straightforward. It would rely on the standard protein sequence comparison methods such as standard BLAST or PSI-BLAST; BLAST matches protein sequences in a pairwise fashion while PSI-BLAST matches a protein sequence to a multiple sequence 'profile' constructed in previous iterations through a database of sequences (Altschul et al., 1997). However, even though the Human Genome Project sequence is more than 91% finished (5 October, 2002), the list of the human genes is still substantially incomplete due to the difficulty of assigning genes based on a raw genomic sequence (Yeh et al., 2001). Therefore, we approached the problem of finding new cysteine cathepsins by matching the 11 known cysteine cathepsins to the raw genomic sequence using the TBLASTN program, followed by an inspection of the matches. This approach was feasible because of the relatively small number of the known cysteine cathepsin genes and their matches in the human genome. The manual inspection allowed us to carefully investigate short stretches of low similarity, as well as to benefit from information that is difficult to use in a computer program, such as the conservation of functional motifs. In principle, this approach is applicable to any gene family in any genome.

The scanning of the human genome identified all 11 of the known cysteine cathepsins, three additional cathepsin L variants, and two remote homologs that are not considered members of the cysteine cathepsin family (Table 2). No other cathepsin-like genes were identified. Our ability to detect all known cysteine cathepsins based on at least one other known cathepsin suggests that all human cysteine cathepsins were already discovered. Nevertheless, we cannot exclude a possibility that new cysteine cathepsins are yet to be identified, although our results suggest that the probability of such a discovery is low. New members of the family with very short exons (less than 30 amino acid residues) and/or new members that are very remotely related to the known members may have escaped our search. If any, they will be discovered experimentally, or by computation when the genes are assigned more reliably and more sensitive sequence matching methods than TBLASTN can be used for finding homologs of the known cysteine cathepsins.

To gauge the relative lack of limitations of our approach for the identification of new cysteine cathepsins, it is useful to discuss the TBLASTN matches of the known cathepsins against each other (Table 2). Cathepsin F was retrieved by only one query, cathepsin C; the match corresponds to a single exon containing the conserved sequence motif NSW and has an almost insignificant similarity score (E=9.4). The main reason for the difficult identification of cathepsin F through querying by its homologs is that cathepsin F has short exons of ca. 30 amino acid residues; there are 8 exons for the mature part of the enzyme. In contrast, there are only 3 exons for cathepsin C, 5 exons for cathepsins L, V, X, K, S and O, and 7 exons for cathepsins B and W, and 8 exons for cathepsin H. Conversely, the ability of cathepsin C to be retrieved by all other known proteases in the family is explained by the small number of exons and the large exon from amino acid residues 63 to 233 of the mature part of the protease. Other factors, in addition to exon length, must be important because cathepsins H, W and B have a number of exons similar to that of cathepsin F, but are easy to retrieve. An inspection of the genomic regions for cathepsins H, W and B that were matched by the queries revealed that they contain the completely conserved motif NSW or the partially conserved motif CGSC (with the two completely conserved cysteines); in cathepsin F, the exon containing the motif NSW is only 19 residues long, compared to 44 residues for cathepsin H and 40 residues for cathepsin B. Therefore, not surprisingly, both the exon length and degree of similarity are important factors limiting the power of our approach to identify new protein family members.

Three genomic regions similar to that of cathepsin L have already been described (Bryce et al., 1994). The authors, who refer to these three regions as cathepsins L1, L2 and L3, reported the complete sequence of cathepsin L1, which is 88% identical to that of cathepsin L, and short segments of cathepsins L2 and L3. Both cathepsins L1 and L2 have a stop codon corresponding to the 51st amino acid residue in the propart of cathepsin L. In cathepsin L3, the stop codon is replaced by a tryptophan codon, suggesting that cathepsin L3 may represent a functional member of the cathepsin L family, although no expression has been detected so far.

Knowledge of the expression of a protein is critical for the understanding of its biological function. Expression profiling based on mRNA arrays provides a major advantage over classical methods based on Northern blot analysis. Whereas the latter is normally limited to a single protein in a limited number of tissues, the former can be simultaneously examined in a much larger number of tissues and is not limited to a single protein. Furthermore, it is a quantitative method for comparison of the mRNA expression levels of a number of proteins, as shown here for cysteine cathepsins. A simultaneous scan for mRNA revealed the presence of all 10 cathepsins examined in at least one of the 46 examined tissues and confirmed

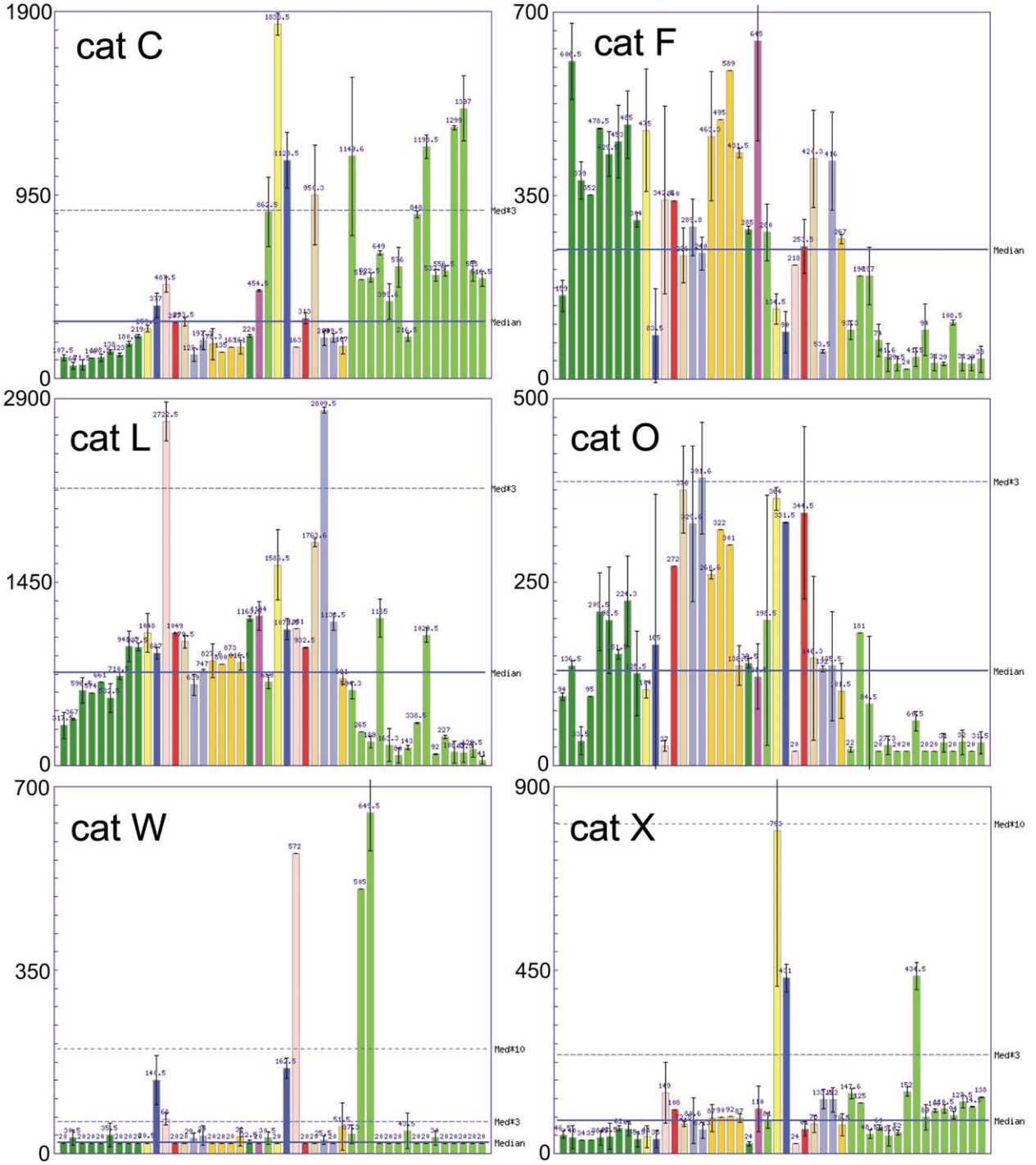

Figure 3 Expression profiles for 10 out of the 11 known cysteine cathepsins in 46 human tissues and cell lines.
Forty-six different human tissues and cell lines were examined for the presence of 10 cysteine cathepsins, except for cathepsin B for which no probe was available. Bars 1–9 (dark green), fetal brain, cerebellum, brain, cortex, caudate nucleus, amygdala, thalamus, corpus callosum, spinal cord; bar 10 (yellow), testis; bar 11 (dark blue), pancreas; 12 (pink) placenta; 13 (red) pituitary gland; 14 (light brown) thyroid; 15–16 (blue) cancer prostate; prostate; 17–20 (orange) ovary, OVR278E, OVR278S, uterus; 21 (dark green) DRG; 22 (magenta) salivary gland; 23 (green) trachea; 24 (yellow) lung; 25 (dark blue) thymus; 26 (pink) spleen; 27 (red) adrenal gland; 28 (light brown) kidney; 29–30 (blue) fetal liver, liver; 31 (orange) heart; 32–46 (light green) HUVEC, THY+, THY−, myelogenous K 562, lymphoblatic molt 4, Burkitt's Daudi, Burkitt's Raji, Hep3b, A2058, DOHH2, GA10, HL60, K422, Ramos and WSN. Signal values are absolute and quantitative (i.e., a value of 200 equals approx. 4 mRNA copies per cell).

some of the previously obtained results. Some of the data are completely new, such as the expression profile of cathepsins in the various parts of brain, in a number of glands, and several cancer cell lines. Although the probes for cathepsins S and X may not be reliable, the human library can be complemented with the mouse library, especially since the functions of the cathepsins seem to be fairly conserved between human and mouse (Turk et al., 2001a). Based on the expression levels of cathepsins B, X and S in mouse, it can be concluded

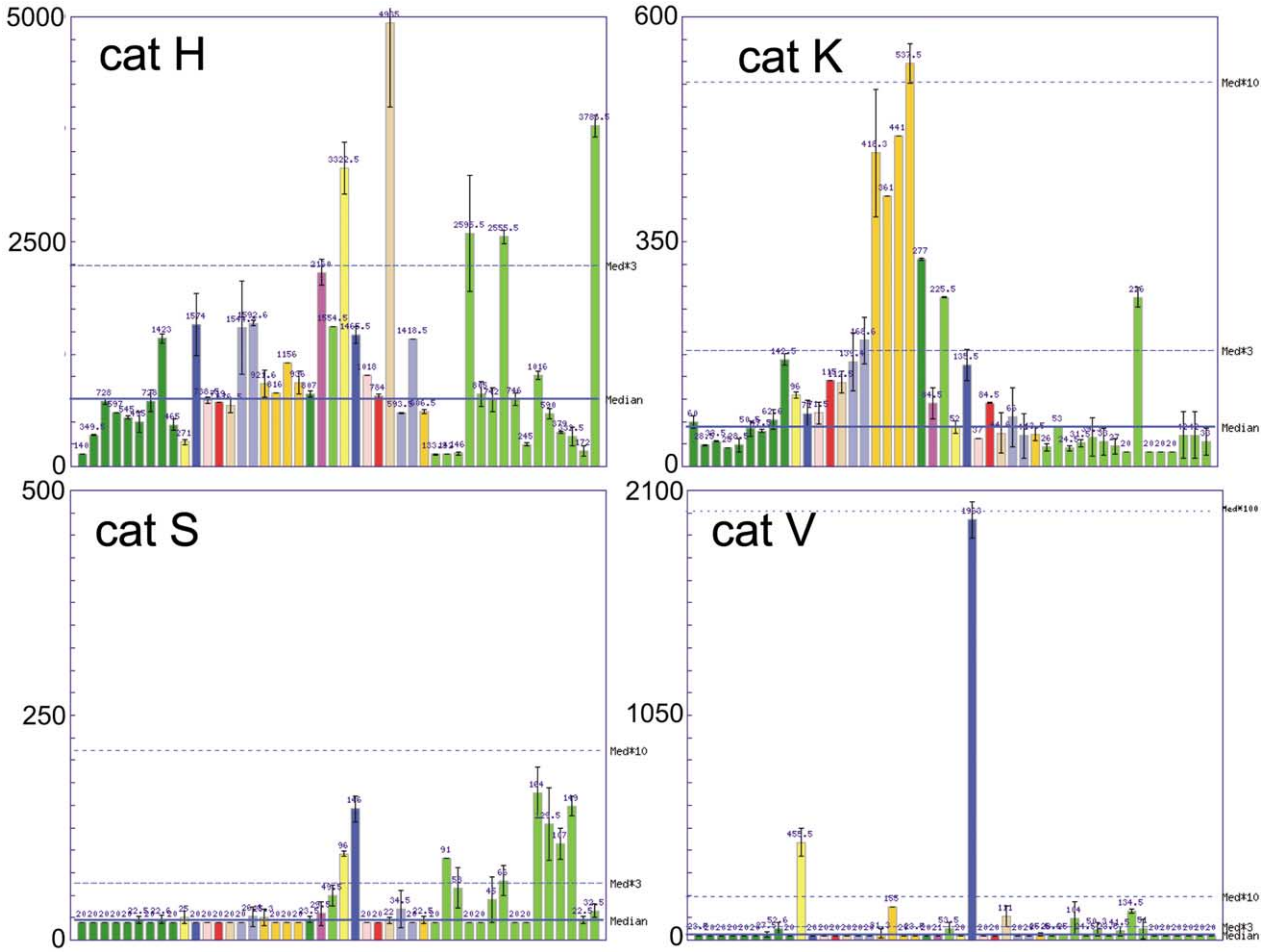

Figure 3 (continued)

that the mouse probes were reliable (data not shown; http://expression.gnf.org). Taking the mouse experiments into account, reliable quantitative expression profiles for all 11 cysteine cathepsins are available.

Given the many important functions of the cysteine cathepsins, it is perhaps puzzling that there may be only 11 of them in the human genome. The mouse genome encodes the orthologs of all human cathepsins, except for a single gene functionally replacing both cathepsins L and V, although being more similar to cathepsin V. Moreover, the mouse genome encodes eight additional cysteine cathepsins located on chromosome 13. These cathepsins were found to be expressed specifically in placenta (i.e., PECs; Deussing et al., 2002; Sol-Church et al., 2002). Mouse chromosome 13 corresponds to human chromosome 9, which, in addition to cathepsins L and V, appears to encode only a pseudogene of cathepsin L. Cathepsin L3 is encoded by chromosome 10q and appears not to be expressed in placenta. In comparison, the only cathepsin found to be highly expressed in the human placenta was cathepsin L. However, mouse cathepsins B and L were also found to be highly expressed in placenta, suggesting that these enzymes probably have only a housekeeping role in placenta.

In conclusion, we combined analyses of the human genome sequence and mRNA expression levels to examine the presence of new cysteine cathepsins. No new cathepsins were found, suggesting that there are only 11 functional human cysteine cathepsins.

Materials and methods

Human genome databases

The two draft sequences of the human genome were (i) the NCBI version of the human genome produced by the Human Genome Project (assembly build 22; Lander et al., 2001) and (ii) the Celera Inc. version of the human genome stored in the Celera Discovery System (CDS) (May 2001; Venter et al., 2001).

Sequence matching

The program TBLASTN, a variant of the standard BLAST (Altschul et al., 1990) (see http://www.ncbi.nlm.nih.gov/BLAST/producttabte.shtml for more information and comparison of different BLAST variants), was used to scan the genomic nucleotide sequences in all 6 reading frames for stretches that may encode segments of cysteine cathepsin-like amino acid residue sequences. It was accessed through the NCBI and CDS web servers, using the default parameters, except for the E-value significance score cutoff which was increased to 10. The results of the remote TBLASTN searches were retrieved and saved locally for further analysis.

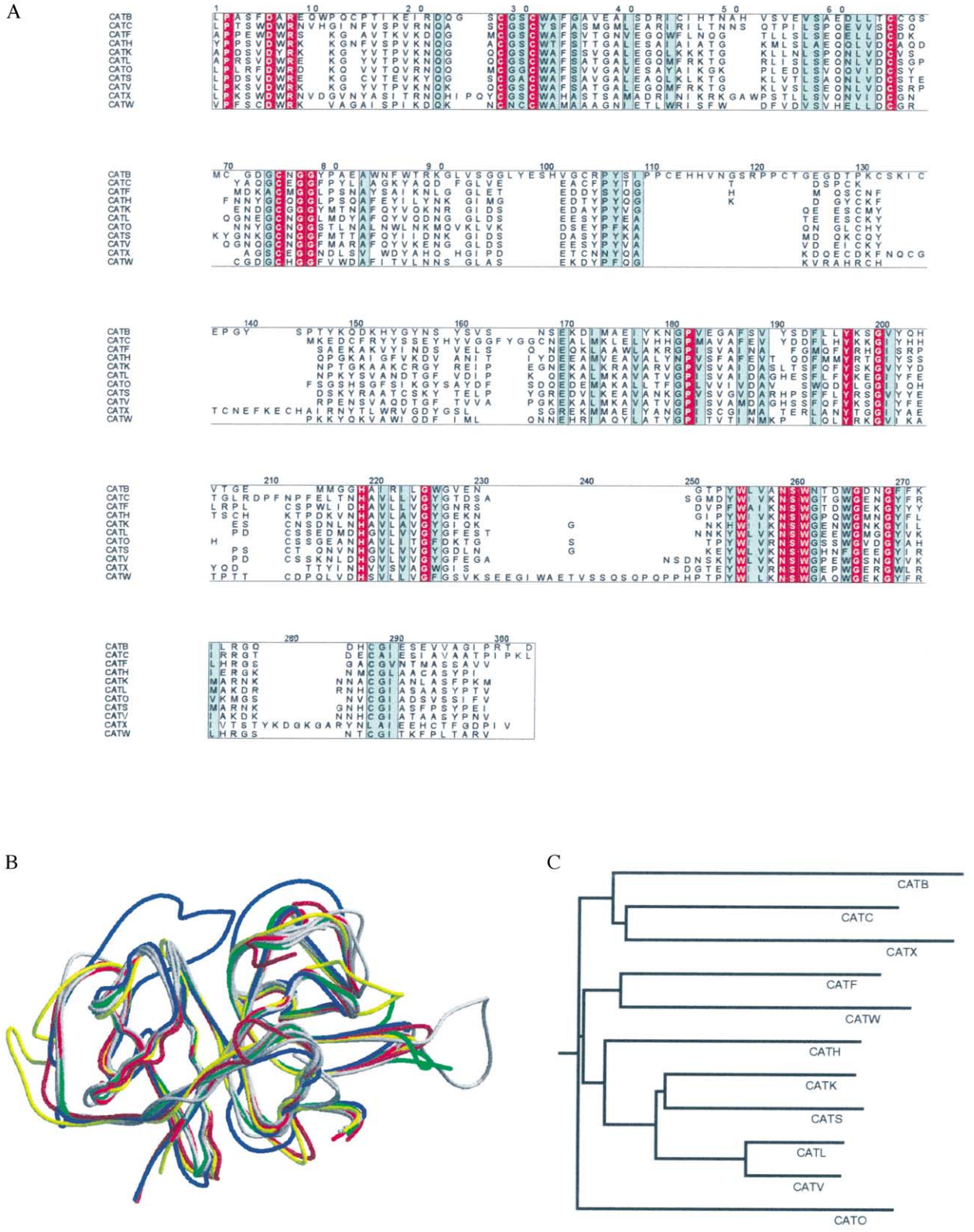

Figure 4 The family of human cysteine cathepsins.
(A) Sequence alignment. Red columns correspond to fully conserved positions, green columns to partially conserved positions. (B) Superposition of the known structures (color coded) and comparative models (grey); blue, CATB; brown, CATC; red, CATK; green, CATL, CATV; yellow, CATX. (C) Phylogenetic tree. The most similar (i.e., L and V) and the most dissimilar sequence pairs (e.g., B and X) share 79% and 25% sequence identity, respectively.

Multiple sequence and structure alignment, clustering, and display

The multiple structure-based alignment of 11 cysteine cathepsins was obtained by the SALIGN command in MODELLER (http://salilab.org/modeller) (Sali and Blundell, 1993; Figure 4). Seven of these 11 structures were determined by X-ray crystallography (Table 1) and the remaining 4 by comparative protein structure modeling. The comparative models were constructed by MODELLER and can be accessed in ModBase, a compre-

hensive database of comparative protein structure models for all protein sequences that are detectably related to at least one known protein structure (http://salilab.org/modbase; Pieper et al., 2002). The phylogenetic tree was calculated by the program CLUSTALW (Thompson et al., 1994), using the percentage sequence identities implied by the multiple alignment. The structures were displayed by the program Molscript (Kraulis, 1991).

Gene expression analysis

Total cellular RNA was prepared from tissue and cultured cells using the RNeasy Kit (Qiagen), and hybridized to oligonucleotide microarrays (U95Av2 GeneChip; Affymetrix Inc., Santa Clara, USA) (Su et al., 2002). Duplicates for all samples were performed. Scanned image files were analyzed with GENECHIP 3.2 (Affymetrix), and images were scaled to an average hybridization intensity of 200. Further details and access to the entire expression database can be found at http://expression.gnf.org. The individual cathepsins were assigned to the specific U95Av2 probe sequences by BLAST, using the cathepsin sequences as queries against the U95Av2 probe sequences.

Acknowledgements

This work was supported by NIH/GM 54762, Mathers Fund Award, and the Sandler Family Supporting Foundation (AS), as well as Slovenian Ministry of Schools, Science and Sports (BT). AS is an Irma T. Hirschl Trust Career Scientist.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Altschul, S.F., Madden, T.L, Schaffer, A.A, Zhang, J.H, Zhang. Z., Miller, W. and Lipman. D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell. J., Rapp, B.A. and Wheeler, D.L. (2002). GenBank. Nucleic Acids Res. *30*, 17–20.

Brömme, D., Li, Z.Q., Barnes, M. and Mehler, E. (1999). Human cathepsin V functional expression, tissue distribution, electrostatic surface potential, enzymatic characterization, and chromosomal localization. Biochemistry *38*, 2377–2385.

Brown, J., Matutes, E., Singleton, A., Price, C., Molgaard, H., Buttle, D. and Enver, T. (1998). Lymphopain, a cytotoxic T and natural killer cell-associated cysteine proteinase. Leukemia *12*, 1771–1781.

Bryce, S.D., Lindsay, S., Gladstone, A.J., Braithwaite, K., Chapman, C., Spurr, N.K. and Lunec, J. (1994). A novel family of cathepsin L-like (Ctsll) sequences on human-chromosome 10Q and related transcripts. Genomics *24*, 568–576.

Chan, S.J., San Segundo, B., McCormick, M.B. and Steiner, D.F. (1986). Nucleotide and predicted amino-acid-sequences of cloned human and mouse preprocathepsin-B cDNAs. Proc. Natl. Acad. Sci. USA *83,* 7721–7725.

Chapman, H.A., Riese. R.J. and Shi, G.P. (1997). Emerging roles for cysteine proteases in human biology. Annu. Rev. Physiol. *59*, 63–88.

Coulombe, R., Grochulski, P., Sivaraman, J., Menard, R., Mort, J.S. and Cygler, M. (1996a). Structure of human procathepsin L reveals the molecular basis of inhibition by the prosegment. EMBO J. *15,* 5492–5503.

Coulombe, R., Li, Y.G., Takebe, S., Menard, R., Mason, P., Mort, J.S. and Cygler, M. (1996b). Crystallization and preliminary X-ray diffraction studies of human procathepsin L. Proteins *25*, 398–400.

Cygler, M., Sivaraman. J., Grochulski, P., Coulombe, R., Storer, A.C. and Mort, J.S. (1996). Structure of rat procathepsin B: model for inhibition of cysteine protease activity by the proregion. Structure *4*, 405–416.

Deussing, J., Kouadio, M., Rehman, S., Werber, I., Schwinde, A. and Peters, C. (2002). Identification and characterization of a dense cluster of placenta-specific cysteine peptidase genes and related genes on mouse chromosome 13. Genomics *79*, 225–240.

Fuchs, R., and Gassen, H.G. (1989). Nucleotide-sequence of human preprocathepsin-H, A lysosomal cysteine proteinase. Nucleic Acids Res. *17,* 9471.

Gal, S., and Gottesman, M.M. (1988). Isolation and sequence of a cDNA for human pro-(cathepsin-L). Biochem. J. *253*, 303–306.

Gunčar, G., Podobnik, M., Pungerčar, J., Štrukelj, B., Turk, V. and Turk, D. (1998). Crystal structure of porcine cathepsin H determined at 2.1 Å resolution: location of the mini-chain C-terminal carboxyl group defines cathepsin H aminopeptidase function. Structure *6*, 51–61.

Gunčar, G., Pungerčič, G., Klemencic, I., Turk, V. and Turk. D. (1999). Crystal structure of MHC class H-associated p41 li fragment bound to cathepsin L reveals the structural basis for differentiation between cathepsins L and S. EMBO J. *18,* 793–803.

Gunčar, G., Klemenčič, I., Turk, B., Turk, V., Karaoglanovic-Carmona, A., Juliano, L. and Turk, D. (2000). Crystal structure of cathepsin X: a flip-flop of the ring of His23 allows carboxy-monopeptidase and carboxy-dipeptidase activity of the protease. Struct. Fold. Des. *5*, 305–313.

Kraulis, P.J. (1991). Molscript-a program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. *24*, 946–950.

LaLonde, J.M., Zhao, B.G., Janson, C.A., D'Alessio, K.J., McQueney, M.S., Orsini, M.J., Debouck, C.M. and Smith, W.W. (1999). The crystal structure of human procathepsin K. Biochemistry *38,* 862–869.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

McGrath, M.E., Klaus, J.L., Barnes, M.G. and Brömme, D. (1997). Crystal structure of human cathepsin K complexed with a potent inhibitor. Nat. Struct. Biol. *4*, 105–109.

Musil, D., Zucic, D., Turk, D., Engh, R.A., Mayr, I., Huber, R., Popovic, T., Turk, V., Towatari, T., Katunuma, N. et al. (1991). The refined 2.15-Å X-ray crystal-structure of human liver cathepsin-B - the structural basis for its specificity. EMBO J. *10,* 2321–2330.

Nagler, D.K., and Menard, R. (1998). Human cathepsin X: a novel cysteine protease of the papain family with a very short proregion and unique insertions. FEBS Lett. *434,* 135–139.

Nakagawa, T., Roth, W., Wong, P., Nelson, A., Farr, A., Deussing, J., Villadangos, J.A., Ploegh, H., Peters, C. and Rudensky, A.Y. (1998a). Cathepsin L: Critical role in li degradation and CD4 T cell selection in the thymus. Science *280,* 450–453.

Nakagawa, T., Rudensky, A., Wong, P., Farr, A., Nelson, A., Roth, W., Deussing, J. and Peters, C. (1998b). Cathepsin L: critical role in li degradation in thymic cortical epithelial cells and CD4 T cell development. FASEB J. *12,* A590.

Olsen, J.G., Kadziola, A., Lauritzen, C., Pedersen, J., Larsen, S. and Dahl, S.W. (2001). Tetrameric dipeptidyl peptidase I directs substrate specificity by use of the residual pro-part domain. FEBS Lett. *506*, 201–206.

Pariš, A., Štrukelj, B., Pungerčar, J., Renko, M., Dolenc, I. and Turk, V. (1995). Molecular cloning and sequence analysis of human preprocathepsin-C. FEBS Lett. *369*, 326–330.

Pham, C.T.N., and Ley, T.J. (1999). Dipeptidyl peptidase I is required for the processing and activation ofgranzymes A and B *in vivo.* Proc. Natl. Acad. Sci. USA *96*, 8627–8632.

Pieper, U., Eswar, N., Stuart, A.C., Ilyin, V.A. and Sali, A. (2002). MODBASE, a database of annotated comparative protein structure models. Nucleic Acids Res. *30,* 255–259.

Podobnik, M., Kuhelj, R., Turk, V. and Turk, D. (1997). Crystal structure of the wild-type human procathepsin B at 2.5 Å resolution reveals the native active site of a papain-like cysteine protease zymogen. J. Mol. Biol. *271,* 774–788.

Roth, W., Deussing, J., Botchkarev, V.A., Pauly-Evers, M., Saftig, P., Hafner, A., Schmidt, P., Schmahl, W., Scherer, J., Anton-Lamprecht, I. et al. (2000). Cathepsin L deficiency as molecular defect of furless: hyperproliferation of keratinocytes and pertubation of hair follicle cycling. FASEB J. *14,* 2075–2086.

Saftig, P., Hunziker, E., Wehmeyer, O., Jones, S., Boyde, A., Rommerskirch, W., Moritz, J.D., Schu, P. and von Figura, K. (1998). Impaired osteoclastic bone resorption leads to osteopetrosis in cathepsin-K-deficient mice. Proc. Natl. Acad. Sci. USA *95*, 13453–13458.

Sali, A., and Blundell, T.L. (1993). Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol. *234,* 779–815.

Santamaria, I., Velasco, G., Cazorla, M., Fueyo, A., Campo, E. and Lopez-Otin, C. (1998a). Cathepsin L2, a novel human cysteine proteinase produced by breast and colorectal carcinomas. Cancer Res. *58*, 1624–1630.

Santamaria, I., Velasco, G., Pendas, A.M., Fueyo, A. and Lopez-Otin. C. (1998b). Cathepsin Z, a novel human cysteine proteinase with a short propeptide domain and a unique chromosomal location. J. Biol. Chem. *273*, 16816–16823.

Santamaria, I., Velasco, G., Pendas, A.M., Paz, A. and Lopez-Otin, C. (1999). Molecular cloning and structural and functional characterization of human cathepsin F, a new cysteine proteinase of the papain family with a long propeptide domain. J. Biol. Chem. *274*, 13800–13809.

Shi, G.P., Munger, J.S, Meara. J.P, Rich, D.H. and Chapman, H.A. (1992). Molecular cloning and expression of human alveolar macrophage cathepsin S, an elastinolytic cysteine protease. J. Biol. Chem. *267*, 7258–7262.

Shi, G.P., Chapman, H.A., Bhairi, S.M, Deleeuw, C., Reddy, V.Y. and Weiss, S.J. (1995). Molecular cloning of human cathepsin O, a novel endoproteinase and homolog of rabbit-Oc2. FEBS Lett. *357*, 129–134.

Shi, G.P., Villadangos. J.A., Dranoff, G., Small, C., Gu, L.J., Haley, K.J., Riese, R., Ploegh, H.L. and Chapman, H.A. (1999). Cathepsin S is required for normal MHC class II peptide loading and germinal center development. Immunity *10,* 197–206.

Sivaraman, J., Lalumiere, M., Menard, R. and Cygler, M. (1999). Crystal structure of wild-type human procathepsin K. Protein Sci. *8,* 283–290.

Sivaraman, J., Nagler, D.K., Zhang, R.L, Menard, R. and Cygler, M (2000). Crystal structure of human procathepsin X: a cysteine protease with the proregion covalently linked to the active site cysteine. J. Mol. Biol. *295*, 939–951.

Sol-Church, K., Picerno, G.N., Stabley, D.L, Frenck, J., Xing, S.X., Bertenshaw, G.P. and Mason, R.W. (2002). Evolution of placentally expressed cathepsins. Biochem. Biophys. Res. Commun. *293*, 23–29.

Somoza, J.R., Palmer, J.T. and Ho, J.D. (2002). The crystal structure of human cathepsin F and its implications for the development of novel immunomodulators. J. Mol. Biol. *322*, 559–568.

Somoza, J.R., Zhan, H.J., Bowman, K.K., Yu, L., Mortara, K.D., Palmer, J-T., Clark, J.M. and McGrath, M.E. (2000). Crystal structure of human cathepsin V. Biochemistry *39*, 12543–12551.

Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A. et al. (2002b). Large-scale analysis of the human and mouse transcriptomes. Proc. Natl. Acad. Sci. USA *99*, 4465–4470.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). Clustal W-improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.

Turk, B., Turk, D. and Turk, V. (2000). Lysosomal cysteine proteases: more than scavengers. Biochim. Biophys. Acta *1477,* 98–111.

Turk, D. Turk, D., Janjič, V., Štern, I., Podobnik, M., Lamba, D., Dahl, S.W., Lauritzen, C., Pedersen, J., Turk. V. and Turk, B. (2001a). Structure of human dipeptidyl peptidase I (cathepsin C): exclusion domain added to an endopeptidase framework creates the machine for activation of granular serine proteases. EMBO J. *20,* 6570–6582.

Turk, D., Podobnik. M., Kuhelj, R., Dolinar, M. and Turk, V. (1996). Crystal structures of human procathepsin B at 3.2 and 3.3 Å resolution reveal an interaction motif between a papain-like cysteine protease and its propeptide. FEBS Lett. *384,* 211–214.

Turk, V., Turk, B. and Turk, D, (2001b). Lysosomal cysteine proteases: facts and opportunities. EMBO J. *20*, 4629–4633.

Turkenburg, J.P., Lamers, M.B.A.C., Brzozowski, A.M., Wright, L.M., Hubbard, R.E., Sturt, S.L. and Williams, D.H. (2002). Structure of a Cys25→Ser mutant of human cathepsin S. Acta Crystallogr. D Biol. Crystallogr. *58*, 451–455.

Velasco, G., Ferrando, A.A., Puente, X.S., Sanchez, L.M. and Lopez-Otin, C. (1994). Human cathepsin O—molecular cloning from a breast-carcinoma, production of the active enzyme in *Eschehchia coli*, and expression analysis in human tissues. J. Biol. Chem. *269*, 27136–27142.

Venter, J.C., Adams. M.D., Myers. E.W., Li, P.W, Mural, R.J., Sutton. G.G., Smith, H.O.,Yandell, M., Evans, C.A., Holt, R.A.etal. (2001). The sequence of the human genome. Science *291,* 1304–1351.

Wex, T., Lipyansky, A., Brömme, N.C., Wex, H., Guan, X.Q. and Brömme, D. (2001). TIN-ag-RP, a novel catalytically inactive cathepsin B-related protein with EGF domains, is predominantly expressed in vascular smooth muscle cells. Biochemistry *40,* 1350–1357.

Yeh, R.F., Lim, L.P.and Surge, C.B. (2001). Computational inference of homologous gene structures in the human genome. Genome Res. *11,* 803–816.

Zhao, B.G., Janson, C.A., Amegadzie, B.Y., DAlessio, K., Griffin, C., Hanning, C.R., Jones, C., Kurdyla, J., McQueney, M., Qiu, X.Y. et al. (1997). Crystal structure of human osteoclast cathepsin K complex with E-64. Nat. Struct. Biol. *4*, 109–111.

Zhou, B., Nelson, T.R., Kashtan. C., Gleason, B., Michael, A.F., Vlassi, M. and Charonis, A.S. (2000). Identification of two alternatively spliced forms of human tubulointerstitial nephritis antigen (TIN-Ag). J. Am. Soc. Nephrol. *11*, 658–668.

Received November 26, 2003; accepted February 11, 2004