Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 3063–3067, April 1993
Biochemistry

Two distinct gene subfamilies within the family of cysteine protease genes

(tetrahymena/propeptide/cathepsin)

KATHLEEN M. KARRER*, STACIA L. PEIFFER†, AND MICHELE E. DITOMAS

Department of Biology, Marquette University, Milwaukee, WI 53233

Communicated by David M. Prescott, January 7, 1993

ABSTRACT A cDNA clone for a physiologically regulated *Tetrahymena* cysteine protease gene was sequenced. The nucleotide sequence predicts that the clone encodes a 336-amino acid protein composed of a 19-residue N-terminal signal sequence followed by a 107-residue propeptide and a 210-residue mature protein. Comparison of the deduced amino acid sequence of the protein with those of other cysteine proteases revealed a highly conserved interspersed amino acid motif in the propeptide region of the protein, the ERFNIN motif. The motif was present in all of the cysteine proteases in the data base with the exception of the cathepsin B-like proteins, which have shorter propeptides. Differences in the propeptides and in conserved amino acids of the mature proteins suggest that the ERFNIN proteases and the cathepsin B-like proteases constitute two distinct subfamilies within the cysteine proteases.

The cysteine proteases are a family of enzymes that play an important role in intracellular protein degradation. These proteases and their cDNA clones have been isolated from phylogenetically diverse organisms ranging from slime mold to mammals. The tertiary structures of two plant cysteine proteases, papain and actinidin, have been solved (1, 2). The enzymes have two protein domains that come together to form the active site. Amino acid sequence homologies suggest this double domain structure is conserved in the animal thiol proteases cathepsins B, H, and L (3).

The phylogenetic range of organisms for which the sequence of cysteine protease genes are known was extended by determination of the sequence of a cDNA clone for a gene from a ciliated protozoan, *Tetrahymena thermophila*.‡ Comparison of the deduced amino acid sequence to those of known cysteine proteases revealed the presence of an amino acid motif in the propeptide region consisting of highly conserved amino acids interspersed with variable residues. The motif was present in 15 of 20 cysteine proteases in the EMBL/GenBank data base (August 1992). The five proteases that lacked the motif were all cathepsin B-like enzymes. Recognition of the differences in the propeptide region prompted comparison of the mature proteins. Alignment of the amino acid sequences of the proteases as two separate groups allowed identification of amino acids that are highly conserved among the proteases with the propeptide motif or among the cathepsin B-like proteases but strikingly different between the two groups. We suggest that the proteins with the interspersed motif and the cathepsin B-like proteases represent two distinct classes of cysteine proteases that can be distinguished by both propeptide and mature protein structure.

MATERIALS AND METHODS

Tetrahymena clone pCyP (formerly BC11) is a cDNA clone of an RNA that is expressed in starved, but not growing, cells (4, 5). The clone was isolated from a cDNA library of RNA from starved cells cloned into the *Pst* I site of pUC9 (4). DNA fragments were subcloned into pBluescript for sequencing. The sequence was scanned for open reading frames by using the DNA INSPECTOR IIE program (Textco), taking into consideration that in *Tetrahymena*, as in several ciliates, TAA and TAG code for Gln (6–8). DNA sequences that code for homologous proteins were identified through a Pearson and Lipman (9) search of the EMBL/GenBank data base by using the TFASTA program.

RESULTS

The 1189-bp nucleotide sequence of *Tetrahymena* pCyp and the derived amino acid sequence of the protein encoded by the single long open reading frame beginning at the first AUG are presented in Fig. 1. The open reading frame encodes a protein of 336 amino acids with a calculated molecular weight of 37,716. A short poly(A) sequence at the 3′ end of the insert is 33 bp downstream from a consensus poly(A) addition site, AAUAAA, and is presumably derived from the poly(A) tail of the RNA.

The amino acid sequence of pCyP is highly homologous to those of cysteine proteases from a variety of eukaryotes. The deduced sequence of the *Tetrahymena* protein is shown in Fig. 2 along with sequences of representative cysteine proteases from a slime mold, a plant, an arthropod, and a mammal. The sequences have been aligned to maintain blocks of homology and to align cysteines that form disulfide bridges in papain (3). The numbers above the *Tetrahymena* sequence have been assigned with positive numbers for the putative mature protein and negative numbers for the prepro-region of the protein. Blocks of highly conserved sequence contain the Gln^{19}, Cys^{25}, His^{162}, Asn^{179}, and Trp^{181} residues that are present at the active site of cysteine proteases (1). Conservation of amino acid sequence predicts that the mature *Tetrahymena* protein has a molecular weight of 22,850 and an N-terminal Leu.

The structure of the *Tetrahymena* cysteine protease gene suggests that it is translated as a preproenzyme. The open reading frame has a preponderance of hydrophobic residues in the first 19 amino acids of the protein, as expected for the signal peptide commonly found at the N terminus of cysteine proteases. Calculations according to the weight-matrix method of von Heijne (15) predict that the signal sequence cleavage site is after the first Ala residue (Fig. 1).

The putative propeptide region contains a block of conserved amino acids from Leu^{-51} to Phe^{-40}, which has previously been noted as a feature of several cysteine proteases (16). When the propeptide regions were aligned with minimal

*To whom reprint requests should be addressed.

†Present address: Department of Genetics, Washington University, St. Louis, MO 63110.

‡The sequence reported in this paper has been deposited in the GenBank data base (accession no. L03212).

```
                                               M  N  K  K  F  I  I  L  S  I  I  M  L  M
1       AAAATAAAAAAAAAATAAAAAAAACTAAAACTTTAAAGTATGAATAAAAAAATTCATCATTTTGAGTATTATCATGCTCATG

                       ▼
         P  L  C  L  A  Q  D  I  S  V  E  K  L  L  A  Y  N  K  W  S  S  Q  N  Q  R  A
80      CCTCTCTGTTTGGCTCAAGATATAAGTGTAGAAAAACTTCTTGCTTATAATAAATGGTCAAGCTAAAATCAAAGAGCC

         Y  L  N  E  D  E  K  L  Y  R  Q  I  V  F  F  E  N  L  Q  K  I  K  E  H  N  S
158     TATCTGAATGAAGATGAAAAACTGTATAGACAAATAGTTTTCTTTGAAAATTTGTAAAAAATTAAGGAGCATAACAGT

         N  P  N  N  T  Y  S  I  H  L  N  Q  F  S  D  M  T  R  E  E  F  A  E  K  I  L
236     AACCCTAATAACACCTATTCTATCCATTTAAACTAATTCTCAGATATGACTAGAGAAGAATTTGCAGAAAAAATTCTT

         M  K  Q  D  L  I  N  D  Y  M  K  G  I  G  Q  Q  A  T  H  N  N  A  N  N  E  T
314     ATGAAATAGGATTTGATTAACGATTATATGAAGGGAATTGGTTAATAGGCTACTCACAATAATGCTAATAACGAAACT

                                ▼
         Q  M  N  S  Q  N  H  T  L  A  A  S  I  D  W  R  T  K  G  A  V  T  S  V  K  D
392     TAAATGAATTCATAAAACCATACTTTAGCTGCTTCTATAGATTGGAGAACAAAAGGTGCTGTAACATCGGTTAAGGAT

         Q  G  Q  C  G  S  C  W  S  F  S  A  A  A  L  M  E  S  F  N  F  I  Q  N  K  A
470     TAAGGTTAATGTGGTTCATGCTGGAGTTTCTCTGCAGCTGCCTTAATGGAGTCATTTAACTTCATTTAAAACAAAGCT

         L  V  N  F  S  E  Q  Q  L  V  D  C  V  T  P  E  N  G  Y  P  S  Y  G  C  K  G
548     TTAGTTAATTTTTCTGAGTAATAACTTGTTGATTGTGTGACCCCTGAAAATGGTTACCCCTCTTATGGATGTAAAGGA

         G  W  P  A  T  C  L  D  Y  A  S  K  V  G  I  T  T  L  D  K  Y  P  Y  V  A  V
626     GGATGGCCTGCTACTTGTCTGGATTATGCCTCCAAAGTAGGTATCACAACACTAGACAAGTATCCCTATGTTGCAGTA

         Q  K  N  C  T  V  T  G  T  N  N  G  F  K  L  K  K  W  I  V  I  P  N  T  S  N
704     CAGAAAAATTGTACTGTGACAGGTACAAATAATGGCTTTAAGCTTAAAAAGTGGATTGTAATTCCTAACACTTCAAAC

         D  L  K  S  A  L  N  F  S  P  V  S  V  L  V  D  A  T  N  W  D  Y  Y  S  S  G
782     GACTTAAAAAGTGCTTTAAATTTCTCTCCTGTTTCTGTTCTTGTTGATGCTACCAATTGGGATTATTATTCGTCTGGA

         I  F  N  G  C  N  Q  T  N  I  N  L  N  H  A  V  L  A  V  G  Y  D  E  K  D  N
860     ATTTTCAACGGATGTAATTAAACTAATATTAATCTTAATCATGCTGTATTAGCTGTAGGCTATGACGAAAAAGATAAC

         W  I  V  K  N  S  W  S  A  G  W  G  E  H  G  Y  I  R  L  A  P  N  N  T  C  G
938     TGGATTGTTAAAAATTCTTGGAGCGCTGGTTGGGGTGAACATGGATATATTAGACTTGCTCCTAACAATACATGTGGT

         I  L  S  S  N  I  Q  V  T  A  *
1016    ATCTTAAGCTCTAATATATAAGTTACTGCTTGAAAATTAGGATAAGCTATTATAATTAAAATTTATTAAAATATATTA


1094    TTCGTATAAACAAAAATATTATATAAAATAAATAGTCTTTCAAATATTAAGATTTGTTTAATTTTA$_{31}$
```

FIG. 1. DNA sequence and deduced amino acid sequence of pCyP. Arrows, putative sites of posttranslational cleavage; ★, stop codon; underlined, poly(A) addition site. The sequence contains 13 TAA and 2 TAG Gln codons.

gaps, a consensus sequence was found between Glu^{-81} and Asn^{-62} that consists of conserved amino acids interspersed with variable ones: $EX_3RX_2(V/I)FX_2NX_3IX_3N$. This "ERFNIN" motif, named for the single letter code of the conserved amino acids, was found in 15 cysteine protease genes identified in a combined search of the literature and the GenBank data base (Table 1).

All of the cysteine proteases with similarity to the mammalian H and L cathepsins contain the ERFNIN motif and amino acid variants within the motif generally display a high degree of structural similarity to the consensus residue. Discounting the *Trypanosoma* protease, the first two amino acids of the motif, Glu and Arg, and the final Asn are perfectly conserved among the 14 enzymes. The Phe is present in 11 of the proteins and in the other 3 Phe is replaced by another amino acid with an aromatic ring, Trp. There are two variations in the first Asn of the motif and the three examples in which the consensus Ile is replaced by Val. These are all highly conservative changes as measured by the scale of Feng *et al.* (14).

The most unusual cysteine protease that bears an identifiable ERFNIN motif is the *Trypanosoma* enzyme (Table 1). In this protein, three residues in the motif are replaced by Ala, which does not bear strong structural resemblance to the consensus amino acid. The significance of this degree of variation is unknown; however, it is noted that the *Trypanosoma* cysteine protease is also unique in another respect. It has a long 108-residue extension at the C-terminal end in the deduced protein. Estimates of the molecular mass suggest that at least some of the C-terminal extension persists in the mature enzyme (17).

The interspersion distance of the conserved amino acids in the ERFNIN motif suggests that they lie along one face of an α-helix. In addition, the interspersed motif is found at a discrete distance from the block of conserved amino acids in the propeptide. In 7 of the 15 cysteine proteases, 14 amino acids are present between the last Asn of the ERFNIN motif and the first Asn of the conserved block. In the other 8 proteases, there are 10 or 11 amino acids, a number consistent with one fewer turn of an α-helix. This suggests that the interspersed ERFNIN motif plus the conserved block previously noted by Ishidoh *et al.* (16) constitute a functional unit.

The five cysteine proteases in the data base that did not contain the ERFNIN motif were the cathepsin B-like proteases. The propeptides of the cathepsin B-like proteases are shorter than those of the proteases that contain the ERFNIN motif. It is difficult to predict from conservation of sequence

```
Tetrahymena cysteine protease (Tt) ..............................MNKKFIILS
Dictyostelium cysteine protease 1 (Dd) ...............................MKVILLF
Papaya papain (Cp)....................MAMIPSISKLLFVAICLFVYMGLSFGDF
Lobster cysteine protease 2 (Ha).............................................
Rat cathepsin L (Rn)...........................................MTPLLLL

             -110                -90                 -70
               .                   .                   .
Tt      IIMLMPLCLAQDISVEKLLAYNKWSSQNQRAYLNEDEKLYRQIVFFENLQKIKEHNSNPN
Dd      VLAVFYVFVSSRGIPPEEQSQFLEFQDKFNLLYSHEEYLERFEIFKSNLGKIEELNLIAI
Cp      SIVGLYSQNDLTSTERLIQLFESWMLKHNKIYKNIDEKIYRFEIFKDNLKYIDETNKKNN
Ha      .MKVAVLFLCGVALAAASPSWEHFKGKYGRQYVDAEEDSYRRVIFEQNQKYIEEFNKKYE
Rn      AVLCLGTALATPKFDQTFNAQWHQWKSTHRRLYGTNEEEWRRAVWEKNMRMIQLHNGEYS
                                            =   =  -   =   =   =

                 -50                 -30                 -10
                  .                   .                   .
Tt      NTYS...IHLNQFSDMTREEFAEKILMKQDLINDYMKGIGQQATHNNANNETQMNSQNHT
Dd      NHKADTKFGVNKFADLSSDEFKNYYLNNKEAIFTDDLPVADYLDDEFINS..........
Cp      SYW....LGLNVFADMSNDEFKEKYTGSIAGNYTTTELSYEEVLNDGDVN..........
Ha      NGEVTFNLAMNKFGDMTLEEFNAVMKGNIPRRSAPVSVFYPKKETGPQ............
Rn      NGKHGFTMEMNAFGDMTNEEFRQIVNGYRHQKHKKGRLFQEPLMLQ..............
        -        -= =-=-- -==

                10                  30                  50
                 .        *  Ø  *    .                   .     Ø
Tt      LAASIDWRTKGAVTSVKDQGQCGSCWSFSAAALMESFNFIQNKALVNFSEQQLVDCVTPE
Dd      IPTAFDWRTRGAVTPVKNQGQCGSCWSFSTTGNVEGQHFISQNKLVSLSEQNLVDCDHEC
Cp      IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNEYSEQELLDCDRRS
Ha      .ATEVDWRTKGAVTPVKDQGQCGSCWAFSTTGSLEG.HFLKTGSLISLAEQQLVDCSRPY
Rn      IPKTVDWREKGCVTPVKNQGQCGSCWAFSASGCLEGQMFLKTGKLISLSEQNLVDCSHDQ
        --  -=== -= ==-==-== =====-==- - -=-   -    =   -== =-==

                     70                   90                   110
                    Ø .                    .          Ø         .
Tt      NGYPSY.....GCKGGWPATCLDYAS.KVGITTLDKYPYVAVQKN.CTVTGTNNGFKLKK
Dd      MEYEGEEACDEGCNGGLQPNAYNYIIKNGGIQTESSYPYTAETGTQCNFNSANIGAKISN
Cp      Y..........GCNGGYPWSALQLVA.QYGIHYRNTYPYEGVQRY.CRSREKGPYAAKTD
Ha      GPQ........GCNGGWMNDAFDYIKANNGIDTEAAYPYEARDGS.CRFDSNSVAATCSG
Rn      GNQ........GCNGGLMDFAFQYIKENGGLDSEESYPYEAKDGS.CKYRAEYAVANDTG
                   ==-==             =-     ===  -     =

                       130                     150
                        .                       .    Ø        *
Tt      WIVIPNTSNDLKSALNF..SPVSVLVDAT..NWDYYSSGIFNG..CNQTNINLNHAVLAV
Dd      FTMIPKNETVMAGYIVST.GPLAIAADAV..EWQFYIGGVFDIP.CNPN..SLDHGILIV
Cp      GVRQVQPYNEGALLYSIANQPVSVVLEAAGKDFQLYRGGIFVGP.CGNK...VDHAVAAV
Ha      HTNIASGSETGLQQAVRDIGPISVTIDAAHSSFQFYSSGVYYEPSCSPS..YLDHAVLAV
Rn      FVDIPQQEKALMKPVATV.GPISVAMDASHPSLQFYSSGIYYEPNCSSK..DLDHGVLVV
                            =---  -=     - = -=--    =      --=--  =

          170                         190
          .                *  *        .             Ø
Tt      GYDEKD........NWIVKNSWSAGWGEHGYIRLAP.....NNTCGILSSNIQVTA
Dd      GYSAKNTIFRKNMPYWIVKNSWGADWGEQGYIYLRRGK....NTCGVSNFVSTSII
Cp      GYGPN.........YILIKNSWGTGWGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN
Ha      GYGSEGGQD.....FWLVKNSWATSWGDAGYIKMSRNR...NNNCGIATVASYPLV
Rn      GYGYEGTDSNKDK.YWLVKNSWGKEWGMDGYIKIAKDR...NNHCGLATAASYPIVN
        ==              --====-  ==  === -          ==- -
```

FIG. 2. Deduced amino acid sequence of the *Tetrahymena* cysteine protease and four representative cysteine proteases from *Dictyostelium* (10), papaya (11), lobster (12), and rat (13). ★, Amino acids in the active site; ∅, cysteines in disulfide bridges in papain; =, conserved in all five proteins; –, similar in all five proteins where any pair of amino acids within the group have a score of 4 or greater on the scale of Feng *et al.* (14); . , gaps introduced to maximize alignment; boldface type, residues conserved in all known cysteine proteases. The numbers above the *Tetrahymena* sequence indicate the number of the amino acid with positive numbers for the putative mature protein and negative numbers for the prepro- region of the protein.

which amino acids in the cathepsin B propeptide are required for function. The high degree of similarity in the propeptides of the three mammalian cathepsin B-like proteases may simply reflect short evolutionary distance. Conservation of the peptide sequence −57 to −37 between the mammalian enzymes and cathepsin B from *Schistosoma mansoni* (Fig. 3) suggests that this region may be important for propeptide function, but the *Haemonchus contortus* propeptide shows no striking homology to the others. The sequences of cDNAs for additional cathepsin B-like proteases are required to determine whether there is a conserved motif in their propeptide regions.

The subfamily of cysteine proteases that contains the ERFNIN motif encompasses all of the enzymes described thus far that are similar to mammalian cathepsins H and L. The cathepsin B-like enzymes apparently constitute a separate class of cysteine proteases with respect to the structure of the mature protein and the propeptide. With regard to the overall structure, the proteases containing the interspersed ERFNIN motif in the propeptide have longer propeptides and are processed to smaller mature enzymes than the cathepsin B-like proteins. In terms of amino acid sequence, there are several examples of active site residues that are highly conserved within the ERFNIN proteases or within the cathepsin B-like proteases but different between the two groups (Table 2). (*i*) The seventh amino acid in the ERFNIN proteases is an invariant Trp, whereas there is a consensus Ala at this position in the cathepsin B-like proteases. (*ii*) Although the three amino acids that precede the active site Gln are structurally similar in the two groups, the specific amino acids are different between the two groups. (*iii*) The active site His is located within a block of hydrophobic amino acids that is interrupted by charged amino acids specific to one group or the other. In the ERFNIN group, the His is preceded by Asp or Asn; in the cathepsin B-like proteins, the third amino acid after the His is Lys or Arg. (*iv*) The amino acid that precedes the active site Asn is a basic Lys in 14 of 15 ERFNIN proteases and an Arg in the last example. There is an invariant nonpolar Ala at this position in the cathepsin B-like proteins.

Table 1. Conserved motif in the propeptide of cysteine proteases

Gene	ERFNIN motif							N–N
Consensus	E X_3	R X_2	(I/V)	F X_2	N X_3	I X_3	N	
Cysteine protease								
Tetrahymena thermophila	–	–	V	–	–	–	–	11
Trypanosoma brucei	–	–	A	–	–	A	A	10
Dictyostelium cp1	–	–	I	–	–	–	–	14
Dictyostelium cp2	–	–	I	–	–	V	–	11
Papaya papain	–	–	I	–	–	–	–	10
Actinidia actinidin	–	–	I	–	T	–	–	11
Vigna mungo	–	–	V	–	–	V	–	10
Barley aleurain	–	–	I	–	S	V	–	10
Lobster cp1	–	–	V	–	–	–	–	14
Lobster cp2	–	–	I	–	–	–	–	14
Lobster cp3	–	–	V	–	–	–	–	14
Mouse cathepsin L	–	–	I	W	–	–	–	14
Rat cathepsin H	–	–	V	–	–	–	–	10
Rat cathepsin L	–	–	V	W	–	–	–	14
Human cathepsin L	–	–	V	W	–	–	–	14
Nonprotease								
Mouse CTLA-2α	–	–	V	W	–	–	–	14
Mouse CTLA-2β	–	–	M	W	–	–	–	14

–, Identity of the amino acid with that of the consensus sequence; N–N, the number of amino acids between the last Asn of the ERFNIN motif and the Asn of the conserved sequence block. References: *Trypanosoma brucei* (17), *D. discoideum* (10), papaya (11), *Actinidia* (18), *V. mungo* (19), barley (20), lobster cp1, cp2, and cp3 (12), mouse (21), rat cathepsin H (16), rat cathepsin L (13), human cathepsin L (22), and mouse CTLA-2α and -β (23).

In the ERFNIN proteases and the cathepsin B-like proteases, there is a motif Gly-Cys-Asn-Gly-Gly (residues 67–71 in *Tetrahymena* and 70–74 in human cathepsin B, Figs. 2 and 3). With the exception of the central Asn residue, this motif is invariant in all of the cysteine proteases examined. In papain, the Cys in this motif is involved in a disulfide bond and is located within a turn. The conservation of this motif between the two families of cysteine proteases suggests that it has an important structural role.

DISCUSSION

Analysis of the deduced amino acid sequence of a cysteine protease gene from *Tetrahymena* suggested that it is translated as a preproenzyme. The putative propeptide region of the protein contains an amino acid motif in which highly conserved amino acids are interspersed with variable ones. This motif was present in all the cysteine protease genes in

Table 2. Consensus sequences of cysteine proteases

Protease	Consensus sequence		
	6 * *	160*	175 * *
ERFNIN protease	**D**WRTKGAVTP...VKN**Q**GQ**C**GS**CW**AFSX$_{19}$SEQNLVDC	L**D**H**G**VLAVGY	**W**LVK**N**S**W**
	|=| =| = || |||||| | = | |	| = =	|||
Cathepsin B	**D**AREQWSNCPTIXIRD**Q**GS**C**GS**CW**AFGX$_{19}$SAEDLLTC	GG**H**AIRILGW	**W**LVA**N**S**W**
	= = === = =	= =	=

Numbers refer to the first amino acid in the sequence according to the *Tetrahymena* numbering. *, Active site amino acids; boldface type, conserved in all cysteine proteases; |, conserved between the consensus sequences; =, conserved within the group and different between the two groups; . . . , gap introduced to align active site residues.

the data base, with the exception of the cathepsin B-like proteases.

One possibility is that the ERFNIN motif in the propeptide serves to inhibit protease activity and that removal of the propeptide converts the protein to the enzymatically active form. Although it has not been demonstrated that the propeptide inhibits enzyme activity of cysteine proteases, proregion peptides of an aspartyl protease and carboxypeptidase A specifically inhibit the respective mature proteases (28, 29).

If the function of the propeptide is inhibition of enzymatic activity, it is not surprising that proteases that lack the ERFNIN motif also show differences in the structure of the enzymatic moiety. It is difficult to identify a region of the propeptide that might serve a similar function for the cathepsin B-like proteases because the number of available sequences and the phylogenetic distribution of the organisms from which they have been obtained are limited.

A search of the data base was done to determine whether the ERFNIN motif was present in proteins other than cysteine proteases. The search uncovered CTLA-2α and CTLA-2β, cDNA clones of mouse RNAs that are specifically expressed in T lymphocytes. The deduced CTLA-2 gene products are small proteins that have striking homology to the cysteine protease propeptides but are not associated with a catalytic moiety (23). Their function is unknown but, because the propeptides of several proteases serve to inhibit protease activity, it was suggested that the CTLA-2 proteins may regulate the activity of unidentified cysteine protease(s).

Analysis of the 20 cysteine protease genes in the EMBL/GenBank data base suggests that they can be divided into two distinct classes. Phylogenetic distribution suggests that the two types of cysteine proteases were established early in evolution. The ERFNIN proteases are found in organisms ranging from protozoa to mammals. It is likely that the cathepsin B-like enzymes evolved before the divergence of the platyhelminthes.

```
Human (Hs) ................................MWQLWASLCCLLVLANARSRPSFHPVS
Mouse (Mm).................................MWWSLILLSCLLALTSAHDKPSFHPLS
Rat (Rn)....................................MWWSLIPLCLLALTSAHDKPSFHPLS
Schistosoma mansoni (Sm) ...............MLTSILCIASLITFLEAHISVKNEKFEPLS
                                                                 = =-=
Haemonchus contortus (Hc) .............MKYLVLALCTYLCSQTGADENAAQGIPLEAQ

         -50                  -30                 -10
          .                    .                   .
Hs      DELVNYVNKR.NTTWQAGHNFYNVDMSYLKRLCGTFLGGPKPPQRVMFTEDLK.....
Mm      DDLINYINKQ.NTTWQAGRNFYNVDISYLKKLCGTVLGGPKLPGRVAFGEDID.....
Rn      DDMINYINKQ.NTTWQAGRNFYNVDISYLKKPCGTVLGGPKLPERVGFSEDIN.....
Sm      DDIISYINEHPNAGWRAEKSNRFHSLDDARIQMGARREEPDLRRKRRPTVDHNDWNVE
        =----=-=-  =- = =- - -   -   -   =-  --=    -    -=
Hc      RLTGEPLVAYLRRSQNLFEVNSAPTPNFEQKIMDIKYKHQKLNLMVKEDPDPEVD...

                10                  30                  50
                 .            *     *.                   .
Hs      LPASFDAREQWPQCPTIKEIRDQGSCGSCWAFGAVEAISDRICIHTNAHVSVEVSAED
Mm      LPETFDAREQWSNCPTIGQIRDQGSCGSCWAFGAVEAISDRTCIHTNGRVNVEVSAED
Rn      LPESFDAREQWSNCPTIAQIRDQGSCGSCWAFGAVEAMSDRICIHTN..VNVEVSAED
Sm      IPSNFDSRKKWPGCKSIATIRDQSRCGSCWSFGAVEAMSDRSCIQSGGKQNVELSAVD
Hc      IPPSYDPRDVWKNCTTFY.IRDQANCGSCWAVSTAAAISDRICIASKAEKQVNISATD

                   70                  90                  110
                    .                   .                   .
Hs      LLTCCGSMCGDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGCRPYSIPPCEHHVNGSR
Mm      LLTCCGIQCGDGCNGGYPSGAWNFWTKKGLVSGGVYDSHIGCLPYTIPPCEHHVNGSR
Rn      LLTCCGIQCGDGCNGGYPSGAGNFWTRKGLVSGGVYNSHIGCLPYTIPPCEHHVNGSR
Sm      LLTCC.ESCGLGCEGGILGPAWDYWVKEGIVTASSKENHTGCEPYPFPKCEHHTKGKY
Hc      IMTCCRPQCGDGCEGGWPIEAWKYFIYDGVVSGGEYLTKDVCRPYPIHPCGHHGNDTY

                     130                 150                 170
                      .                   .                   .
Hs      PPCTGEG.DTPKCSKICEPGYSPTYKQDKHYGYNSYSVSNSEKDIMAEIYKNGPVEGA
Mm      PPCTGEG.DTPRCNKSCEAGYSPSYKEDKHFGYTSYSVSNSVKEIMAEIYKNGPVEGA
Rn      PPCTGEG.DTPNCNKMCEAGYSTSYKEDKHYGYTSYSVSDSEKEIMAEIYKNGPVEGA
Sm      PPCGSKIYNTPRCKQTCQRKYKTPYTQDKHRGKSSYNVKNDEKAIQKEIMKYGPVEAS
Hc      YGECRGTAPTPPCKRKCRPGVRKMYRIDKRYGKDAYIVKQSVKAIWSEILRNGPVVAS

                       190                 210                 230
                        .        *           .        * *        .
Hs      FSVYSDFLLYKSGVYQHVTGEMMGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFF
Mm      FTVFSDFLTYKSGVYKHEAGDMMGGHAIRILVWGVENGVPYWLAANSWNLDWGDNGFF
Rn      FTVFSDFLTYKSGVYKHEAGDVMGGHAIRILGWGIENGVPYWLVANSWNVDWGENGFF
Sm      FTVYEDFLNYKSGIYKHITGEALGGHAIRIIGWGVENKTPYWLIANSWNEDWGENGYF
Hc      FAVYEDFRHYKSGIYKHTAGELRGYHAVKMIGWGNENNTDFWLIANSWHNDWGEKGYF

                          250
                           .
Hs      KILRGQDHCGIESEVVAGIPRTDQYWEKI
Mm      KILRGENHCGIESEIVAGIPRTDQYWGRF
Rn      KILRGENHCGIESEIVAGIPRTQ
Sm      RIVRGRDECSIESEVIAGRIN
Hc      RIIRGTNDCGIEGTIAAGIVDTESL
```

FIG. 3. Derived amino acid sequence for cathepsin B-like cysteine proteases from human (24), mouse (24), rat (25), *Schistosoma* (26), and *Haemonchus* (27). ★, Amino acids in the active site; =, conserved in the propeptide region of the first four proteins; –, similar in the first four proteins where any pair of amino acids within the group have a score of 4 or greater on the scale of Feng *et al.* (14); . , gaps introduced to maximize alignment; boldface type, residues conserved in all known cysteine proteases. The numbers above the human sequence indicate the number of the amino acid with positive numbers for the putative mature protein and negative numbers for the prepro- region.

We thank Kevin Miller and Alan Friedman for assistance with computer analysis. This work was supported in part by Grant GM39890 from the National Institutes of Health and by National Science Foundation Grant MCB-9110651. K.M.K. is a Clare Boothe Luce Professor and gratefully acknowledges the support of the Henry Luce Foundation.

1. Kamphuis, I. G., Kalk, K. H., Swarte, M. B. A. & Drenth, J. (1984) *J. Mol. Biol.* **179,** 233–256.
2. Baker, E. N. (1980) *J. Mol. Biol.* **141,** 441–484.
3. Dufour, E. (1988) *Biochimie* **70,** 1335–1342.
4. Karrer, K. M. & Stein-Gavens, S. (1990) *J. Protozool.* **37,** 409–414.
5. Stargell, L. A., Karrer, K. M. & Gorovsky, M. A. (1990) *Nucleic Acids Res.* **18,** 6637–6639.
6. Horowitz, S. & Gorovsky, M. A. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 2452–2455.
7. Kuchino, Y., Hanyu, N., Tashiro, F. & Nishimura, S. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 4758–4762.
8. Hanyu, N., Kuchino, Y. & Nishimura, S. (1986) *EMBO J.* **5,** 1307–1311.
9. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
10. Pears, C. J., Mahbubani, H. M. & Williams, J. G. (1985) *Nucleic Acids Res.* **13,** 8853–8866.
11. Cohen, L. W., Coghlan, V. M. & Dihel, L. C. (1986) *Gene* **48,** 219–227.
12. Laycock, M. V., MacKay, R. M., Di Fruscio, M. & Gallant, J. W. (1991) *FEBS Lett.* **292,** 115–120.
13. Ishidoh, K., Towatari, T., Imajoh, S., Kawasaki, H., Kominami, E., Katunuma, N. & Suzuki, K. (1987) *FEBS Lett.* **223,** 69–73.
14. Feng, D. F., Johnson, M. S. & Doolittle, R. F. (1985) *J. Mol. Evol.* **21,** 112–125.
15. von Heijne, G. (1986) *Nucleic Acids Res.* **14,** 4683–4690.
16. Ishidoh, K., Imajoh, S., Emori, Y., Ohno, S., Kawasaki, H.,

Minami, Y., Kominami, E., Katunuma, N. & Suzuki, K. (1987) *FEBS Lett.* **226,** 33–37.

17. Mottram, J. C., North, M. J., Barry, J. D. & Coombs, G. H. (1989) *FEBS Lett.* **258,** 211–215.
18. Podivinsky, E., Forster, R. L. S. & Gardner, R. C. (1989) *Nucleic Acids Res.* **17,** 8363.
19. Akasofu, H., Yamauchi, D., Mitsuhashi, W. & Minamikawa, T. (1989) *Nucleic Acids Res.* **17,** 6733.
20. Rogers, J. C., Dean, D. & Heck, G. R. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 6512–6516.
21. Portnoy, D. A., Erickson, A. H., Kochan, J., Ravetch, J. V. & Unkeless, J. C. (1986) *J. Biol. Chem.* **261,** 14697–14703.
22. Gal, S. & Gottesman, M. M. (1988) *Biochem. J.* **253,** 303–306.
23. Denizot, F., Brunet, J.-F., Roustan, P., Harper, K., Suzan, M., Luciani, M.-F., Mattei, M.-G. & Golstein, P. (1989) *Eur. J. Immunol.* **19,** 631–635.
24. Chan, S. J., San Segundo, B., McCormick, M. B. & Steiner, D. F. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 7721–7725.
25. San Segundo, B., Chan, S. J. & Steiner, D. F. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 2320–2324.
26. Klinkert, M.-Q., Felleisen, R., Link, G., Ruppel, A. & Beck, E. (1989) *Mol. Biochem. Parasitol.* **33,** 113–122.
27. Cox, G. N., Pratt, D., Hageman, R. & Boisvenue, R. J. (1990) *Mol. Biochem. Parasitol.* **41,** 25–34.
28. San Segundo, B., Martinez, M. C., Vilanova, M., Cuchillo, C. M. & Avilès, F. X. (1982) *Biochim. Biophys. Acta* **707,** 74–80.
29. Evin, G., Devin, J., Castro, B., Menard, J. & Corvol, P. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 48–52.