Volume 258, number 2, 211–215 FEB 07883 December 1989

A cysteine proteinase cDNA from *Trypanosoma brucei* predicts an enzyme with an unusual C-terminal extension

Jeremy C. Mottram, Michael J. North°, J. David Barry[1] and Graham H. Coombs*

Wellcome Unit of Molecular Parasitology, Institute of Genetics, University of Glasgow, Church Street, Glasgow G11 5JS, °School of Molecular and Biological Sciences, University of Stirling, Stirling FK9 4LA and **Department of Zoology, University of Glasgow, Glasgow G12 8QQ, Scotland*

Received 7 September 1989, revised version received 26 September 1989

A cDNA for a *Trypanosoma brucei* cysteine proteinase has been cloned and sequenced The deduced protein can be divided into four domains, based on homologies with other cysteine proteinases the pre-, pro- and central regions show considerable homology to the cathepsin L class of mammalian enzymes, whilst the long C-terminal extension distinguishes the trypanosome enzyme from all mammalian cysteine proteinases reported This 108 amino acid extension, which includes 9 contiguous prolines near the junction with the central domain, appears likely to be processed in part to produce the mature enzyme, and may be involved in targetting the protein within the cell The trypanosome genome contains more than 20 copies of the cysteine proteinase gene arranged in a long tandem array

Trypanosome, Cysteine proteinase

1. INTRODUCTION

The cysteine proteinases (EC 3.4.22) are proving a valuable group for investigating the relationship between enzyme structure and function. Mammalian lysosomal cathepsins (B, H, L) and plant proteinases such as papain have been the most extensively studied, but similar enzymes occur widely in invertebrates and protozoa [1]. Notably, they occur at high activity in a number of parasitic protozoa, many of which contain multiple enzymes that in some species are developmentally regulated [2,3]. The abundance of cysteine proteinases in a variety of parasites, their potential role in the host-parasite interaction and virulence, and the advent of families of specific irreversible inhibitors [4], have made these enzymes attractive targets for chemotherapeutic attack.

2. EXPERIMENTAL

Trypanosome DNA was isolated, digested with restriction endonucleases and blotted as described previously [5] Oligonucleotide labelling and hybridisations were carried out at 42°C for 24 h, washed in 5 × SSC (1 × SSC = 150 mM NaCl, 15 mM Na citrate), 0.1% SDS at 55°C for 1 h and autoradiographed [6] Hybridisation of the cDNA clone pTCP-F1 was performed at 55°C (fig.2C) or 65°C (fig 2D) as above and washed in 2 × SSC with 0 1% SDS at 55°C for 1 h (C) or 0.1 × SSC with 0.1% SDS at 65°C for 1 h (D) The sequence of the oligonucleotide probes were

Probe O-428. 5′ GC CCA PCA GCC GCC PCA YTG GCC YTG PTC YTT 3′

Probe O-429. 5′ AT GXA GCC PTC YTC GCC CCA YTG NGC NGT CCA 3′

where P = G or A, Y = T or C, X = T or A and N = G, A, T or C

A *Trypanosoma brucei* λgt11 cDNA library [5] was screened with oligonucleotide probes O-428 and O-429 and positive cDNAs subcloned into Bluescript plasmid for mapping and sequencing according to manufacturer's instructions (Stratagene)

Correspondence address. J C Mottram, Wellcome Unit of Molecular Parasitology, Institute of Genetics, University of Glasgow, Church Street, Glasgow G11 5JS, Scotland

These sequence data will appear in the EMBL nucleotide sequence database under the accession number X16465

3. RESULTS AND DISCUSSION

Two degenerate DNA oligonucleotides were designed from two sets of data: the amino acid sequence of 2 conserved regions of eukaryotic cysteine proteinases, and the published partial amino acid sequence of a *T. cruzi* cysteine proteinase (fig.1). When used as probes on Southern blots of genomic DNA of *T. brucei* and *T. cruzi* (fig.2), both oligonucleotides recognised a single 1.75 kbase (kb) *Hin*cII band (fig.2A and B, lanes 1 and 4), a >23 kb doublet for the *T. brucei Bam*HI digest (lane 6) and a single >23 kb band for *T. cruzi* (lane 3). A *T. brucei* cDNA library in λgt11 was also screened and several clones were isolated. The largest insert was subcloned into Bluescript plasmid (pTCP-F1). On genomic Southerns, the cDNA hybridised to the same fragments as the oligonucleotides for the *Hin*cII and *Bam*HI digests, with an extra band present in the *Pst*I digest (fig.2C). Although the oligonucleotides hybridised to DNA of both trypanosomes with equal intensity, the *T. brucei* cDNA hybridised only weakly to the *T.*

Published by Elsevier Science Publishers B V (Biomedical Division)
00145793/89/$3 50

	20 25	195 200 205	Refs
T. cruzi	K D Q G Q C G G C W A	? N S W T A Q W G E D G Y I	[7]
CP1	- N - - - - - S - - S	K - - - G A D - - - Q - - -	[18]
Actinidin	- S - - E - - - - - -	K - - - D T T - - - E - - F	[14]
Papain	- N - - S - - S - - -	K - - - G T G - - - N - - -	[19]
Cathepsin Lh	- N - - - - - S - - -	K - - - G E E - - M G - - V	[21]
consensus	K Q G C G C W *	K N S W W G G Y	

Fig 1 Conserved regions of cysteine proteinases used for the design of the oligonucleotides The numbers are the amino acid positions for *Dictyostelium* CP1 *, active site cysteine, –, residue identical with that in *T cruzi*

cruzi DNA even at low stringency (fig.2C, lanes 1–3). Clearly, outside the conserved regions recognised by the oligonucleotides there is considerable divergence between the two trypanosome proteinases.

The first 22 nucleotides (nts) of the cDNA pTCP-F1 (fig.3) are identical to the spliced leader [8,9], a sequence that is thought to be present on the 5′ end of all trypanosome mRNAs [10]. The first ATG, 32 nts 3′ of the splice junction, has been taken to be the initiator methionine, although there is a second in-frame ATG a further 15 nts downstream. The open reading frame predicts a protein of 450 amino acids and 48.4 kDa which is highly homologous to the cathepsin L class of cysteine proteinases (fig.4).

The organisation of the cysteine proteinase gene was investigated by genomic Southerns of *T. brucei* DNA digested with *Hind*111 (fig.2D). Complete digests showed a 1.75 kb band and a 3.75 kb band (arrowed, track 11, fig.2D) which hybridised only weakly. Partial digests revealed a ladder of bands increasing in 1.75 kb units, 10 of which could be resolved by agarose gel electrophoresis, clearly indicative of a tandem repeat. Lambda clones isolated from an EMBL 4 library showed that the 3.75 kb band contained one copy of the cysteine proteinase gene located at the 3′ flank of the repeat unit. Comparison of the 1.75 and 3.75 kb bands by densitometric scans indicated that there are more than 20 copies of cysteine proteinase genes repeated in tandem. It is not known whether all copies are transcribed or if some are pseudogenes. As enzyme studies have suggested that there are only a few isoenzymes (unpublished), it seems likely that there are multiple copies of essentially the same gene. The reason for the large number of cysteine proteinase genes is not known; the enzyme is not present at unusually high activity and only slight differences have been detected

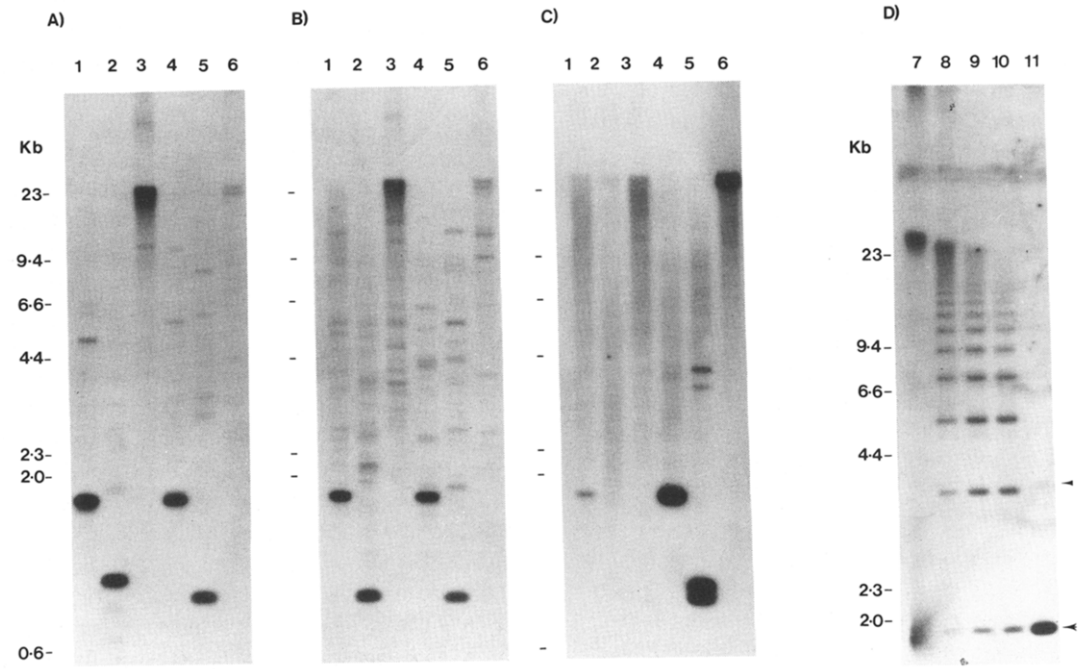

Fig 2 Southern blot analysis of trypanosome cysteine proteinase genes *T cruzi* (lanes 1–3) and *T brucei* (lanes 4–6) genomic DNA was probed with Oligo 428 (A), Oligo 429 (B) and cDNA clone pTCP-F1 (C,D) Restriction enzymes *Hinc*11 (lanes 1 and 4), *Eco*R1 (lane 2), *Bam*H1 (lanes 3 and 6), *Pst*1 (lane 5) (D) Partial digest of *T brucei* genomic DNA with *Hind*111 probed with pTCP-F1 Lane 7, uncut, lanes 8–10, partial digests at time points (in min) 5, 45 and 120, lane 11, complete digest

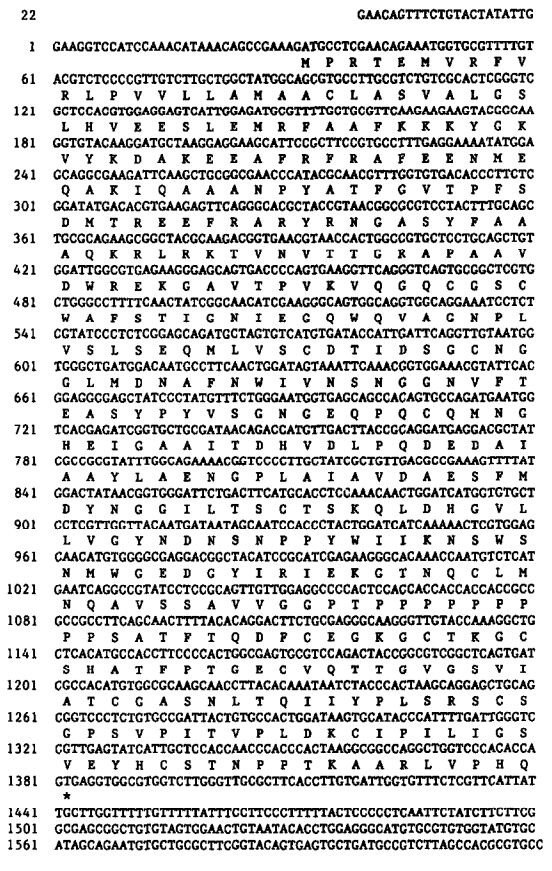

Fig.3 Nucleotide sequence of a *T brucei* cysteine proteinase cDNA with its predicted amino acid sequence As the first 22 nts of the cDNA clone TCP-F1 are identical to the spliced leader sequence [8,9], the nucleotides are numbered from the left starting at the splice junction.

between the activities in different stages of the parasite life-cycle [11,12].

Amino acid sequence comparison shows human cathepsin L to be the most similar of the mammalian cysteine proteinases to the trypanosomal protein, although the *Dictyostelium* CP1 product has greater homology with the parasite enzyme (fig.4, table 1). The trypanosomal sequence can be considered to be made up of four sections. At the N-terminus there is a 20-residue, hydrophobic pre-sequence and a 105-residue, hydrophilic pro-sequence typical of those present in other cysteine proteinases. On the basis of the known N-terminal sequence of the *T. cruzi* cysteine proteinase [7], and by analogy with the majority of other cysteine proteinases, the N-terminal residue of the mature *T. brucei* proteinase is likely to be the alanine residue at position 126. The 217-residue central domain of the trypanosomal protein corresponds to the mature forms of papain, actinidin and mammalian cathepsins L and H. All of the highly conserved residues are present, including residues involved in the catalytic mechanism of papain and the six cysteine residues which form the three conserved disulphide bridges. The most surprising feature is a 108-residue extension at the C-terminal end which has few equivalents in other cysteine proteinases. Much shorter extensions (6 and 25 amino acids) that are processed to produce the mature enzymes occur with cathepsin B [13] and actinidin [14], respectively, but the only sequence with which the trypanosomal gene shows clear homology (22%) is that of a mRNA induced in tomatoes by low temperature [15]. The tomato and trypanosomal sequences are most closely related at the start of the extension. In this section the trypanosomal protein has a remarkable sequence made up of nine consecutive prolines. The only homology detected in a search of the NBRF protein sequence data base is with the proline-rich regions of the gag polyprotein of T-cell leukemia virus (HTLV-II) and the human proline-rich peptide P-B [16].

It seems likely that part of the C-terminal extension remains in the final trypanosomal proteinase. The only cysteine proteinases readily detectable in *T. brucei* extracts using gelatin-SDS-PAGE and fluorogenic substrates appear not to be glycosylated and have apparent molecular masses of approximately 28 and 31 kDa ([12], Robertson et al., unpublished). The predicted total molecular mass of the trypanosomal protein is 48.4 kDa, the different regions being pre-, 2.3 kDa; pro-, 11.9 kDa; central, 23.1 kDa and C-terminal extension, 11.1 kDa. As the pro-region is probably removed, as in *T. cruzi*, the prediction is that approximately 50 residues of the *T. brucei* C-terminal extension must be present in the mature enzyme. No information is yet available on the fate of the C-terminal extension of the tomato cysteine proteinase.

The functional significance of the C-terminal extension is unknown. The targetting of other trypanosomatid enzymes to glycosomes may in some cases be mediated by a C-terminal extension [17], which however has no apparent sequence homology with that of the cysteine proteinase. Current evidence suggests that like most other cysteine proteinases, those of *T. brucei* are lysosomal [12] and, as suggested by the presence of a hydrophobic signal sequence, would be synthesised on membrane-bound ribosomes. Glycosomal proteins are synthesised on free ribosomes. The mechanism of targetting proteins to lysosomes in trypanosomatids is yet to be elucidated and it will be interesting to see whether C-terminal extensions are common features of many lysosomal enzymes in this group of organisms or a specific character of cysteine proteinases. We are currently attempting to discover how the unusual feature of the parasite enzyme is reflected in its structure and activity. The results of these studies should provide a greater insight into the factors that influence the catalytic activity of this family of enzymes and identify possible new approaches to the design of drugs to exploit the peculiarities of the parasite enzyme.

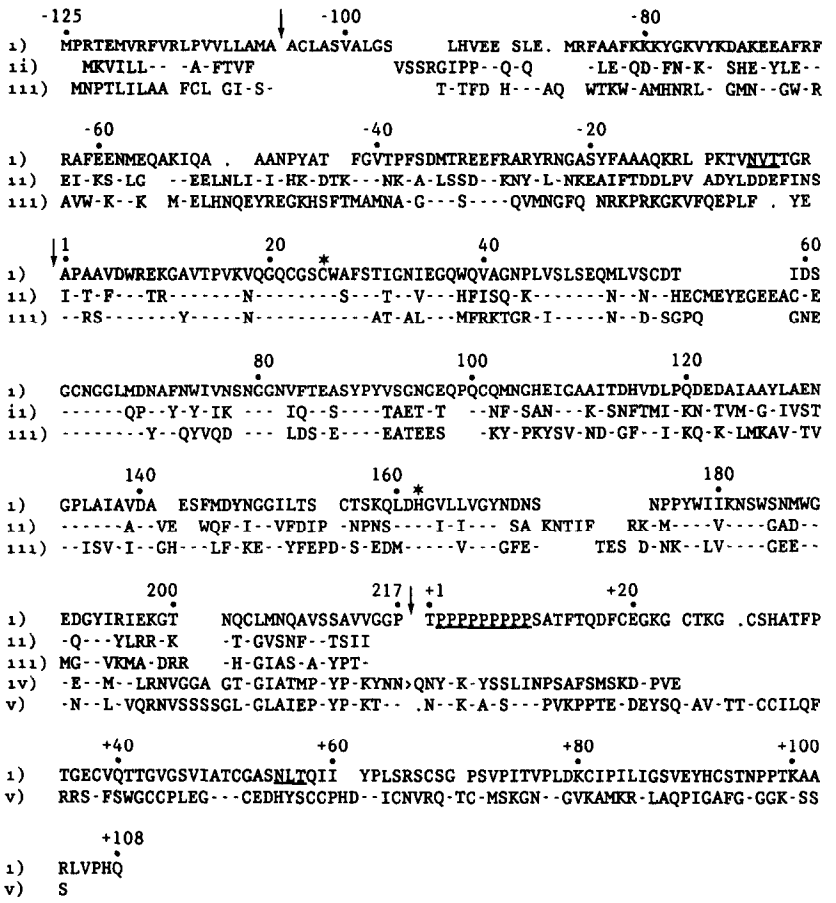

Fig.4. Predicted amino acid sequence comparisons for cysteine proteinases. (i) *T. brucei* (this paper); (ii) *Dictyostelium discoideum* CP1 [18]; (iii) human cathepsin L [21], (iv) actinidin [14], (v) tomato cold-induced proteinase [15]. Only the C-terminal portions of actinidin and the tomato proteinase are given. The sequences were aligned by eye, taking into account other published cysteine proteinase sequences not included here Some of these other sequences contain additional residues to those in the sequences shown. Gaps have been included to maximise homology, these are indicated by * All numbers refer to the trypanosomal sequence: – refers to the residues in the putative prepro-region, + refers to residues in the putative C-terminal extension, the remaining residues comprise the central domain corresponding to the mature form of other cysteine proteinases. The residues shown as – are identical to those in the trypanosomal sequence; *, active site cysteine and histidine residues, >, C-terminus of actinidin protein, ↓ boundary between domains Putative N-linked glycosylation sites in the trypanosomal sequence and the oligoproline sequence are underlined

Table 1

Homology between the trypanosomal proteinase and other cysteine proteinases

Proteinase	Percentage identity				
	Pre	Pro	Central	C-terminal	Reference
Dictyostelium CP1	20.0 (26.7)	25.7 (26 7)	51.6 (49 6)	none present	18
Papain	15 0 (13.0)	24.8 (23 6)	37.8 (38.7)	none present	19
Actinidin	[a]	[a]	38.2 (37 7)	2.8 (12.0)	14
Tomato	[a]	[a]	40.1 (39.5)	22.2 (22.0)	15
Cathepsin H (rat)	20.0 (21 1)	23.8 (26.6)	38.7 (38 2)	none present	20
Cathepsin L (human)	10 0 (11.8)	20.0 (21 8)	48.4 (47 7)	none present	21
Cathepsin B (human)	15.0 (17 6)	4 8 (11.3)	26.7 (23.0)	0 (0)	13

[a] Complete sequence not yet available

The figures represent the percentage of residues in each section of the trypanosomal protein which are present at the identical position in the other proteins Those in brackets are the percentage of residues in the other proteins present in the trypanosomal protein

Acknowledgements· We thank Dr C.M.R. Turner for providing the trypanosomes and V Graham for technical assistance. This work was supported by the Wellcome Trust. J D.B. is a Wellcome Trust Senior Lecturer

REFERENCES

[1] North, M J. (1982) Microbiol. Rev. 46, 308–340
[2] Lockwood, B C., North, M.J , Mallinson, D J. and Coombs, G.H. (1987) FEMS Microbiol. Lett 48, 345–350.
[3] Pamer, E.G , So, M. and Davis, C E (1989) Mol. Biochem Parasitol 33, 27–32.
[4] Zumbrunn, A , Stone, S. and Shaw, E. (1988) Biochem. J. 250, 621–623.
[5] Mottram, J.C., Murphy, W J. and Agabian, N (1989) Mol Biochem Parasitol., in press
[6] Mottram, J.C., Perry, K L , Lizardi, P.M , Luhrmann, R , Agabian, N and Nelson, R.G (1989) Mol. Cell Biol 9, 1212–1223
[7] Cazzulo, J.J., Couso, R , Raimondi, A., Wernstedt, C and Hellman, U (1989) Mol. Biochem. Parasitol 33, 33–42.
[8] Van der Ploeg, L.H.T , Liu, A Y C , Michels, P.A.M , De Lange, T , Borst, P., Majunder, H K , Weber, H., Veeneman, G.H and Van Boom, J (1982) Nucleic Acids Res 10, 3591–3604
[9] Boothroyd, J C. and Cross, G.A (1982) Gene 20, 281–289.
[10] Walder, J.A , Eder, P S , Engman, D M , Brentano, S T., Walder, R Y., Knutzon, D.S , Dorfman, D.M and Donelson, J E (1986) Science 233, 569–571.
[11] North, M.J., Coombs, G.H and Barry, J D (1983) Mol Biochem. Parasitol. 9, 161–180.
[12] Lonsdale-Eccles, J D and Grab, D.J (1987) Eur J Biochem 169, 467–475.
[13] Chan, S J., San Segundo, B., McCormick, M.B. and Steiner, D F (1986) Proc. Natl. Acad Sci. USA 83, 7721–7725.
[14] Praekelt, V.M., McKee, R.A and Smith, H (1988) Plant Mol Biol. 10, 193–202
[15] Schaffer, M A and Fischer, R L (1988) Plant Physiol 87, 431–436
[16] Isemura, S., Saitoh, E. and Sarada, K.J. (1979) Biochemistry 86, 79–86
[17] Swinkels, B W , Evers, R. and Borst, P. (1988) EMBO J 7, 1159–1165.
[18] Williams, J G , North, M J and Mahbubani, H (1985) EMBO J. 4, 999–1006.
[19] Cohen, L W., Coghlan, V M and Dihel, L C. (1986) Gene 48, 219–227
[20] Ishidoh, K., Imajoh, S., Emori, Y , Ohno, S , Kawasaki, H , Minami, Y , Kominami, E., Katunuma, N and Suzuki, K (1987) FEBS Lett 226, 33–37
[21] Gal, S and Gottesman, M M. (1988) Biochem J 253, 303–306