Amplification and sequencing of genomic DNA fragments encoding cysteine proteases from protozoan parasites

Ann E. Eakin¹,², Jacques Bouvier¹, Judy A. Sakanari¹, Charles S. Craik² and James H. McKerrow¹

Departments of Pathology¹ and Pharmaceutical Chemistry², University of California, San Francisco, CA, U.S.A.

(Received 20 June 1989; accepted 18 September 1989)

Cysteine protease gene fragments from three protozoan parasites Trypanosoma cruzi, Trypanosoma brucei, and Entamoeba histolytica were amplified by the polymerase chain reaction (PCR) from genomic DNA using degenerate oligonucleotide primers. The primers used for the amplification were designed based upon amino acid sequences flanking the active site cysteine and asparagine residues that are conserved in the eukaryotic cysteine proteases analyzed to date. The amplified DNA fragments, representing approximately 70% of the coding regions of the cysteine protease genes, were subcloned and sequenced. Sequence analysis and alignment showed significant sequence similarity to other members of the eukaryotic cysteine protease family (45% identical to chicken cathepsin L) and conservation of the cysteine, histidine, and asparagine residues which form the catalytic triad. These gene fragments provide molecular probes for further analysis of the structure and function of these important metabolic enzymes.

Key words: Trypanosoma cruzi; Trypanosoma brucei; Entamoeba histolytica; Cysteine protease; Polymerase chain reaction

Introduction

Cysteine (thiol) proteases are widely distributed in nature. They have been identified in plants [1], animals [2], viruses [3], bacteria [4] and eukaryotic microorganisms [5]. While there are structural variations among the cysteine proteases of different phyla, all are endopeptidases that range in molecular weight from 15,000–80,000 and have a critical cysteine residue at the active site involved in the hydrolysis of the substrate. Cysteine proteases have been identified in a number of protozoan parasites, including Trypanosoma cruzi [6,7], Trypanosoma brucei [8,9], Leishmania mexicana [10], Leishmania tarentolae [9], Crithidia fasciculata [9], Entamoeba histolytica [11–14], and Plasmodium falciparum [15] and in several helminth parasites such as Schistosoma mansoni [16], Paragonimus westermani [18] and Fasciola hepatica [19].

Sakanari et al. [20] showed that conserved structural motifs, identified by alignment of several members of the serine protease family, can be used to design generic molecular probes for amplification of serine protease gene fragments from the DNA of parasitic organisms using the polymerase chain reaction (PCR). For the present study, we have used the same strategy to amplify fragments of genes encoding cysteine proteases in parasitic protozoa. In theory, this technique should allow amplification of a fragment of virtually any eukaryotic cysteine protease gene directly from a minute amount of cDNA or genomic DNA. This approach circumvents the lack of an abundant supply of organisms and consequent difficulties in protein purification. Our goal has been to determine the
primary structure of the cysteine proteases of *T. cruzi*, *T. brucei* and *E. histolytica*. This is a critical first step in characterizing these enzymes and defining their function.

Based upon conserved amino acid sequences in the regions flanking the active site cysteine-25 and asparagine-175 residues [21], two sets of degenerate oligonucleotides were synthesized. These primers were used in the PCR to amplify protease gene fragments from genomic DNA of the parasites. The PCR-amplified fragments were then subcloned and sequenced. In addition to sequence data obtained directly, these fragments provide homologous molecular probes with which the entire gene sequences encoding these cysteine proteases and related members of the family may be readily isolated. Isolation and expression of these genes will provide additional information on this class of enzymes and will facilitate analysis of their structure, function, and molecular evolution.

**Materials and Methods**

*Isolation of parasite DNA*. Genomic DNA from *T. cruzi* epimastigotes and *T. brucei* trypomastigotes was prepared as described by Castro et al. [22]. Cultures of *E. histolytica* were maintained as described by Keene et al. [11]. Genomic DNA from *E. histolytica* was isolated from 10^7* trophozoites. Washed cells were suspended in 1 ml of 10 mM Tris-HCl, pH 8.5, 50 mM EDTA, and lysed by the addition of SDS to 0.2%. Debris was removed by centrifugation at 13,000 g for 10 min. The DNA was purified using the Geneclean™ kit (Bio 101, La Jolla, CA), ethanol-precipitated, and resuspended in water to 100 µg ml^-1. The Abs 260/280 ratio was 1.8.

**Construction of oligonucleotide primers.** Design of the oligonucleotide primers used to amplify the cysteine protease gene fragments (Fig. 1) was based upon the consensus sequences flanking the active site cysteine (position 25) and the active site asparagine (position 175) of a cysteine protease of the slime mold *Dictyostelium discoideum* [23]. Mixed oligonucleotides were synthesized, using inosines in positions where all four bases were possible in the codon, in order to minimize the degeneracy of the DNA primer, but maximize its ability to form stable hybrids with the target DNA. The degeneracy of the oligonucleotides was 1024 for the cysteine primer and 384 for the asparagine primer. The cysteine primer was synthesized as the sense strand primer and the asparagine primer was synthesized as the anti-sense strand primer (Fig. 1). Recognition sites for the restriction endonucleases *EcoRI* and *HindIII* were added to the 5' ends of each primer to allow rapid subcloning in a known orientation for double-stranded DNA sequencing. Three additional bases (ACA in the cysteine primer and TTA in the asparagine primer) were added to the 5' ends to insure polymerization through the restriction sites. Oligonucleotides were synthesized using a betacyanoethyl phosphoramidite procedure and purified by FPLC anion exchange chromatography (Operon, San Pablo, CA).

**Polymerase chain reaction.** 200 ng of genomic DNA were used per 50 µl reaction volume and amplified in 60 cycles of the polymerase chain reaction with primers annealed at 25°C, as previously described [20].

**Isolation and sequencing of gene fragments.** Amplified fragments were resolved in 4% agarose gels

**Cys 5' Primer:**

5' ACA GAA TTC CAR GGI CAR TGY GGI TCI TGY TGG 3'  
EcoRI   gln gly gln cys gly ser  cys  trp  
R = A or G  
Y = T or C  
I = inosine

**Asn 3' Primer:**

5' TTA AAG CTT CCA 1GA RTT YTT IAC RAT CCA RTA 3'  
HindIII trp ser  asn lys val ile trp tyr

Fig. 1. The oligonucleotide primers constructed in order to amplify the cysteine protease gene fragments. Details of their construction are given in Materials and Methods.
(3% Nu Sieve® GTG/1% SeaKem® GTG; FMC Rockland, ME). Fragments of expected length (400–600 bp) were excised from the gel, purified using ‘glassmilk’ (Geneclean™), digested with EcoRI and HindIII, extracted once with phenol/chloroform, ethanol-precipitated, and subcloned in a Bluescript™ vector (Stratagene, La Jolla, CA). Plasmid DNA was isolated using the boiling-lysis method [24] and sequenced directly using M13 forward and reverse primers and the Sequenase™ Kit (United States Biochemical, Cleveland, OH).

Amino acid sequences of the protease gene fragments were deduced from the nucleic acid sequences and analyzed using the Sequence Analysis Package designed and provided by the Computer Graphics Laboratory, Department of Biochemistry and Biophysics, UCSF.

Results

Fig. 2 shows the results of the polymerase chain reaction using the generic cysteine protease primers with genomic DNA from T. cruzi, T. brucei and E. histolytica. Ethidium bromide-stained gels showed 5 bands from T. cruzi, 5 bands from T. brucei, and 7 bands from E. histolytica. Based on published cysteine protease sequences, and assuming there were no introns (which has been the case in protozoans thus far), the expected size of the cysteine protease gene fragments was between 400 and 600 bp in length.

Two fragments were excised and subcloned into Bluescript™ from both T. cruzi (400 and 500 bp) and T. brucei (500 and 600 bp). Sequence analysis indicated that the 400-bp fragment of T. cruzi and the 600-bp fragment of T. brucei were identical to their respective 500-bp fragments, suggesting that subsequent digestion and subcloning of the fragments eliminated the size difference. Three fragments from E. histolytica of 350, 400, and 450 bp were subcloned into Bluescript™ and sequenced. Only the 450-bp fragment contained sequences of the original primers and showed sequence similarity with other cysteine proteases. The sequences of the other fragments were unrelated to known cysteine protease sequences, and therefore were probably the result of non-specific primer amplification due to the low stringency of the DNA hybridization conditions in the PCR. Since the degeneracy of the PCR primers can result in several amplified DNA fragments that do not originate from the target gene, we routinely use predicted size to discriminate among the amplified products.

Fig. 3 shows the alignment of the predicted amino acid sequences of these gene fragments with each other and with a spectrum of eukaryotic cysteine proteases including the ‘Sm31’ protease from the trematode S. mansoni [25], papain from papaya [21], the cysteine protease of D. discoideum [23], and chicken cathepsin L [26].

![Fig. 2](image_url)  
**Fig. 2.** Analysis of gene fragments amplified by the polymerase chain reaction using generic cysteine protease primers with genomic DNA from T. brucei (lane 2), T. cruzi (lane 3), and E. histolytica (lane 4). 16% (8 μl) of total PCR reaction volume (50 μl) was analyzed by 4% agarose-TAE gel electrophoresis using 1-kb ladder DNA markers (BRL, Bethesda, MD; lane 1). Cysteine protease gene fragments migrated at 500 bp in T. cruzi and T. brucei and at 450 base pairs in E. histolytica.
Fig. 3. Alignment of predicted amino acid sequences from PCR-amplified gene fragments, encoding protozoan cysteine proteases, with published amino acid sequences of representative eukaryotic enzymes [23, 25, 26]. Also included in the alignment are the previously published partial amino acid sequence of a T. cruzi cysteine protease [28] and the partial genomic sequence of a T. cruzi clone isolated from a λgt11 bacteriophage library using the PCR-amplified gene fragment as a probe. Amino acid residues representing the catalytic triad are indicated in bold type. Non-conserved loops present in S. mansoni and D. discoideum were omitted to conform to the papain numbering system [21], so that a = ECMEYEGEEAC, b = KENHTGCEYPFPKCEH, c = VMAGYIVSTGPL, and d = FRKNMP.

gions of identity with other eukaryotic cysteine proteases could be observed around amino acids representing the active site residues as well as other structural motifs common to one or more of the other cysteine proteases. Most notably, the histidine at position 159, which is required for ion pair formation [27] with the active site cysteine, is present in all three sequences. Also, the glycine at position 66, which is involved in substrate binding in papain, is present in all three protozoan sequences.

Computer analysis of sequence identity (Table I) showed that the T. cruzi and T. brucei protease gene fragments are 64% identical to each other at the amino acid level but are less similar to the E. histolytica gene fragment - 35% and 36%, respectively. All three of these proteases demonstrate greater similarity to chicken cysteepsin L (45%) than to any other protein represented in the Dayhoff (National Biomedical Research Foundation) protein sequence library. The percent identity values were calculated as the number of identical amino acids per total number of amino acids, allowing gaps in the sequence to achieve the best alignments.


### TABLE I

Percent identity of predicted amino acid sequences of protozoan cysteine protease gene fragments with corresponding regions of representative eukaryotic enzymes

<table>
<thead>
<tr>
<th></th>
<th><em>T. cruzi</em></th>
<th><em>T. brucei</em></th>
<th><em>E. histolytica</em></th>
<th><em>Cathepsin L. chick</em></th>
<th><em>Dictyostelium</em></th>
<th><em>Cathepsin B. rat</em></th>
<th><em>Cathepsin B. human</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>T. cruzi</em></td>
<td>–</td>
<td>64%</td>
<td>35%</td>
<td>45%</td>
<td>41%</td>
<td>34%</td>
<td>30%</td>
</tr>
<tr>
<td><em>T. brucei</em></td>
<td>64%</td>
<td>–</td>
<td>36%</td>
<td>45%</td>
<td>41%</td>
<td>38%</td>
<td>34%</td>
</tr>
<tr>
<td><em>E. histolytica</em></td>
<td>35%</td>
<td>36%</td>
<td>–</td>
<td>45%</td>
<td>41%</td>
<td>31%</td>
<td>29%</td>
</tr>
<tr>
<td>Cathepsin L. chick [26]</td>
<td>45%</td>
<td>45%</td>
<td>45%</td>
<td>–</td>
<td>63%</td>
<td>32%</td>
<td>32%</td>
</tr>
<tr>
<td><em>D. discoideum</em> [23]</td>
<td>41%</td>
<td>41%</td>
<td>41%</td>
<td>63%</td>
<td>–</td>
<td>22%</td>
<td>20%</td>
</tr>
<tr>
<td>Cathepsin B. rat [34]</td>
<td>34%</td>
<td>38%</td>
<td>31%</td>
<td>32%</td>
<td>22%</td>
<td>–</td>
<td>83%</td>
</tr>
<tr>
<td>Cathepsin B. human [35]</td>
<td>30%</td>
<td>34%</td>
<td>29%</td>
<td>32%</td>
<td>20%</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

**Discussion**

We have shown that molecular probes based upon consensus sequences of eukaryotic cysteine proteases can be used to isolate gene fragments from related enzymes of three protozoan parasites. Alignment of the predicted amino acid sequences shows significant sequence identity with other members of the eukaryotic cysteine protease family and conservation of the catalytic triad.

Recently, Cazzulo et al. [28] described the isolation and further characterization of a cysteine protease from the epimastigote stage of *T. cruzi*. They determined the sequence of 32 amino acids of the N-terminus of the protein and 14 amino acids of an internal trypsin-generated peptide. The cysteine protease gene fragment we have isolated from *T. cruzi* can be identified as the gene encoding the same protease purified by Cazzulo et al. as demonstrated by the overlap of our predicted amino acid sequence with their N-terminal sequence (Fig. 3). Together with the data of Cazzulo et al. [28], the amino acid sequence of the major portion of this protein is now known.

In order to extend our molecular analysis of the cysteine protease of *T. cruzi*, we used the PCR-amplified fragment as a homologous molecular probe to screen a genomic library (Agt11) for the entire gene. Preliminary sequence analysis of the cloned gene confirms the amino acid sequence of the original primer used for PCR and also confirms the published N-terminal sequence [28], including the identification of the ambiguous residues. The N-terminal sequence contained a glycine at position 27 [28], while our clone encodes a serine residue which is conserved in published cysteine protease sequences.

The preliminary sequence data from the cloned gene also allowed us to assess the accuracy of the sequence data derived directly from the PCR-amplified gene fragment. Because the Taq polymerase lacks 3' exonuclease activity, there is no mechanism for editing of misincorporated bases generated during PCR. Nevertheless, Innis et al. [29] reported that when using 50–200 μM of each dNTP, the accuracy of Taq DNA polymerase is quite high (one error in 4000–5000 bp sequenced after 35 cycles of PCR) and is comparable with the accuracy of other DNA polymerases. To maximize the accuracy of our sequences we used 200 μM of each dNTP. The sequence data from the cloned gene differs from our PCR-amplified sequence by only one base, which does not change the amino acid sequence. This difference may be due to misincorporation by the Taq DNA polymerase during the PCR or, perhaps more likely, be due to the occurrence of a silent mutation between the DNA sequences of the two different strains of *T. cruzi*.

The localization, function and substrate specificity of the enzyme from *T. cruzi* have yet to be fully characterized. Some investigators have suggested that this protease may be localized within the lysosome-like vacuoles of the epimastigote stage of the parasites and may play a role in the degradation of cellular components [30]. Alternatively, the protease may be associated with the cell-surface membrane and may play a role in destruction of host tissue or evasion of the host cell response [31].

Londsdale-Eccles and Mpimbaza [8] identified
cysteine protease activity in several species of African trypanosomes. Maximal enzyme activity was associated with the metacyclic and bloodstream forms of the parasites but activity was also detected at lower levels in the procyclic insect form. Pamer et al. [32] recently reported the presence of a developmentally regulated cysteine protease of T. brucei having the same molecular weight as that reported by Lonsdale-Eccles and Mpmibaza [8]. This cysteine protease appeared to be prevalent in the short stumpy form of the parasite and less prevalent in the long slender stage found in the mammalian bloodstream, suggesting that changes in its expression may be associated with developmental differentiation. Our isolation of a cysteine protease gene fragment from T. brucei will allow more direct analysis of stage-specific expression and regulation.

Cysteine protease activities have been detected in trophozoites of virulent laboratory strains of E. histolytica by several investigators [11–13]. Two cysteine proteases of 56 and 27 kDa have been purified from E. histolytica of the HM-1 strain [11,12]. The 56-kDa E. histolytica protease has been shown to degrade a number of host macromolecules including collagen, laminin, and fibrinectin and appears to be the primary factor in producing the cytopathic effect of virulent strains [11]. It has also been identified in clinical isolates of E. histolytica from patients with colitis or extraintestinal amebic disease but not in amebae isolated from patients with asymptomatic or mild infections [33].

The enzymological properties of the 56 kDa and 27 kDa proteases are very similar [11,12], but it is unclear whether the larger species represents a pro-enzyme form, a different gene product, or a dimer of the smaller species. We will now be able to address this issue using the cysteine protease gene fragment we have isolated.

The cysteine proteases predicted by the gene fragments of T. brucei and T. cruzi are closely related, with 64% amino acid identity, and both are more similar to chicken cathepsin L (45%) than to other known cysteine protease sequences (Table 1). The cysteine protease predicted by the gene fragment of E. histolytica is also similar to chicken cathepsin L (45%) but is clearly divergent from the trypanosome proteases (<36% identity).

Amino acid conservation among this family of proteases may indicate functional importance of certain residues. Therefore, analysis of these conserved residues may provide additional information about the general mechanism of cysteine protease activity as well as how the proteases from parasites may differ in their structural and functional properties from proteases of higher eukaryotes.

The DNA sequences presented in this study were generated from 200 ng of genomic DNA. By virtue of the engineered restriction sites in the oligonucleotide primers, the PCR-amplified sequences coding for the cysteine protease genes were subcloned and sequenced within one week. By analogy with the size of known eukaryotic cysteine proteases, these gene fragments represent approximately 70% of the coding region of the mature proteases. Furthermore, using the DNA fragments as hybridization probes, protease genes from subspecies and strains of trypanosomatids can also be rapidly isolated and sequenced to provide additional information about the genetic drift of these organisms as reflected in the sequence of a key metabolic enzyme. Comparisons can be made to determine if expression of these proteases may be related to the degree of pathogenicity of the parasite. Enzymes from organisms which cannot be readily accessed in large enough quantity for classical biochemical analysis can now be studied in detail using a minimum amount of readily extracted genomic DNA.

Acknowledgements

Supported by grants from the World Health Organization and NIAID (AI20452) to JHM, Academic Senate and School of Medicine of UCSF to JAS, and NSF (DMB8608086 and EET8807179) to CSC. JB is supported by a fellowship from the Swiss National Science Foundation. T. cruzi cultures were maintained and DNA was isolated in the laboratory of Dr. Mario Castañeda at Universidad Nacional Autónoma de México, Mexico City. T. brucei genomic DNA generously provided by Dr. Syd Craig at University of California, San Francisco and T. cruzi genomic library generously provided by Dr. Steve Reed at Seattle Biomedical Research Institute.
References


33 Reed, S.L., Keene, W.E. and McKeever, J.H. (in Press) Thiol proteinase expression and pathogenicity of Enta-


Acad. Sci. USA 80. 3666–3670.