The Sequence, Organization, and Expression of the Major Cysteine Protease (Cruzain) from Trypanosoma cruzi*

Ann E. Eakin‡§, Alea A. Mills¶‖, Guenter Harth¶‖, James H. McKerrow‡, and Charles S. Craik‡

From the Departments of ‡Pharmaceutical Chemistry and ¶Pathology, University of California, San Francisco, California 94143, the ¶Department of Immunology and Infectious Disease, Palo Alto Medical Foundation, Palo Alto, California 94301, and the †Stanford University Medical Center, Stanford, California 94305

The complete sequence of the gene encoding the major cysteine protease from Trypanosoma cruzi is reported. The amino acid sequence predicted from the gene sequence aligns well with members of the papain family of cysteine proteases, suggesting the name cruzain. The sequence is most closely related to the cysteine protease of Trypanosoma brucei (59.3%) and the murine cathepsin L (42.2%). At least six copies of the gene are present in the genome and are organized in a tandem array of copies which are identical in all restriction endonuclease sites tested. The gene appears to be expressed in all developmental stages of T. cruzi with mRNA levels approximately 2-fold higher in the intracellular amastigote form. A copy of the T. cruzi gene was expressed in bacteria as an inactive, insoluble fusion polypeptide to approximately 5% of the total cell protein. The fusion protein was readily purified, solubilized in urea, and successfully refolded to produce a polypeptide which processed autocatalytically to yield approximately 1 mg of active protease per 3 g of wet cell paste. The processed form of the recombinant protease has an NH₂-terminal sequence identical to that of the mature form of the protease purified from T. cruzi (Murta, A. C. M., Persechini, P. M., Souto-Padron, T., de Souza, W., Guimarães, J. A., and Scharfstein, J. (1990) Mol. Biochem. Parasitol. 43, 27–38; Cazzulo, J. J., Cuomo, R., Raimondi, A., Wernstedt, C., and Hellman, U. (1989) Mol. Biochem. Parasitol. 33, 33–42). This suggests that the recombinant protease possesses the requisite specificity and activity to correctly process the proform of the protease in vitro. Kinetic assays with peptide substrates demonstrate that the substrate specificity and kinetic parameters for the recombinant protease are consistent with those of the endogenous protease. The proteolytic activity of the recombinant protease is enhanced by dithiothreitol, inhibited by leupeptin, N³-p-tosyl-L-lysine chloromethyl ketone and trans-epoxysuccinyl-L-leucylamido(4-guanidino)butane (E-64) but is unaffected by phenylmethylsulfonyl fluoride, pepstatin, and 1,10-phenanthroline. More specifically, the recombinant enzyme was inhibited by benzoyloxy-carbonyl-Phe-Arg-fluoromethyl ketone, which inhibits replication and differentiation of T. cruzi within mammalian cells in culture.

Proteases have been shown to be key factors in the pathogenicity of many parasitic diseases, either by inducing tissue damage and facilitating invasion or by empowering the parasites to salvage metabolites from host proteins (McKerrow, 1989). Elucidating the biological roles of these proteases and attaining a thorough knowledge of their biochemistry and structure is required for the design of antiparasitic drugs which specifically block proteolytic action. Trypanosoma cruzi is the flagellated protozoan parasite that is the etiologic agent of Chagas’ disease, a debilitating disease which infects 24 million individuals in South and Central America (Garcia-Zapata et al., 1991). Infection by T. cruzi usually culminates in the life-long chronic disease, which is characterized by an accumulation of host tissue damage over several years as the parasites sustain their life cycle by infecting host cells, multiplying intracellularly, and rupturing the cells to infect new cells. The cumulative damage to the host tissues most frequently results in myocarditis due to the destruction of cardiac muscle and associated ganglia. Alternatively, the chronic phase culminates in megasymphdromes (enlarged or distended intestines and esophagus) caused by the derervation and destruction of the smooth muscle cells in the gut. Chagas’ disease is the leading cause of early death due to heart disease in Latin America. The proteolytic activity in T. cruzi may contribute to these pathologic effects of Chagas’ disease.

Several independent lines of research have converged on a thiol-dependent protease of T. cruzi as a key metabolic enzyme (Bontempi et al., 1984, 1989) and target of the host immune response (Murta et al., 1990). This protease was initially purified by Rangel et al. (1981a) and was immunologically labeled on the three developmental stages of T. cruzi, epimastigotes (insect forms), trypomastigotes (bloodstream forms), and amastigotes (intracellular forms, Rangel et al. (1981b)). A similar protease, if not identical, is also expressed in human cells (Scharfstein et al., 1986). The cysteine proteases have been implicated to play an important metabolic role in the parasite involved in protein turnover and salvage mechanisms as it is localized to the lysosomal organelles of T. cruzi (Bontempi et al., 1989). Amino-terminal sequence analysis of the purified protein located the probable active site cysteine, which aligns with the active site Cys in papain (Cazzulo et al., 1989). Recently, Murta et al. (1990) discovered that the glycoprotein GP7/51, which had been described previously as a major antigen from T. cruzi in infected humans (Scharfstein et al., 1986), was the same cysteine protease. The protease was suggested to participate in host tissue damage directly by
secretion from the parasites, which may facilitate rupture of host cells or incidentally by leakage of the protease upon parasitic death and lysis, thus stimulating the observed host immune response (Murta et al., 1990). The partial amino acid sequence data, coupled with enzymatic characterization, classified this protease as a member of the papain superfamily of cysteine proteases (Cazzulo et al., 1989, 1990a; Murta et al., 1990).

Isolation of genes encoding homologous proteins can be accomplished if functionally and/or structurally conserved regions of the proteins can be demarcated. Classification of the *T. cruzi* thiol protease as a member of the cysteine protease family permitted the design of degenerate oligonucleotides that encoded the highly conserved amino acid sequences flanking the active site cysteine and asparagine of related cysteine proteases. These primers were used in the polymerase chain reaction (PCR) with total DNA from three species of parasitic protozoa, including *T. cruzi*, to amplify fragments representing the major portion of the protease genes of these organisms (Bakin et al., 1990). Part of the encoded sequence of the gene fragment isolated from *T. cruzi* by this method was identical to the NH2-terminal sequence of the endogenous protease purified from *T. cruzi* epimastigotes (Cazzulo et al., 1989), confirming that this gene fragment encoded part of the same cysteine protease.

In this paper we present the complete sequence of the gene for this cysteine protease from *T. cruzi*. We further show that the gene is organized in a tandem repeat of at least six copies in the *T. cruzi* genome. The mRNA encoding the enzyme is present in all three developmental stages. Expression of a copy of the gene fragment encoding the protease in bacteria and preliminary kinetic characterization of the purified enzyme verifies its classification as a member of the papain superfamily and suggests the name cruzain for the enzyme.

The expression system herein described provides active recombinant protease in reagent quantities and may have general applications for the expression of recombinant proteins whose activity, or other qualities, makes high yield soluble expression difficult.

**EXPERIMENTAL PROCEDURES**

*Materials—* Genomic DNA from epimastigotes of *T. cruzi* (Tulahuén strain, clone tetc2) was generously provided by Dr. Steve Reed of Seattle Biomedical Research Foundation in Seattle, WA. The *Escherichia coli* strains and genotypes used for expression were: dg8F (F' lacZΔM15, proAB thi, hsdR17, endA4, supE44Am, lacQ, proC: T10); DH5α (F' φ80dΔlacZΔM15, ΔlacZYA-argF[U169], endA1, recA1, hsdR17(r− m−), deoR, thi, supE44, λ×pyrA61, relA1); HB101 (F' mcrB, mrr, hsdR20, relA1, recA13, leuB6, ara-14, proA2, lacY1, gusA, xyl-1, mtl-1, rpsL200, supE44 λ−); and X90 (F' lacI, lacZΔM15, proAB, Δlac-proA, ara, nth, argEΔam, thi, rff'). The DH5α strain was used as a host for subcloning procedures. The RNA was purified from Y strain *T. cruzi*. All restriction endonucleases were purchased from New England Biolabs (Beverly, MA) or Gibco/BRL. Reagents for oligonucleotide synthesis and sequencing were purchased from Applied Biosystems (Foster City, CA). The Pro Blot membrane used for electrophoretic proteins for NH2-terminal sequence analysis was obtained from Applied Biosystems. Ampli-Taq polymerase and a DNA thermalycler were acquired from Perkin-Elmer Cetus Instruments. Protein-stained protein molecular weight markers were from Gibco/BRL. DEAE-Sepharose was purchased from Pharmacia (Upssala, Sweden).

*Genomic blot—* Genomic DNA (1 µg/reaction) from *T. cruzi* was digested for 30 min with increasing dilutions of PstI restriction

dendonuclease. The reactions were stopped with 50 mM EDTA, pH 8.0, and the digests run on a 1% agarose, Tris-acetate-buffered gel. The DNA was transferred to nitrocellulose by the method of Southern (Southern, 1975). A PCR-amplified gene fragment encoding part of the *T. cruzi* cysteine protease (Cruzain, Fig. 1) was labeled with 32P by random priming and hybridized to the blot under stringent conditions in 5 x SSC, 30% formamide at 50°C (Sambrook et al., 1989).

*Library Construction, Cloning, and Sequencing—* A library of high molecular weight fragments (15–20 kb) of *T. cruzi* genomic DNA was constructed in λEMBL phage using the AGER-11™ system from Pharmacia (Maddison, WI) according to the manufacturer's instructions. The library was screened by plaque hybridization using the PCR-generated 32P-labeled gene fragment and the same hybridization conditions as for the Southern blot. One purified phage clone, containing an insert of approximately 20 kb of *T. cruzi* DNA, spanned the entire region of the tandemly repeated gene copies (see below).

Digestion of the phage DNA with several different restriction endonucleases (i.e., *Pvu*II, *Sal*I, and EcoRI) indicated that the six copies of the gene had identical restriction sites so that the size of the 1845-bp fragment was generated by each enzyme (see Fig. 1). Because there were no restriction endonuclease sites within the intergenic regions of the tandem array that are normally found in plasmid polynucleotide cloning sites, the DNA from the phage clone was digested with *Sal*I. This endonuclease cuts at a single site within the gene so that a 1845-bp fragment is generated which contains 620 bp of the 3′ end of one copy of the gene, 444 bp of intergenic sequence, and 781 bp of the 5′ end of the next copy in the repeated to encode the entire fragment was cloned into a Bluescript KS− plasmid (pBS) from Stratagene (La Jolla, CA), and the double-stranded DNA was sequenced using the Sequenase enzyme and reagents from U. S. Biochemical Corp. The sequencing strategy is shown in Fig. 1, with 100% of the sequence being confirmed for both strands of the DNA. Sequence analysis was performed using programs designed by the Bio-computational Laboratory at University of California, San Francisco. To create a plasmid construction with a full-length copy of the gene, two copies of *Sal*I DNA fragments were ligated in tandem to create the plasmid pStStc.

*Cell culture—* *T. cruzi* (strain Y) were grown, and the different developmental stages were purified as previously described (Harth et al., 1987). The purified forms were >95% homogeneous, live organisms. The cells were harvested and stored at −20°C prior to use. Total RNA (15 µg each) from *T. cruzi* epimastigotes, trypomastigotes, and amastigotes was resolved by formaldehyde-denaturing agarose gel electrophoresis and transferred to nitrocellulose by standard methods (Sambrook et al., 1989). The blot was probed under stringent conditions with a 32P-labeled DNA fragment containing sequence proximal to and present in the entire translated protease. An autoradiogram of the Northern blot was scanned with an LKB densitometer, and the relative quantities of the hybridized radioactive probe were determined indirectly by integration of the digitized output signals. Each lane was scanned in three locations, and the digitized output values were averaged to determine the specific mRNA levels.

*Expression Plasmid Construction—* A 1438-bp fragment of DNA purifying a cDNA corresponding to the NH2-terminus of cruzain (from Cys96 to 100 bp downstream of the stop codon) was amplified with the polymerase chain reaction. The oligonucleotides used in this amplification added a 5′ XhoI site, upstream DNA sequences encoding an enteropeptidase cleavage site, and an *XbaI* site at the 3′ end of the gene. The expression plasmid, *pCheY15Lox* (Sigal et al., 1989), was digested with *Sal*I and *XbaI* to remove the lipoxigenase gene. Since *XhoI* and *Sal*I generate compatible sticky ends, the *XhoI*-XbaI PCR-amplified fragment was ligated into the plasmid *pCheY* to yield the expression of the proform of the protease as a fusion with the *E. coli* CheY protein under control of the lac promoter (see Fig. 5).

*Expression—* Overnight cultures of *E. coli* strain dg89, DH5α, HB101, and X90 transformed with *pCheY* were diluted 10-fold into fresh LB medium plus 100 µg/ml ampicillin and allowed to recover for 1 h at 37°C. IPTG (isopropyl β-d-thiogalactopyranoside) was then added to 1 mM to induce expression and the cultures were grown with agitation at 37°C for 4 more h. The cells were harvested by centrifugation and resuspended in 0.1 volume of lysis buffer (50 mM Tris, pH 8, 10 mM EDTA, 0.5 mg/ml lysozyme, and 10 µg/ml

1 The abbreviations used are: PCR, polymerase chain reaction; SDS, sodium dodecyl sulfate; PAGE, polyacrylamide gel electrophoresis; IPTG, isopropyl β-d-thiogalactopyranoside; h, base pair(s); kb, kilobase pair(s); DTT, dithiothreitol; AMC, 7-amino-4-methyl coumarin; Z, benzoylxy carbonyl;
DNase). The samples were lysed by freezing in dry ice/ethanol and thawing at 37 °C four times. The insoluble material was pelleted by centrifugation for 10 min in a microcentrifuge, and the supernatant was collected as the soluble fraction of the cells. Proteins were suspended in loading buffer with 100 mM dithiothreitol (DTT), heated at 100 °C for 5 min, and analyzed by SDS-PAGE (Laemmli, 1970).

Protein Purification—For large scale protein preparations, cultures of E. coli (strain DH5α containing the expression plasmid) were grown overnight, diluted 10-fold into fresh LB medium plus 100 μg/ml ampicillin, and allowed to recover at 37 °C for 1 h. IPTG was added to 1 mM, and the cultures were induced at 37 °C with shaking for 4 h. Cell lysis, urea solubilization, and refolding were performed as described by Marston et al. (Marston et al., 1984) with the following modifications. Insoluble proteins were solubilized in 7 M urea, and after a pH 10.7 refolding step and subsequent incubation at pH 8.0 (Marston et al., 1984), the soluble proteins were precipitated with ammonium sulfate at 40% saturation. The precipitated proteins were collected by centrifugation and resuspended in 0.1 M sodium acetate, pH 5.6, and dialyzed against two changes of 10- to fold excess of the same sodium acetate buffer to remove other salts. The proteins were then fractionated by ion exchange chromatography on DEAE-Phosphorose using a 0-1 M gradient of NaCl.

Autopeptolysis—The fusion protein purified by anion exchange chromatography was diluted 2-fold into 0.1 M sodium acetate, pH 5.5, 5 mM DTT and incubated at 37 °C. Samples were collected hourly. Activity was tested by fluorometric assay (see below), and the remaining sample was precipitated with trichloroacetic acid using perchloric acid as a carrier. The precipitate was then run on a standard 12% polyacrylamide gel and visualized with Coomassie Brilliant Blue stain (Laemmli, 1970).

NH₂-terminal Amino Acid Sequencing—Standard SDS-PAGE was performed with 12% polyacrylamide gels (Laemmli, 1970). The proteins were then electrophoresed to polyvinylidene difluoride membrane (Yue et al., 1998) for sequence analysis. Sequential Edman degradation of the proteins was performed using an Applied Biosystems model 470A gas phase protein sequencer equipped with an on-line phenylthiohydantoin-derivative analyzer. The phenylthiohydantoin-derivatives were separated by reverse phase chromatography on a Brownlee C-18 column (Hunkapiller et al., 1985). The sequencing yields were in good agreement with that estimated from intact protein on SDS-PAGE.

Activity Analysis—The activity of the recombinant cysteine protease was assayed by the liberation of the fluorescent leaving group, 7-aminomethyl coumarin (AMC, Enzyme Systems Products, Livermore, CA), from the peptide substrate, Z-Phe-Arg-AMC (where Z is benzoyloxycarbonyl), upon incubation with the enzyme at 25 °C in 0.1 M sodium acetate, pH 5.5, 5 mM DTT. Alternatively, the enzyme activity was assayed by monitoring the release of the cleaved leaving group, p-nitroaniline, from the peptide substrate, benzoyl-Pro-Phe-Arg-p-nitroanilide (Sigma). Active site titration was performed with Z-Phe-Arg-fluoromethyl ketone by allowing the titrant to bind the enzyme in 1 M sodium acetate buffer, pH 5.5, 5 mM DTT for 30 min at 25 °C before measuring activity by adding Z-Phe-Arg-AMC to 20 μM. The inhibitor profile was determined by monitoring enzyme activity with 20 μM Z-Phe-Arg-AMC after a 5-min preincubation of enzyme with 10 μM of the following inhibitors: leupeptin, trans-epoxyoxycinn-1-leucylamido(4-guanidinobutyrate (E-64), 1,10-phenanthroline, and L-Phe-Lys-OEt-CHN₂-dimethylcarbamate (TLCX) from Sigma, pepstatin from Bachem Biocience, Inc. (Philadelphia, PA), phenylmethylsulfonyl fluoride from Boehringer Mannheim, and Z-Phe-Arg-fluoromethyl ketone from Protek (Dublin, CA).

RESULTS

Genomic Blot—Shown in Fig. 1 is a Southern blot of partially digested genomic DNA from T. cruzi. The genomic DNA was probed with the radiolabeled 450 bp PCR-generated restriction fragment described previously as encoding a major portion of the T. cruzi cysteine protease gene (Eakin et al., 1990). The lanes were loaded left to right with DNA digests with increasing amounts of the restriction endonuclease PvuII, so that partial to full digestion of the DNA is visible. The symmetric ladder of evenly spaced (2 kb) bands in the lanes with partially digested DNA indicates that the protease is encoded by a tandem repeat of at least six copies of the gene.

The ladder was generated by partial digestion at the conserved PvuII restriction enzyme sites, releasing fragments of 1.8, 3.7, 5.5, 7.4, 9, and 23 kb. The sizes of the fragments depend upon the number of tandem copies of the gene they contain (see Fig. 1). The completely digested sample (lane 1) shows that the size of a single copy of the gene is approximately 1.85 kb. The termini of the gene array migrate at 22 and 9 kb. Since each terminus has only one copy of the gene, the greater distance to the adjacent PvuII site results in the larger DNA fragments.

Gene Sequence—The 450-bp PCR-generated fragment was used to screen a library of T. cruzi genomic DNA constructed in λEMBL phage. A clone was isolated which contains the entire region of the tandem repeats as demonstrated by a Southern blot of the cloned phage DNA (similar to the Southern blot in Fig. 1, data not shown). The phage DNA was digested with SaI, a resulting 1845-bp fragment was subcloned into a plasmid vector, and the DNA sequence was determined. The sequence is shown in Fig. 2 with the deduced amino acid sequence and the DNA sequence flanking the proposed coding region. The entire coding region predicts a preproform of the protein with a mass of 49,800 Da. The site of the putative signal peptide cleavage, as predicted by the von Heijne rules (von Heijne, 1986), is indicated with arrow a. The resulting proform of the protease is predicted to have a mass of 48,000 Da. The amino terminus of the protease isolated from trypansomes (Murtza et al., 1990; Cazzulo et al., 1989) is indicated with the arrow b, and the sequence predicts a mass of 36,500 Da for the mature protease. Arrow c indicates a site of autocatalysis (Hellman et al., 1991; Cazzulo et al., 1990b) which would release a 13,800-Da COOH-terminal domain leaving a 22,700-Da core protein. The amino acid numbering of the protease was adjusted to conform to the papain numbering system (Drenth et al., 1970).

In the DNA sequence of the upstream flanking region, the boldface, italicized sequence (Fig. 2) indicates a polyproline tract which may signal the location of a splice acceptor site at the next downstream AG (Ohashi and Gotoh, 1987; Padgett et al., 1986) where the spliced leader sequence may be trans-spliced onto the 5′ end of each gene copy in the polycistronic transcripts of the array. In the DNA sequence downstream of the stop codon, the boldface, underlined bases identify a possible poly(A) addition signal. The DNA and amino acid sequences reported here contain regions which are identical to the PCR-amplified gene fragments reported previously (Aslund et al., 1991; Eakin et al., 1990) and to the several reported peptide sequences (Hellman et al., 1991; Murtza et al., 1990; Cazzulo et al., 1989).

Sequence Alignment—The amino acid sequence deduced from the gene sequence is aligned with the Trypanosoma brucei (Pamer et al., 1990; Mottram et al., 1989), mouse cathepsin L (Joseph et al., 1988), and papain (Cohen et al., 1986) sequences in Fig. 3. Relative to the T. cruzi sequence, the degree of sequence identity (shown with boxes) in regions of alignment is 59.3, 42.2, and 32%, respectively. Of note, the trypanosomal sequences extend 114 amino acids (T. brucei) and 130 amino acids (T. cruzi) beyond the carboxyl terminus of related cysteine proteases. These COOH-terminal extensions have unusual stretches of repeated amino acids (prolines for T. brucei and threonines for T. cruzi). There are also 8 conserved cysteine residues in each extension. A cDNA sequence from Lycopersicon esculentum (common name, tomato; Schaffer and Fischer (1988)) predicts a cysteine protease with a COOH-terminal extension of 111 amino acids that begins with a polyproline stretch similar to that in the T. brucei sequence (North, 1981). However, no other published
cysteine protease sequences have this extension.

RNA Analysis—Shown in Fig. 4 is a Northern blot of total RNA from the developmental stages of T. cruzi (trypomastigotes, amastigotes, and epimastigotes). The RNA was probed with a PCR-amplified fragment of DNA encoding the predicted proform of the T. cruzi cysteine protease. A message of 1.9 kb is present in each stage (Fig. 4). Equal quantities (15 μg) of total RNA, as determined by optical density at 260 nm, were loaded in each lane. The results of a densitometric scan of a autoradiogram of the Northern blot are shown graphically in Fig. 4. The amount of hybridization, as determined by integrating the output signals from the densitometer, indicates that there may be as much as twice the level of mRNA transcribed from the protease gene present in the amastigote stage as compared with either the trypomastigote or epimastigote stages.

Expression Vector—A diagram of the relevant portion of the plasmid vector which was used to express the T. cruzi cysteine protease in bacteria is shown in Fig. 5. The diagram depicts the portion of the pUC vector which was expressed under control of the lac promoter in E. coli. The proform of cruzain (beginning with amino acid Cys-104) was expressed as a fusion with 45 amino acids of the CheY protein of E. coli (Gan et al., 1989). An enteropeptidase site was included at the junction of this fusion so that it could be removed subsequent to expression and isolation without altering the amino terminus of the T. cruzi protease. The domains of the protease are demarcated, and the sequences at the CheY-protease and protease domain junctions are shown above the diagram. The sites of autoproteolysis which remove the proregion and the COOH-terminal extension are indicated with arrows above the sequence. The active site Cys-36, His-19, and Asn-57 side chains are shown.

Expression and Purification—Four strains of E. coli (dg98, DH5α, HB101, and X90) were tested as hosts for the expression of the CheY-protease fusion protein. Only two of the hosts (dg98 and DH5α) produced sufficient levels of protein that could be detected by Coomassie Brilliant Blue staining of SDS-polyacrylamide gels or by immunoblot analysis. Although no fusion protein could be visualized in extracts of transformed HB101, these extracts produced the highest level of active protease above background in the soluble protein fraction, as demonstrated by activity against Z-Phe-Arg-AMC (data not shown). This observation suggests that the production of active soluble protease in HB101 bacteria may be disadvantageous to the host. Alternatively, the activity of the protease may result in autodegradation and prevent its accumulation within the bacteria. The expression of the fusion protein in dg98 and DH5α is demonstrated by a Coomassie Brilliant Blue-stained polyacrylamide gel of proteins from lysed cells which contained the expression plasmid (pCheY2, Fig. 6). After lysis the bacterial cells were fractionated into soluble (S) and insoluble (I) proteins by centrifugation to determine the solubility of the recombinant fusion protein. The fusion protein was found entirely in the insoluble fraction of each strain and appeared to be quite prominent at 4 hr post-induction with IPTG (pCheY2 lanes, Fig. 6). The size of the fusion protein was predicted from the DNA sequence to be 54 kDa. However, the migration of the fusion protein through SDS-polyacrylamide gels estimates its size to be 80 kDa. The fusion protein could not be detected in the uninduced samples (data not shown), nor was there a prominent band of the same size in the cells that did not contain the expression plasmid (pBS lanes, Fig. 6).

The estimated yield of the fusion protein produced in DH5α cells is approximately 15–20 mg/liter of bacterial culture (3 g wet weight, of cell pellet). After refolding, purification, and activation (described below), the yield of the mature, active protease is approximately 1–1.5 mg from the initial 3 g of bacterial cells. Considering that the mature form of the recombinant T. cruzi cysteine protease comprises only 46% of the original fusion protein (28,000 Da/60,000 Da), the insoluble protein yield was 7–10 mg from which ~1 mg (~10%) of purified active protease was recovered.

Autoproteolytic Processing—Incubation of the purified fusion protein at 37 °C in 0.1 M sodium acetate, pH 5.5, 5 mM DTT resulted in an increase of activity that was monitored by the release of the fluorogenic leaving group (AMC) from the peptide substrate Z-Phe-Arg-AMC. Maximal activity was attained in 4 hr and remained stable for up to a total of 7 hr of incubation (Fig. 7A). This incubation permitted autoproteolysis and resultant autoactivation and a concomitant reduc-
FIG. 2. DNA sequence of one copy of the repeated gene unit including the deduced amino acid sequence. The DNA sequence shown begins and ends at the PvuI site, which occurs in the noncoding region of the repeat. Within the noncoding sequence preceding the gene, the putative spliced leader recognition signal of a polypyrimidine stretch followed by an AG splice acceptor sequence is indicated in boldface italics. The putative start methionine initiates the coding region which is translated into the three-letter code for the amino acids, which are numbered and shown above the DNA sequence. The numbering of the amino acids is based on the papain numbering system (Drechsel et al., 1971). The active site amino acids are indicated by boxes. Three possible N-linked glycosylation sites are marked with an asterisk above the sequence. Arrow a denotes the site of signal peptide cleavage, as predicted by the von Heijne rules (von Heijne, 1986), which results in a proform of the protease beginning at Cys⁴¹⁰⁴. Arrow b denotes the cleavage point which removes the proregion of the protease and leaves the mature form of the protease as identified by the published NH₂-terminal sequence (Cazzulo et al., 1988; Murta et al., 1990). Arrow c indicates the site of autoproteolysis which removes the COOH-terminal extension (Cazzulo et al., 1990b; Hellman et al., 1991). The arrows marked with an x indicate the positions of autoproteolytic cleavage sites of the recombinant protease (see Fig. 7). The stop codon terminates the protease at Leu⁶⁶⁰⁴. The putative poly(A) addition signal is boldface and underlined in the noncoding region following the gene.

tation in the size of the fusion protein. This autocatalysis obviated the need for the enteropeptidase cleavage site and greatly simplified the production of purified, active T. cruzi protease.

The autoproteolytic processing events were demonstrated by SDS-PAGE and NH₂-terminal sequencing of the proteins resulting from increasing incubation times (Fig. 7B). As shown in Fig. 7B, amino-terminal sequencing of the autoproteolytically generated 38 kDa protein band after 3 h incubation of the fusion protein indicates that there are two cleavage sites which result in two NH₂-terminal sequences. The major (90%) sequence begins within the proregion of the protease with Ala⁴⁰⁴ (refer to Fig. 2) and results from proteolysis of a preferred recognition sequence SLH-AEETTL where “...” denotes the scissile bond. The secondary (10%) sequence begins with Ser⁵⁷⁷, six amino acids carboxy-terminal from the primary cleavage site and has the sequence SQFAX (refer to Fig. 2). The fifth amino acid of the secondary sequence was predicted to be glutamic acid, but the residual presence of the large Glu peaks from the primary sequence made the definitive assignment impossible. The observed size of this protein form indicates that the COOH-terminal domain has been removed as it migrates at a much smaller size than would be predicted for the full-length protein beginning with this NH₂ terminus.

The size of the protein observed as a result of early processing events (Fig. 7B, 0–1 h) suggests that the recognition sequences in the proregion (described above) are the first to be cleaved, releasing the remaining proform of the protease from the CheY fusion. The 38-kDa protein is the next form to be generated by the cleavage of the COOH-terminal extension. The prominent band at 28 kDa, which appears after 4 h of incubation and remains the major band for 3 additional h, also was subjected to NH₂-terminal sequencing and found to possess the sequence identical to that determined for the mature, endogenous protease purified from trypansomes (Murta et al., 1990; Cazzulo et al., 1989). The NH₂-terminal sequence begins with Ala¹ as APAAVDWRAR (see Figs. 2 and 7B), and the recognition sequence for this autoproteolytic
event is VVG-AP. Each of the proteins generated during these proteolytic events, as well as the original fusion protein, migrates through the SDS-polyacrylamide gels at a reduced rate; therefore, their observed molecular sizes are consistently greater than those predicted by the gene sequence by 5–7 kDa. This aberrant mobility may be due to unusual charge or structural features of the protease. By correlating the activity changes with the processing of the recombinant proteins, the most active form of the protease appears to be the 28-kDa protein that is prominent in the incubation mixture after 4 h and stable for up to a total of 7 h of incubation.

Activity and Inhibition—Shown in Table I are the kinetic parameters for the recombinant protease compared with published data of those for the endogenous protease for two substrates (Cazzulo et al., 1990a; Murtz et al., 1990). Although the absolute values are not identical, the data for the recombinant protease are consistent with those for the endogenous protease preparations. The calculations for the recombinant protease were based upon a mass of 28,000 Da for the mature, active protease. However, the calculations for the endogenous protease were based upon masses of 60 kDa (Cazzulo et al., 1990a) and 51 kDa (Murtz et al., 1990).

The inhibitor profile for cruzain is shown in Table II. The sensitivity of the recombinant protease activity toward leupeptin and E-64, and the persistence of the activity in the presence of phenylmethylsulfonyl fluoride, pepstatin, and 1,10-phenanthroline classify the recombinant enzyme as a cysteine protease. In addition, the increased activity in the presence of DTT supports this classification. Inhibition of the recombinant protease activity by TLCK is consistent with the rest of the profile as this inhibitor is not exclusively an inhibitor of trypsin-like proteases but will also inhibit many of the cysteine proteases. The fluoromethyl ketone substrate analogue was the most potent inhibitor tested and was therefore used as the active site titrant to determine the concentration of active enzyme in the previous kinetics calculations.

**FIG. 3. Alignment of the T. cruzi protease, cruzain, with related cysteine proteases.** The deduced amino acid sequence of the cysteine protease from T. cruzi is aligned with the T. brucei cysteine protease (Mottram et al., 1989; Pamer et al., 1990), the mouse cathepsin L (Joseph et al., 1988), and the papain (Cohen et al., 1986) deduced amino acid sequences. The numbering system for mature papain (Drenth et al., 1971) is shown below the sequences. The active site cysteine, histidine, and asparagine amino acids are in bold print. Amino acid sequence identity with cruzain is shown with boxes.

**FIG. 4. Northern analysis of the expression of the cysteine protease gene in the developmental stages of T. cruzi.** Total RNA (15 μg) from T. cruzi trypomastigotes (T), amastigotes (A), and epimastigotes (E) was electrophoresed on a formaldehyde-denaturing gel, transferred to nitrocellulose, and hybridized with RNA encoding the proform of cruzain. The resulting autoradiogram is shown. The sizes of the RNA molecular weight standards are indicated to the left of the autoradiogram. The bar graph shows the results of a densitometric scan of the Northern autoradiogram, which compares the amount of hybridization of the radioactive probe in each lane by integrating the peaks generated by the output signals from the scan. Each developmental stage (T, A, E) is indicated below the bar representing the output level from the respective lanes. The standard deviations of the values calculated from three scans of each lane are indicated with error lines above each bar.

**DISCUSSION**

In this paper, we present for the first time the complete sequence, including flanking regions, of the cloned gene for the cysteine protease from T. cruzi, cruzain. Nucleic acid characterization shows a complex genomic organization and a uniform developmental regulation of expression. Heterologous expression of the protease sequence in bacteria and characterization of the recombinant protein verified the authenticity of the gene and provides an extremely efficient system for studying the encoded enzyme.

**Nucleic Acid Characterization**—The gene encoding cruzain is organized in the T. cruzi genome as a series of at least six tandemly repeated copies. These copies appear to be identical to the extent that their restriction endonuclease sites are conserved so that digestion of the genomic DNA consistently generates a 1845-base pair fragment containing one copy of the repeat. Partial sequences of this gene have been reported in the literature (Aslund et al., 1991; Bakin et al., 1990) which have a very high nucleic acid sequence identity (>98%) with the complete gene herein reported. However, there are some differences which result in amino acid changes. These changes could be a result of: 1) cloning or sequencing artifacts, 2) interstrain differences, or 3) intrastrain differences. The
amino acid substitutions may be inconsequential or may result in functional differences among the expressed proteases. The entire tandem array is probably transcribed in a single polycistronic message which is subsequently cleaved into single-copy mRNA by trans-splicing via a mechanism common to many trypanosomal genes (Muhich and Boothroyd, 1988). Our estimate for the size of the mRNA in Northern blots is consistent with this interpretation (see below). The function of this type of organization and RNA processing is the subject of active research in a number of laboratories. Others have shown that many, if not all, trypanosomal mRNAs are trans-spliced so that a conserved 39-nucleotide leader sequence is added to the 5' end of each mRNA (Murphy et al., 1986; Sutton and Boothroyd, 1986; van der Ploeg, 1986). This leader sequence possesses a 7-methylguanosine cap so that the trans-splicing event assures that each mRNA will acquire this cap at the 5' end (Sutton and Boothroyd, 1986; Perry et al., 1987). This type of genomic organization and messenger RNA processing may provide a direct means of controlling the level of gene expression and protein translated by regulating the post-transcriptional processing machinery instead of regulating initial transcription of the gene.

We show that the in vivo expression of the cruzain gene, as determined by Northern analysis, indicates that the message is present in all three developmental stages of the life cycle of T. cruzi. Additionally, our results show that the level of expression may be roughly 2-fold greater in the intracellular amastigote form of the parasite. The size of the message is approximately 1.9 kb, indicating that the majority of the mRNA for the protease is in a single-copy form as opposed to a stable polycistronic message containing several copies. Puff hybridization can be detected in longer exposures of the Northern blot shown in Fig. 4, corresponding to a message size of 4.4 kb. These bands may represent incompletely trans-
Cruzain, the Major Cysteine Protease from T. cruzi

Table I

Kinetic parameters for the recombinant protease compared with published results for the endogenous enzyme

<table>
<thead>
<tr>
<th>Inhibitor</th>
<th>( K_m ) (( \mu M ))</th>
<th>( k_{cat} ) (s(^{-1} ))</th>
<th>( k_{cat}/K_m ) (s(^{-1} ) M(^{-1} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z-Phe-Arg-AMC</td>
<td>10</td>
<td>3.1</td>
<td>3.1 ( \times ) 10(^{5} )</td>
</tr>
<tr>
<td>Endogenous( ^a )</td>
<td>0.96 (0.049)( ^b )</td>
<td>4.48 (0.13)</td>
<td>46.82 ( \times ) 10(^{5} ) (1.12 ( \times ) 10(^{6} ))</td>
</tr>
<tr>
<td>Recombinant</td>
<td>35</td>
<td>9.89</td>
<td>292 ( \times ) 10(^{5} )</td>
</tr>
<tr>
<td>Benzoyl-Pro-Phe-Arg ( p )-nitroanilide</td>
<td>89 (7.99)( ^b )</td>
<td>3.23 (0.11)</td>
<td>36.44 ( \times ) 10(^{5} ) (32.8 ( \times ) 10(^{5} ))</td>
</tr>
</tbody>
</table>

\( ^a \) Data published by Murta et al. (1990); assuming 100% of detectable protein is active. Mass of enzyme was 51 kDa.

\( ^b \) Data in parentheses are the standard deviation of three data points averaged for the values given.

\( ^c \) Data published by Cazzulo et al. (1989); active site titration. Mass of enzyme was 60 kDa.

Table II

Inhibitor profile for the recombinant protease

<table>
<thead>
<tr>
<th>Inhibitor (10 ( \mu M ))</th>
<th>Activity</th>
<th>Relative Activity</th>
<th>% of uninhibited</th>
</tr>
</thead>
<tbody>
<tr>
<td>No inhibitor</td>
<td>9.71</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Pepstatin</td>
<td>9.20</td>
<td>94.7</td>
<td>94.7</td>
</tr>
<tr>
<td>Phenylmethylsulfonyl fluoride</td>
<td>8.79</td>
<td>90.5</td>
<td>90.5</td>
</tr>
<tr>
<td>1,10-Phenanthroline</td>
<td>8.71</td>
<td>89.7</td>
<td>89.7</td>
</tr>
<tr>
<td>TLCK( ^d )</td>
<td>0.066</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>E-64</td>
<td>0.055</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>Leupeptin</td>
<td>0.023</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td>Z-Phe-Arg-FMK( ^d )</td>
<td>0.0046</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>No DTT</td>
<td>7.33</td>
<td>75.5</td>
<td>75.5</td>
</tr>
</tbody>
</table>

\( ^d \) Inhibitor and enzyme were preincubated 5 min before activity was assayed.

\( ^d \) N\(^{\text{-}} \text{p}-\text{tosyl}-\text{t}-\text{lysine chloromethyl ketone.} \)

\( ^d \) Benzoyloxycarbonyl-Phe-Arg-fluoromethyl ketone.

spliced RNAs which contain two copies of the gene-repeat.

The presence of mRNA encoding the cysteine protease gene in each developmental stage of the parasite is consistent with both immunologic (Murta et al., 1990; Souto-Padrón et al., 1990) and activity studies (Campetella et al., 1990). However, the proteolytic activity studies suggest that there is roughly 10-fold more protease in the epimastigote stage than in either amastigotes or trypomastigotes (Campetella et al., 1990). The mRNA levels do not reflect this difference in activity, suggesting that other factors such as the stability or post-translational processing of the protease or expression of an endogenous inhibitor may account for the differences.

Production of Cruzain in Bacteria—The expression of the protease gene in bacteria proved to be very difficult. Several common expression vectors, utilizing different promoters (i.e., tac, phoA, T7) for control of either secretion or intracellular expression, were used without success. A probable reason for the success of the CheY fusion in generating relatively large quantities of recombinant enzyme is that this system initially produces inactive and insoluble protein. The potential for toxicity and instability of heterologously expressed proteolytic enzymes is great. In our current expression system, the inactivation of the protease by precipitation in inclusion bodies appears to be required for the accumulation of significant amounts of the recombinant enzyme. In addition, the insolubility of the recombinant protein provides a useful and extremely convenient purification step. The use solubilization of the inclusion bodies and subsequent refolding steps allow the recovery of the fusion protein from which active protease can be generated by autoproteolysis.

The autocatalytic pattern exhibited by the recombinant protein indicates that cruzain has the capability and specificity to process its proform to the fully active mature protease with the same NH\(_2\) terminus as that found on the endogenous enzyme (Murta et al., 1990; Cazzulo et al., 1989). Although the proform of the protease has never been purified from the trypanosomes, autoproteolysis has been observed in the endogenous protein which cleaves a 25-kDa protein fragment from the COOH terminus of the mature protein and leaves a 35-kDa fragment which has the same NH\(_2\) terminus as the original 60-kDa protein (Hellman et al., 1991; Cazzulo et al., 1989). The NH\(_2\)-terminal sequence of the 25-kDa protein fragment localizes the cleavage site to the region where the COOH-terminal extension of the trypanosomal sequence diverges from the majority of cysteine proteases (Fig. 2, arrow c) (Hellman et al., 1991; Cazzulo et al., 1989).

Sequence identities indicate that the gene and recombinant protein reported herein is the same enzyme that is the subject of other studies which report the size of the mature protease to be in the range of 55–60 kDa (Hellman et al., 1991; Murta et al., 1990; Cazzulo et al., 1989). The reasons for the size differences between the endogenous and recombinant proteases may be due to post-translational modifications that occur in vivo. Within the predicted amino acid sequence for the protease are three potential N-linked glycosylation sites predicted by the conserved signal sequence NXS/T (see Fig. 2). The endogenous protease has been shown to be a glycoprotein that changes in size from 60 to 54 kDa after deglycosylation with endo-\( \beta \)-N-acetylgalactosaminidase H (Cazzulo et al., 1990b). However, the difference between the predicted sequence and observed molecular masses is still great, 36 kDa for the predicted protein versus 54 kDa for the deglycosylated endogenous protein. The persistent size difference may be due to carbohydrate moieties remaining attached to the protein which are not of the high mannose class and therefore endo-\( \beta \)-N-acetylgalactosaminidase-resistant.

Notably, both the mature protease and the COOH-terminal extension fragment in endogenous preparations appear to be approximately 2-fold larger than would be predicted by the gene sequence (60 versus 36 kDa for the mature enzyme and 25 versus 14 kDa for the COOH-terminal domain). This observation could be explained if the endogenous enzyme formed covalent dimers resistant to the reducing agents in the SDS-PAGE buffers. The cross-linking might be coordinated through the glycosylation moieties as there is precedence for a typical O-linked glycosylation moiety participating in a protein-glycosaminoglycan-protein linkage (Engelchild et al., 1991). The electrophoretic mobility of the COOH-terminal extension fragment increases upon treatment with endo-\( \beta \)-N-acetylgalactosaminidase to remove high mannose oligosaccharides (Cazzulo et al., 1990b). The diagnostic antigen GP25, which is a peptide fragment from the GP57/51 preparation of the endogenous protease and presumably corresponds to the COOH-terminal domain, was demonstrated to be modified by O-linked glycosylation moieties which may contribute to the
aberrant electrophoretic mobility or possible dimerization of this domain (Murtza et al., 1980; Mendonça-Previo et al., 1989).

Not only does the COOH-terminal domain migrate at twice the expected molecular weight but also it is very resistant to proteinolytic digestion (Hellman et al., 1991). Furthermore, the ambiguous sequence generated during Edman degradation of this fragment (Hellman et al., 1991) may be due to amino acids in this region of the sequence being modified. These observations suggest that many of the modifications and/or possible dimerization contacts may be localized to this domain. Within bacteria the post-translational modifications required to glycosylate the enzyme or generate the dimer would be absent, and thus the recombinant protease generated has the predicted electrophoretic mobility in SDS-gels.

Kinetic Characterization and Significance—The increase in activity, coinciding with autophosphorylation of cruzain in vitro, suggests that the most fully active form of the protease is the 28-kDa protein. It is unclear from results so far whether the activity of the protease is influenced by the removal of the COOH-terminal extension. The activity of the fully processed 28-kDa recombinant protein, presumably without the COOH-terminal extension, is similar to the activity observed with the endogenous enzyme, thus suggesting that the COOH-terminal extension may have little effect on the level of activity of the protease. In agreement with this finding is the observation that the cysteine protease from T. brucei, when expressed in bacteria, does not require the COOH-terminal extension for the production of active protease (Pamer et al., 1991).

The substrate specificity of the recombinant fusion protein was partially revealed by the amino acid sequences flanking cleavage sites for autoproteolysis in generating the mature enzyme. Notably, both of the cleavage sites for the autoproteolytic removal of the prosequence and the COOH-terminal extension have the same recognition sequence of Val-Val-Gly in the P1–P3 sites (see Fig. 2). Although the absolute values of the kinetic parameters for cruzain produced in bacteria are not coincident with the published values of the endogenous activity isolated from the parasite, the relative values are quite similar, especially in light of the different sources of the enzyme. A significant difference in the calculations of the data is the assignment of the mass of the protease. The mass of the active, recombinant protease is 28,000 Da, whereas the mass used in the calculations with the endogenous enzyme is 60,000 Da (Cazzulo et al., 1990a; Murtza et al., 1990). This disparity could account for some of the discrepancies in the data. The kinetic data and inhibitor profile support the assignment of the cloned gene as that for the cysteine protease from T. cruzi.

The function of cruzain within the parasite and its involvement in the pathology of the infection remains unknown. Recently, some clues have been revealed by studying the effect of protease inhibitors on in vitro cultures of T. cruzi with mammalian cells. The addition of the inhibitor Z-Phe-Arg-fluoromethyl ketone, at a concentration of 10 μM, to media with infective trypomastigotes and mammalian cells resulted in a dramatically reduced number of intracellular amastigotes as compared to control cells without inhibitor. This same fluoromethyl ketone peptide is shown here to be a very effective inhibitor and active site titrant of the recombinant protease in vitro. Effects of protease inhibition at a different stage in the life cycle of T. cruzi were examined by Bonaldo et al. (1991). The cysteine protease inhibitors antipain, leu-

peptin, and Z-Phe-Ala-diazomethane, were found to block the differentiation from the epimastigote stage of the parasite into the metacyclic trypomastigote stage (metacyclogenesis) in a dose-specific fashion. This differentiation event normally occurs in the insect vector whereby the insect stage of the parasite is transformed into the developmental stage which is infective to humans. These results support the previous findings which suggest that the cysteine protease may play an important role in the development and differentiation of the parasites at several stages of their life cycle.

The amino acid sequence of the cysteine protease from T. cruzi deduced from the gene sequence has significant identity to members of the papain superfamily of cysteine proteases: 59.3% identical to the cysteine protease from T. brucei, 42.2% identical to murine cathespin L, and 30% identical to papain. Because of the high degree of sequence identity to papain and the generalization that homologous proteins have similar three-dimensional structures, a molecular structure of the cysteine protease from T. cruzi was modeled based upon the published crystal structures of the related proteins papain and actinidin. This model will assist in the elucidation of the precise crystal structure of cruzain as well as enable the design of experiments for more definitive structure/function studies. Detailed structural and kinetic analysis of cruzain should provide further insights into its biochemical function and substrate preferences which may facilitate the design of specific inhibitors which could be used as chemotherapeutic agents for the treatment of Chagas' disease.

Acknowledgments—We thank Steve Reed for the generous gift of genomic DNA from T. cruzi. We sincerely acknowledge Syd Craig, David Corey, and Scott Willett for careful reading of the manuscript and for numerous suggestions during the course of the work. We thank Ralph Reid and the Biomedical Research Center at the University of California, San Francisco for invaluable help with the protein sequence determination.

REFERENCES


---

Cruzain, the Major Cysteine Protease from T. cruzi


Yuen, S. W., Chui, A. H., Wilson, K. J., and Yuan, P. M. (1989) Biotechniques 7, 74–83