Multiple cysteine proteinases of the pathogenic protozoon *Tritrichomonas foetus*: identification of seven diverse and differentially expressed genes

David J. Mallinson,¹† Jean Livingstone,¹ Kameedea M. Appleton,¹ Sharon J. Lees,¹ Graham H. Coombs² and Michael J. North¹

Author for correspondence: Michael J. North. Tel: +44 1786 467764. Fax: +44 1786 464994.
e-mail: mjn1@stirling.ac.uk

¹ Department of Biological and Molecular Sciences, University of Stirling, Stirling FK9 4LA, UK

² Biochemical Parasitology, Institute of Biomedical and Life Sciences, Joseph Black Building, University of Glasgow, Glasgow G12 8QQ, UK

The cattle protozoan parasite *Tritrichomonas foetus* has multiple forms of cysteine proteinases. To investigate their diversity, PCR and reverse transcriptase PCR were used to isolate genomic DNA and cDNA fragments, respectively, encoding different cysteine proteinases. Seven genes have been identified, *TFCP3–6* from amplification of genomic DNA and *TFCP7–9* from amplification of cDNA. Comparison of the predicted amino acid sequences indicates that the *T. foetus* enzymes are cathepsin-L-like rather than cathepsin-B-like in structure. However, there is considerable diversity among the proteinases. *TFCP7* and *TFCP8* are most similar to one another (78% identity), while *TFCP3* and *TFCP9* are the least closely related (30% identity). All but one of the genes are single-copy, the exception being *TFCP3*, which was present in multiple copies in one of the three isolates examined. Single transcripts were detected for each of the seven genes. *TFCP8* was expressed at the highest levels, while transcripts for *TFCP4* were only just detectable. In *T. foetus* F2, the strain from which the genomic DNA and mRNA were isolated, transcripts of the five other genes were present at intermediate levels. When two other isolates were compared with F2, differences in the expression of individual genes were apparent, with either one or two of them not expressed. In spite of these differences the major cysteine proteinases detected in the three isolates using substrate-SDS-PAGE appeared identical. The data show that the multiplicity of cysteine proteinases in *T. foetus* is due, in part at least, to the presence of multiple genes and that some of the genes encode cysteine proteinases which are not among the high-activity enzymes detected previously.

Keywords: *Tritrichomonas foetus*, trichomonad, cysteine proteinase, PCR, reverse transcriptase PCR

INTRODUCTION

Tritrichomonas foetus is a flagellate protozoan parasite

† **Present address:** Beatson Institute for Cancer Research, CRC Beatson Laboratories, Garscube Estate, Switchback Road, Bearsden, Glasgow G61 1BD, UK.

Abbreviations: CP, cysteine proteinase; RT-PCR, reverse transcriptase polymerase chain reaction.

The nucleotide sequences of *TFCP3*, *TFCP4*, *TFCP5*, *TFCP6*, *TFCP7*, *TFCP8* and *TFCP9* have been assigned the EMBL accession numbers X87776, X87777, X87778, X87779, X87780, X87781 and X87782, respectively.

which is the causative agent of bovine trichomoniasis, colonizing both male and female reproductive tracts. It is asymptomatic in bulls, but the infection in cows is more serious and usually leads to transient infertility and, more rarely, abortion and pyometra. The disease is common throughout the major cattle-rearing areas of the world where natural breeding is relied upon and results in substantial economic losses in these areas (Yule *et al.*, 1989; Corbeil, 1994). Like other trichomonads, *T. foetus* contains high proteolytic activity due to at least 10 cysteine proteinases (CPs), the majority of which have apparent molecular masses of 18–34 kDa and appear to be

lysosomal (Müller, 1973; Lockwood *et al.*, 1984, 1987). Proteinases, exclusively of the cysteine type, are also released into the culture medium during axenic growth *in vitro* (Lockwood *et al.*, 1988; North *et al.*, 1990a; Talbot *et al.*, 1991). CPs have also been reported to be present in the cervico-vaginal mucus of beef heifers artificially infected with *T. foetus* (Yule *et al.*, 1989). It has been suggested that both the intracellular and extracellular CPs may play a role in host-parasite interactions (Lockwood *et al.*, 1987, 1988; Yule *et al.*, 1989; Burgess *et al.*, 1990; Talbot *et al.*, 1991). Indeed, CP inhibitors such as leupeptin have been shown to reduce the contact-mediated cytotoxicity of *T. foetus* to mammalian cells (Burgess *et al.*, 1990), and *T. foetus* extracellular CPs cleave proteins relevant to the defence of the bovine female reproductive tract (Talbot *et al.*, 1991).

Although it is likely that proteinases have a role in the interaction of *T. foetus* with its host, the enzymes have not been characterized in detail. Substrate and inhibitor studies on electrophoretically separated intracellular and extracellular enzymes have allowed some features of individual enzyme activities to be determined and have demonstrated differences in substrate specificity and inhibitor sensitivity between enzymes (North *et al.*, 1990a). This has suggested that some of the enzymes are distinct from one another, and this indirect evidence supports the view that at least some of the multiplicity in enzyme activities is due to multiple genes (North, 1991). As little work has been done on CP purification (McLaughlin & Müller, 1979; Irvine *et al.*, 1993), and it is only recently that a single short primary sequence has become available (Irvine *et al.*, 1993), the molecular basis of CP multiplicity has remained a matter of speculation.

The polymerase chain reaction (PCR) has proved a useful starting point for the analysis of CP genes from a number of organisms, including many protozoan parasites (see Mallinson *et al.*, 1994). The aim of the work reported here was to use PCR to amplify genomic DNA, and reverse transcriptase PCR (RT-PCR) to amplify cDNA, to generate fragments which would allow the diversity among the CP genes to be investigated. The fragments were then used to analyse the variation in the organization and expression of the *T. foetus* genes.

METHODS

Growth of trichomonads and preparation of DNA and RNA.

A clonal cell line of *T. foetus*, F2 (Lockwood *et al.*, 1984), was used unless otherwise indicated. The other isolates used were Manley and Belfast; these were kindly provided by Dr M. Taylor (Central Veterinary Laboratory, Weybridge, UK). The parasites were grown in modified Diamond's medium (Diamond, 1957) as previously described (Lockwood *et al.*, 1984) and harvested for DNA and RNA preparation in the exponential phase of growth by centrifugation. Genomic DNA, total RNA and poly(A)⁺ RNA were prepared as previously described for *Trichomonas vaginalis* (Mallinson *et al.*, 1994). The cDNA used in RT-PCR was synthesized using 5 µg total RNA primed with an oligo dT primer using a Pharmacia First-strand cDNA Synthesis Kit following the manufacturer's instructions. Five reactions were performed, and the products of these were combined; the cDNA was purified by spermine precipitation as described (Davis *et al.*, 1986) and resuspended in 20 µl water.

PCR and RT-PCR. The degenerate oligonucleotide primers used in the PCR reactions to amplify CP gene fragments were based on consensus sequences of highly conserved areas flanking the active-site cysteine (primers 1 and 3) and asparagine (primer 2) (North *et al.*, 1990b; Mallinson *et al.*, 1994). Those used with genomic DNA as the target (primers 1 and 2) were provided by Dr J. C. Mottram (Wellcome Unit of Molecular Parasitology, University of Glasgow); those used with cDNA as the target (primers 2 and 3) were supplied by Genosys Biotechnologies (Cambridge, UK). The primers were synthesized with inosines in positions where all four bases were possible in a codon and with *EcoRI* or *MluI* recognition sites added to the 5' ends to facilitate cloning. The sequences were as follows, using the IUB Group Codes for the redundancies, and the cloning sites are underlined: primer 1, 5' CCGAATTCCARGGICARTGYGGIWSITGYTGG 3'; primer 2, 5' CCGAATTCCCAISWRTYTTIACDATCCARTA 3'; primer 3, 5' CCACGCGTCARGGICARTGYGGIWSITGYTGG 3'. Primer 3 was essentially the same as primer 1 except that an *MluI* cloning site replaced the *EcoRI* site at the 5' end to allow directional cloning and to minimize cloning of multiple inserts which sometimes occurred when cloning PCR products amplified using primers 1 and 2.

The primers were used in 100 µl PCR reactions using the Perkin-Elmer Cetus GeneAmp kit components. When genomic DNA was the target, each reaction contained 100 ng parasite DNA and primers 1 and 2, each at a final concentration of 0.5 µM. The reaction conditions were as follows: denaturing, 94 °C for 1 min; annealing, 55 °C for 1 min; extension, 72 °C for 3 min, for a total of 30 cycles. When cDNA was the target, each reaction contained 2 µl cDNA, which was heated at 94 °C for 5 min and quenched on ice immediately before addition, and primers 2 and 3 each at a final concentration of 0.5 µM. The reaction conditions were as above except that the annealing step was carried out at 35 °C. Appropriate controls were included in which PCR reactions contained the primers alone or in combination, but without the target DNA.

The PCR products were run on 4% (w/v) agarose gels (3:1, wide range/standard, Sigma). The band of approximately 500 bp obtained from the PCR reactions was excised from the gel and cloned into the *EcoRI* site of pUC18. With the cDNA PCR reactions, the whole PCR reaction mix was purified and concentrated using Glassmilk (Life Technologies), and the fragment cloned into the *EcoRI/MluI* site of pSport1 (Life Technologies). DNA from the resulting recombinant bacteria was analysed by restriction mapping, differential hybridization and partial sequencing.

DNA sequencing. All sequencing was double-stranded and was carried out using the Sanger method (Sanger *et al.*, 1977) with the Sequenase enzyme (United States Biochemicals). Partial sequencing of the 5' and 3' ends of inserts in pUC18 was performed using pUC forward and reverse primers, and for pSport1 with pUC forward primers and T7 primers (Promega). Clones of interest were sequenced completely in both directions using the primers appropriate for the vector or custom-synthesized oligonucleotides (18-mers) based on previously determined sequences within individual inserts. The latter were purchased from Genosys Biotechnologies (Cambridge, UK) or synthesized by Dr C. S. Gillespie (University of Stirling). Overlapping sequences were combined to give complete sequences and the predicted amino acid sequences determined using the computer program DNASIS (Pharmacia). Alignment of the amino acid sequences was performed by eye and using MultiAlin (Cherwell Scientific), which uses the algorithm described by Lipman & Pearson (1985). Comparisons of the amino acid sequences with other known CP sequences were

carried out using the FASTA program available on the University of Wisconsin GCG package and by searching the Owl composite protein database (available through the SERC facility at Daresbury, UK).

Southern and Northern blotting. Southern and Northern blotting and nucleic acid hybridizations were performed exactly as previously described (Mallinson *et al.*, 1994).

Gelatin-SDS-PAGE. Analysis of proteinase activity on gelatin gels was carried out exactly as described previously (North, 1994).

RESULTS

PCR amplification of genomic DNA encoding *T. foetus* CPs

By utilizing PCR and the standard CP primers (1 and 2), a single band of DNA of approximately 500 bp was amplified from genomic DNA of *T. foetus* F2 (data not shown). This was the expected size for a cathepsin-L-like CP gene if no introns were present. The band was excised from the gel and cloned into pUC18. A total of 71 recombinant clones were selected, and by using a combination of restriction enzyme analysis, partial DNA sequencing and differential hybridization, four different

cloned DNA fragments were identified, each of which had a sequence resembling those of known CPs. Complete sequencing of the four cloned fragments and translation into amino acid sequences revealed that, as expected, each encoded the central region of a CP (Fig. 1). The genes were designated *TFCP3*–*6*. Note that the designations *TFCP1* and *TFCP2* have been given to two other *T. foetus* CP genes which have recently been cloned in another laboratory (J. S. Ikeda, J. W. Thomford & L. B. Corbeil, personal communication; GenBank accession numbers U13153 and U13154, respectively).

Subsequent Northern blot analysis indicated that some of the genes identified after amplification of genomic DNA were poorly expressed (see Fig. 3), and so to ensure that well-expressed CP genes were among those examined, cDNA fragments were amplified by RT-PCR. Again a single band of approximately 500 bp was obtained (data not shown). Analysis of recombinant clones revealed three new sequences, each corresponding to the central region of a typical CP (Fig. 1). The genes were designated *TFCP7*–*9*. The *TFCP8* sequence was frequently encountered among the recombinants (in 13 of 23 examined), suggesting that the mRNA was very abundant in the RNA preparation used for RT-PCR and that the gene was highly expressed.

TFCP3	SFAACAAFEAGAWFASSGKLVKISEQLFVDCCKY---CFGCGYGSADAAYNWAIHENDGKVLCHEDYPYTGTOGV-CRYKSSM-
TFCP4	AFATTQCMESINALRFLKSLFSFSEQLNVDCDPQ---SNGCAGGSPFSAFMFISRTQNGQINLEDDYPYTGTDNDCKFDPSK-
TFCP5	AFSTIVAQEGCHQIETGELLRLSEQLNVDCADN---CHGCDGGWPIEAFNYVLNKQGGKYCTDDYPYTABQALLCYFYRVQQ
TFCP6	AFSAIQAIESVYAIGTGLLSLSEQLNVDCVDT---CEGCNGGLMDAAYDYVIEKQNGQFNTEASYWYIGIDET-CMFDKYE-
TFCP7	AFSAIQAAESANAISTGTLESYSEQLNVDCVTA---CYGCNGGLMDASYEYIVAKQGGKMNYESDYVYTALDGT-CKFTQYT-
TFCP8	AFSAIQAAESVNCIKSGKLERYSQNLVDCVTA---CYGCNGGLMDASYEYIIDSQNGHLNLEADYPYTAVDGT-CKYAOYT-
TFCP9	AFSAVCAQEGQWARTKGEGLLSLSVQNLDCDD---SEGCNGGWPFSGIFHVISEQNGEWMLENDYPYTSNQCYPFDASK-
TFCP1	AFSAIQAAESAYAIETGTLESYSEQLNVDCVQ---CYGCNGGLMDAYAYKIIDRQKGMILESDYVYTALDGV-CKFAQFQ-
TFCP2	AFSAIAAQESCHAIATGELLRFSEQLNVDCVTS---CYGCNGGWSKAIIDYVVKHQAGKFMLTADYPYTARDGT-CKFHASK-
TvCP1	TFSTIQAMESQWAVKHTKLYSLSEQLNVDCVTT---CYGCNGGLMELAYDYVVKTYQKGFMTADYPYKAIDQS-CKFNAKV
TvCP2	AFSAIQAESERYAQANKQLDLAEQNIIVDCVTS---CYGCNGGWPASKAIDYVVKHQAGKFMLTADYPYTARDGT-CKFHASK-
TvCP3	AFSVVQVQESQWALKKQQLSLAEQNMVDCVDT---CYGCDGGDEYLADYVVIKHQKGLWMLTDYPYTARDGS-CKFKAAC-
TvCP4	AFSAIQAESQYIISFKTLQSLSEQLNVDCVTT---CYGCNGGLMDAAYDYVVKHQSGKFMTEADYPYTARDGS-CKFNAKV
DdCP2	SFSTTGSTEGAHALKTKKLVSLSEQLNVDCSGPEE---NFGCDGGLMNAFDYII-KNKG-IDTESSYPYTAETGTCFLNKS-
Drosophila	AFSSTGALEGOHFRKSGVLVSLSEQLNVDCSTKYG NNGCNGGLMDNAFFYIK-DNKG-IDTEKSYPYEAIDDS-CHFNRAQ-
HCL	AFSATGALEGOHFRKSGVLVSLSEQLNVDCSGPQG-NEGCNGGLMDYAFQYVQ-DNKG-LDSESYPYEATEES-CKYNPKY-
	30 40 50 60 70 80 90 100
TFCP3	AYGHVSQYVVRVFSLSSEIDELMCQTLEEIGPLTVAIDADGAKFRLYDSGIYYDDTCVQGD-ANHAVAVVGYGEED--NGEQ
TFCP4	GYGRITGFMSVQA---QSEEDLFKCVASVGPVAVCIDASLASFSNYSYSSGIYNDRCSSSTV-LDHAVGCGIYGAE--GA
TFCP5	PVSNIASVYQIPQ---GDEEAMKEVVANWGPVAINVDSNYGSFNFYDGGIYVEESCQVKYVYSHAMGIIGYGSAA--GQD
TFCP6	KAGISGGYVNA---SSEDDLAAKVEQYGPAAVAIDASAVGFQLYWGGIYDNSGCSSTM-LDHGVGCVGFGVEG--GTQ
TFCP7	AVGSVSKYVNAQ---GDEDDLASKCETYGPIAVIDASNWSFQLYSGGIYDEKSCSSYS-LDHGVGCVGFGVEG--STK
TFCP8	PVASITKYVNVNQ---NDEDDLAAKVETYGPAVAIDASNWSFQLYTGGVYDEPSCSPYS-LDHGVGCVGFGAE--STK
TFCP9	GVSKTTKIVQLPI---NEEKILAA-CAEYGVISCCIDSSPIDFMYSEGIFDTDQCNWAE-LDHAVNIVGYGAEE--GTK
TFCP1	TGVNVASFLYIAE---NDEEDLAANVETHGPPVAVDAIDASHQSFQLYKSGIYDEPECSATF-LNHGVGCGIYFGSDN--DTK
TFCP2	KVSNIVAVTMFPQ---SDEQNLKGHIAANGPVSCNVDAHGYSFQLYQGGIYWSWFCRTQYIYNHAMGIVGYGVEG--SEE
TvCP1	AEPTVTGYITVTE---GDEKDLMNKVAQYGPAAIAIDASHYSFQLYSGGIYDESSCSPEG-LDHAVGCVGYGSEG--SKN
TvCP2	SVGLTKGYDEVKD---TEAELAK-AASKGVVSVCIDASHYSFQLYTSGIYDEPSCSAWN-LDHAVGLVGYGTEG--SKN
TvCP3	GVTLTKSYVRPTT---TQNEDELKAGCAKGGVSVIAIDASHYSFQLYSSGIYNPKSCSSTF-LDHAVGLVGYGTEN--KVD
TvCP4	GTSQIKSYVNVAE---GDEKDLATKVSTLGPAAIAIDASAWSFQLYSGGIYDESACSSYN-LDHGVGCVGYGTEG--SKN
DdCP2	IGATIKGYVNTTA---GSEISLEN-GAQHGPVSVDAIDASHNSFQLYTSGIYEPKCSPTPE-LDHGVLVVGYGVGQ**KANN
Drosophila	VGATDRGFTDIPQ---GDEKKMPEAVATVGPVSVDAIDASHESFQFYSEGVIYNEPQCDQON-LDHGVLVVGFGTDE--SGED
HCL	SVANDTGFVDIPK---QEKALMKAVATVGPISVAIDAGHESFLFYKEGIYFEPDCSSD-MDHGVLVVGYGFESESDNN
	110 120 130 140 150 160 170 180

Fig. 1. Comparison of the predicted amino acid sequences of *Tritrichomonas foetus* CPs with each other and with *T. foetus* TFCP1 and TFCP2 (J. S. Ikeda and co-workers, personal communication), *Trichomonas vaginalis* TvCP1–4 (Mallinson *et al.*, 1994), *Dictyostelium discoideum* CP2 (DdCP2, Pears *et al.*, 1985), *Drosophila melanogaster* CP1 (Matsumoto *et al.*, 1995) and human cathepsin L (HCL, Gal & Gottesman, 1988). Numbering is that for mature single-chain human cathepsin L. Δ , Active-site histidine. Putative *N*-glycosylation sites are underlined. Dashes indicate gaps made to maximize alignments. Residues shown in bold match across all sequences. **, 36 additional residues in DdCP2.

Table 1. Identity between TFCPs and CPs of other organisms

Identity was calculated as the percentage of the total amino acid residues predicted from the TFCP3–9 PCR fragments (Fig. 1) for which an identical amino acid was present in other CPs.

Proteinase	Identity (%)							Ref.*
	TFCP3	TFCP4	TFCP5	TFCP6	TFCP7	TFCP8	TFCP9	
TFCP3	–	36	31	38	41	39	31	1
TFCP4	34	–	35	47	45	44	43	1
TFCP5	31	36	–	38	39	38	37	1
TFCP6	36	47	37	–	65	63	39	1
TFCP7	39	45	38	65	–	78	39	1
TFCP8	38	44	37	63	78	–	39	1
TFCP9	30	43	36	39	39	39	–	1
TFCP1	36	44	41	55	67	63	35	2
TFCP2	34	35	48	45	43	44	38	2
TvCP1	34	47	39	57	59	57	39	3
TvCP2	33	45	37	49	53	52	43	3
TvCP3	35	43	37	46	49	47	47	3
TvCP4	36	47	41	63	68	63	40	3
<i>Dictyostelium discoideum</i> CP2	35	43	36	47	46	46	33	4
<i>Trypanosoma brucei</i>	31	32	35	39	38	39	34	5
<i>Entamoeba histolytica</i> ACP1	32	39	29	33	38	39	30	6
<i>E. histolytica</i> ACP2	34	32	26	36	34	36	32	6
<i>Fasciola hepatica</i> Fcp4	31	35	28	36	38	37	30	7
<i>F. hepatica</i> Fcp6	31	35	31	36	34	36	31	7
<i>Drosophila</i> CP1	32	39	41	42	43	46	35	8
Papain	28	32	32	34	32	32	34	9
Pea 15a	34	36	30	41	43	42	32	10
Pea tpp	33	37	34	43	43	45	33	11
Human cathepsin L	33	40	37	43	44	45	38	12
Human cathepsin S	34	32	36	39	39	43	31	13
Human cathepsin H	31	36	37	34	38	34	30	14
Human cathepsin B	24	22	25	27	26	29	22	15

* References: 1, this paper; 2, J. S. Ikeda, J. W. Thomford & L. B. Corbeil, personal communication [GenBank accession numbers U13153 (TFCP1); U13154 (TFCP2)]; 3, Mallinson *et al.* (1994); 4, Pears *et al.* (1985); 5, Mottram *et al.* (1989); 6, Reed *et al.* (1993); 7, Heussler & Dobbelaere (1994); 8, Matsumoto *et al.* (1995); 9, Cohen *et al.* (1986); 10, Guerrero *et al.* (1990); 11, Granell *et al.* (1992); 12, Gal & Gottesman (1988); 13, Shi *et al.* (1992); 14, Fuchs & Gassen (1989); 15, Chan *et al.* (1986).

The nucleotide sequences of the *TFCP* genes are not presented here but are available in the EMBL database (accession numbers X87776–X87782 for *TFCP3–9*).

Predicted amino acid sequences of TFCP3–9

Fig. 1 shows the predicted amino acid sequences for TFCP3–9 aligned with sequences for two other cloned TFCPs, four *Trichomonas vaginalis* cysteine proteinases (TvCPs), *Dictyostelium discoideum* CP2, *Drosophila melanogaster* CP1 and human cathepsin L. The latter three are among the non-trichomonad CPs which showed greatest similarity to most of the TFCPs (Table 1). For six of the genes (*TFCP3*, *TFCP5–9*) the cloned fragment included the primer sequences at both ends, but these are not shown in Fig. 1. During cloning of *TFCP4*, however, a

deletion occurred and a short section, which included the whole of primer 2 and part of the vector, was missing. It is most likely that besides the primer sequence only one other base had been lost from the insert, hence one less amino acid is predicted for TFCP4, although the possibility that additional C-terminal amino acids were encoded by the original PCR fragment cannot entirely be excluded.

All the genes encoded CPs which are cathepsin-L-like in structure and their presence in *T. foetus* provides further evidence that CPs of the papain superfamily were present in the early eukaryotes (Mallinson *et al.*, 1994; Berti & Storer, 1995). Those residues which are highly conserved among eukaryotic CPs (see Berti & Storer, 1995) are present in most of the TFCP sequences. These include the active-site histidine (residue 163; this and all subsequent residue numbers correspond to those of human cathepsin

Fig. 2. Southern blot analysis of TFCP3–9 genes. *T. foetus* genomic DNA [10 µg (1, 3); 5 µg (2)] was digested with *EcoRV*, electrophoresed on a 0.8% agarose gel, transferred to nitrocellulose and hybridized individually to (a) TFCP3, (b) TFCP4, (c) TFCP5, (d) TFCP6, (e) TFCP7, (f) TFCP8 or (g) TFCP9. The strains used were (1) F2, (2) Manley and (3) Belfast. The sizes (in kb) of markers (from *lambda/HindIII*; Life Technologies) are indicated on the right.

L as given in Fig. 1) and other residues known to be important in the catalytic mechanism. The active-site cysteine and asparagine were encoded by the PCR primers and so lie outside the sequences shown. All cysteines involved in conserved disulphide bridges (positions 56, 65, 98 and 156) are also present. While detailed analysis of the amino acid sequences is not provided and will await the availability of complete sequences, some features are worth noting here. Two unusual cysteines previously found in TvCP sequences (Mallinson *et al.*, 1994), are also present in the majority of the TFCP sequences. These are present at position 62 in all trichomonad sequences, except TFCP4 and TFCP9, and position 167 in TFCP4, TFCP6, TFCP7 and TFCP8 as well as in TFCP1, TvCP1 and TvCP4. All the TFCP sequences include two extra residues between human cathepsin L residues 78 and 83, a feature characteristic of all the published flagellate CP sequences. A putative *N*-linked glycosylation site was present at position 140 in both TFCP7 and TFCP8.

Table 1 shows the similarity between the TFCPs and other trichomonad CPs and some representative non-trichomonad CPs. The seven TFCPs were considered to fall into one of three groups, a conserved group (TFCP6, TFCP7, TFCP8) which show the greatest similarity to one another (> 62% identity), a diverged group (TFCP3, TFCP5 and TFCP9), each of which has less than 44% identity with any of the other TFCP sequences and, in an intermediate position, TFCP4 which has 44–47% identity

with members of the conserved group. Within the conserved group, TFCP7 and TFCP8 are most closely related (119 residues in common, 78% identity). The CPs in this group are also similar to one of the other cloned *T. foetus* CPs, TFCP1 (J. S. Ikeda and co-workers, personal communication), and to the four TvCPs (Mallinson *et al.*, 1994). Among the diverged group, TFCP3 and TFCP9 are the least similar to one another (47 residues in common, 30% identity). The closest relationships between members of this group and other trichomonad CPs are those of TFCP5 and TFCP2 (48% identity) and TFCP9 and TvCP3 (47% identity). In general, the diverged TFCPs are almost as close to non-trichomonad CPs as they are to other TFCPs or TvCPs. In contrast, members of the conserved group are significantly closer to one another than to any non-trichomonad CP, but they are also, in most cases, more similar to non-trichomonad CPs than are the diverged TFCPs.

Genomic organization of *T. foetus* CP genes

The organization of the seven *T. foetus* CP genes was examined in F2, the clone whose DNA and RNA were used to derive the PCR fragments, and in two independent isolates, Manley and Belfast. DNA was digested using *EcoRV*, an enzyme which did not cut within any of the fragments, and Southern blot analyses were performed using complete cloned fragments as probes (Fig. 2). It

Fig. 3. Northern blot analysis of *TFCEP3-9* gene expression. Poly(A)⁺ RNA (2.5 µg) was electrophoresed on a 1.4% (w/v) agarose gel, transferred to nitrocellulose and hybridized to individual probes with times of exposure (in h) for the autoradiograms given in parentheses: (a) *TFCEP3* (120); (b) *TFCEP4* (120); (c) *TFCEP5* (48); (d) *TFCEP6* (72); (e) *TFCEP7* (42); (f) *TFCEP8* (1); (g) *TFCEP9* (72). The same strips were used for *TFCEP3* and *TFCEP7*, *TFCEP5* and *TFCEP8*, and *TFCEP6* and *TFCEP9*. The strains used were (1) F2, (2) Manley and (3) Belfast. RNA ladder sizes in kb (Life Technologies) are indicated on the right. The arrowhead on the left marks the position of the *TFCEP3* and *TFCEP4* transcripts (a, b); the arrowhead on the right marks the position of all other transcripts.

should be noted that the hybridization signal for the Manley strain is not as intense as for the other strains, as less DNA was loaded for this strain. Each probe gave a different pattern of bands with the three isolates. This confirmed that each was recognizing a unique gene under the stringency conditions used.

For three of the *TFCEP* genes (4, 6 and 8), the *EcoRV* fragment size was the same in all three isolates (approximately 8.2, 2.3 and 4.2 kb, respectively). With *TFCEP5* and *TFCEP7*, clone F2 differed from the Manley and Belfast isolates. The *TFCEP7* fragment size was 5.0 kb in F2 and 4.4 kb in Manley and Belfast. For *TFCEP5*, the band in F2 was 2.3 kb, and a faint band equivalent to this was also apparent in Belfast. However, the major *TFCEP5* band in Belfast was approximately 4.4 kb and a band of similar size was the only one found in Manley. The Manley strain shared a common *TFCEP9* band (approximately 10 kb) with F2, while in Belfast the band size was 8 kb.

TFCEP3 showed a more complex pattern. Three bands were apparent in F2 DNA (approximately 5.0, 6.6 and 7.2 kb), suggesting a multiple-copy gene in this clone. In each of the other strains only one of these bands was apparent, the 7.2 kb band in Manley and the 6.6 kb band in Belfast. No other *TFCEP3* bands were seen in Manley or Belfast DNA, even on prolonged exposure of the autoradiogram, indicating that these strains had only a single copy of the gene.

Expression of *T. foetus* CP genes

Northern blot analysis of poly(A)⁺ RNA from *T. foetus* F2 showed that all four of the CP genes amplified from genomic DNA were expressed, albeit at contrasting levels (Fig. 3). It also confirmed the presence of the mRNA for the three genes identified by RT-PCR. Single transcripts were recognized in every case. The transcripts of five of the genes (*TFCEP5-9*) were approximately 1.1 kb; those of *TFCEP3* and *TFCEP4* genes were slightly larger at

approximately 1.3 kb. These were similar in size to the *Trichomonas vaginalis* transcripts which encode CPs with typical proregions but no C-terminal extension (Mallinson *et al.*, 1994). It seems likely that the products of the *TFCEP* genes are similar in this respect.

The transcript of the *TFCEP8* gene was by far the most abundant (note the much shorter exposure time used), as had been predicted from the frequency with which its cDNA was encountered among RT-PCR clones. That of *TFCEP7*, another gene identified after RT-PCR, appeared to be the second most abundant. In clone F2, transcripts for *TFCEP5*, *TFCEP6* and *TFCEP9* were also readily detectable. Thus all three of the conserved group of genes (*TFCEP6-8*) were well expressed, as were two of the diverged genes (*TFCEP5*, *TFCEP9*). The transcripts of *TFCEP3* and, in particular, *TFCEP4* were present at very low levels. Gene expression was strain specific, and transcripts for two of the genes, both from the diverged group, were not detectable in one or both of the other isolates (Fig. 3). *TFCEP5* transcripts were absent from the Manley strain but were as abundant in the Belfast strain as in F2. Neither the Manley nor the Belfast strain appeared to express *TFCEP3*. This gene was multiple-copy in F2 but single-copy in the other two strains (Fig. 2), and it is possible that only the 5 kb fragment, present in F2 but absent from the other strains, represented a gene which could be expressed. Transcripts of the three conserved genes were present in both Manley and Belfast strains, as were those of two of the diverged genes (*TFCEP4*, *TFCEP9*).

Proteinase patterns

SDS-PAGE analysis of *T. foetus* proteinases using gels incorporating gelatin reveals multiple high-activity CPs (Lockwood *et al.*, 1987; North *et al.*, 1990a). Although the three isolates analysed in this study showed differential expression of some of the CP genes (Fig. 3), the proteinase

Fig. 4. Proteinase patterns in lysates of (1) F2, (2) Manley and (3) Belfast run on a gelatin-containing gel. Each lane was loaded with a sample containing 10 µg protein. The apparent molecular masses (in kDa) of the major proteinase bands (see North *et al.*, 1990a) are indicated on the right.

patterns were very similar (Fig. 4). Similar results have been obtained using fluorogenic substrates (not shown). Among the major proteinases the only difference was with a fast-running band seen in Manley and Belfast but apparently absent from F2. Since all the major proteinase bands detected in F2 were also present in Manley and Belfast, those genes which were not expressed in one or both of the latter strains, namely *TFCP3* and *TFCP5*, cannot encode any of these enzymes. These two genes are in the diverged group, and the enzymes encoded could have atypical properties which affect their stability and/or substrate specificity, preventing them from being observed using substrate-SDS-PAGE methods. Novel assay methods may be required to detect the products of these genes.

DISCUSSION

As judged by the complexity of the band patterns obtained by substrate-SDS-PAGE, trichomonads produce a greater number of CPs than most other organisms (North, 1991). This raises two major questions, the first of which concerns the molecular basis of the multiple forms, and the second the nature of the cellular processes which would require such a diverse array of CPs. Although it is not yet possible to provide full answers to either of these questions, the results presented here have offered some clues.

With respect to the molecular basis of the multiplicity, it is now clear that, like *Trichomonas vaginalis* (Mallinson *et al.*, 1994), *Tritrichomonas foetus* has several different CP genes. In the present study we have identified seven

genes, and as these are distinct from two others recently cloned from another *T. foetus* strain (J. S. Ikeda and co-workers, personal communication), it appears that this parasite has at least nine CP genes. At present the link between individual genes and enzymes remains to be established. The absence of the *TFCP3* transcript from the Manley and Belfast strains, and of the *TFCP5* transcript from the Manley strain, was not reflected in the proteinase band pattern (Fig. 4), and so the products of these two genes are not among those detected to date. An N-terminal amino acid sequence has been obtained for a 25 kDa *T. foetus* CP (Irvine *et al.*, 1993), but because the CP N-terminus lies outside the region encoded by the PCR fragments it was not possible to match this enzyme to any of the genes. However, another CP has recently been purified from *T. foetus* conditioned buffer (Thomford *et al.*, 1996), and this has an internal sequence (15 amino acids) which matches exactly the sequence predicted for the product of the *TFCP8* gene. Purified proteinase preparations give three low-molecular-mass activity bands (≤ 31 kDa) (Thomford *et al.*, 1996), and since the *TFCP8* sequence includes a putative N-linked glycosylation site (Fig. 1), these multiple activity bands could be due to differential carbohydrate modification of the product of this gene. *TFCP8* is one of the conserved group of genes, and it is possible that this group is entirely responsible for the high activity CPs detected using substrate-SDS-PAGE.

The diversity revealed here among the *TFCP* genes is much greater than that found previously for the *TvCP* genes. Nevertheless the proteinase pattern of *T. foetus* is less complex than that of *Trichomonas vaginalis*, with respect to both the number of enzymes and the range of apparent molecular masses (Lockwood *et al.*, 1987; North *et al.*, 1990a). Neale & Alderete (1990) have reported as many as 23 distinct activities in *T. vaginalis*. The four *TvCP* genes which have been characterized in most detail encode similar enzymes (the two most distantly related are 56% identical within the section shown in Fig. 1) and could be included with the conserved group of *TFCP* genes. It is possible that PCR amplification of genomic DNA, as used here, results in a wider range of CP DNAs being isolated than does screening a cDNA library, as used with *T. vaginalis*, not least because the latter will be biased in favour of well-expressed genes. However, PCR amplification was used to obtain a *T. vaginalis* CP DNA probe, and preliminary analysis of the PCR products suggested that the genes amplified were not as diverse as the *TFCPs* (D. J. Mallinson, unpublished results). One explanation for the difference between the species may be that many as yet undetected *T. vaginalis* CP genes are so distinct from cathepsin-L-like genes that they are not recognized by the standard PCR primers. However, it remains a possibility that *T. vaginalis* does not have as wide a range of CP genes but that the gene products undergo more diverse post-translational modification.

Multiple CP genes have been reported in a variety of species. There are at least 12 different mammalian CPs (Berti & Storer, 1995), including cathepsins B, C, H, K, L and O and calpains. By way of comparison, for the section

shown in Fig. 1, TFCP7 and TFCP8 are more similar to one another (119 identical residues) than is cathepsin L to either cathepsin S (78 identical residues) or cathepsin K (Inaoka *et al.*, 1995) (85 identical residues). Indeed, they are as close to one another as are human and mouse cathepsin Ls (116 identical residues). On the other hand, TFCP3 and TFCP9 (47 identical residues) are more distant from one another than is cathepsin H from cathepsin L (61 identical residues), and they show a similar degree of similarity as cathepsin O (Velasco *et al.*, 1994) does to cathepsin L and cathepsin S (50 and 47 identical residues, respectively).

The degree of diversity reported for cathepsin-L-like CPs in other parasites is less than that in *Tritrichomonas foetus*. Multiple CP genes are present in kinetoplastids (trypanosomes and leishmanias) but these include tandemly-linked copies of identical or closely related genes (Mottram *et al.*, 1989; Souza *et al.*, 1992; Lima *et al.*, 1994), and there is no evidence for equivalent sets of genes in trichomonads. A recent study based on RT-PCR has identified five cathepsin-L-like CP genes in the liver fluke *Fasciola hepatica* (Heussler & Dobbelaere, 1994). The two most diverged of these share 47% identity (within the section shown in Fig. 1) and the two most similar share 84% identity. Thus these represent a more closely related set of CPs than the TFCPs.

In multicellular organisms or unicellular organisms with complex life cycles, CP gene expression is often developmentally regulated (Berti & Storer, 1995). The presence of multiple CP genes may reflect differing requirements for proteolysis at particular developmental stages and in various specialized cells. It is widely thought that trichomonads do not undergo differentiation, and so there would appear to be no possibility of stage-specific CP expression. There is some evidence for two different morphological forms of *Trichomonas vaginalis*, although their significance is unclear, but the major CPs produced by these two forms are the same (Scott *et al.*, 1995). Anyway, the present study has shown, at least for *Tritrichomonas foetus* clone F2, that the seven TFCP genes are expressed simultaneously, suggesting that all the CPs are likely to be required at the same time. Whether CP gene expression is selective under conditions which are more physiological than axenic culture remains to be established. The major CPs of trichomonads are lysosomal and extracellular (Lockwood *et al.*, 1988), and so their major role is likely to be in the breakdown of extracellular proteins. It is difficult to envisage, however, why so many different CPs would be required to fulfil a general digestive role, and this raises the possibility that some of the CPs might have more selective functions. If this is so, then it might be predicted that the diverged group of CPs would be involved, since these are more likely to have the distinct properties required for a specialized role. It will be interesting to see whether the conserved and diverged TFCP groups not only have different functions but are also targeted to different locations within the cell. Analysis of the complete structure of individual genes should shed further light on this.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Wellcome Trust. We are grateful to Dr Mike Taylor (Weybridge) for providing us with the Belfast and Manley strains and to Dr Michael Leaver (Stirling) for assistance with some of the sequencing.

REFERENCES

- Berti, P. J. & Storer, A. C. (1995). Alignment/phylogeny of the papain superfamily of cysteine proteases. *J Mol Biol* **246**, 273–283.
- Burgess, D. E., Knoblock, K. F., Daugherty, T. & Robertson, N. P. (1990). Cytotoxic and hemolytic effects of *Tritrichomonas foetus* on mammalian cells. *Infect Immun* **58**, 3627–3632.
- Chan, S. J., San Segundo, B., McCormick, M. B. & Steiner, D. F. (1986). Nucleotide and predicted amino acid sequences of cloned human and mouse preprocathepsin B cDNAs. *Proc Natl Acad Sci USA* **83**, 7721–7725.
- Cohen, L. W., Coghlan, V. M. & Dihel, L. C. (1986). Cloning and sequencing of papain-encoding cDNA. *Gene* **48**, 219–227.
- Corbeil, L. B. (1994). Vaccination strategies against *Tritrichomonas foetus*. *Parasitol Today* **10**, 103–106.
- Davis, L. G., Dibner, M. D. & Battey, J. F. (1986). *Basic Methods in Molecular Biology*. New York: Elsevier.
- Diamond, L. (1957). The establishment of various trichomonads of animals and man in axenic culture. *J Parasitol* **43**, 488–490.
- Fuchs, R. & Gassen, H. G. (1989). Nucleotide sequence of human preprocathepsin H, a lysosomal cysteine proteinase. *Nucleic Acids Res* **17**, 9471.
- Gal, S. & Gottesman, M. M. (1988). Isolation and sequence of a cDNA for human procathepsin L. *Biochem J* **253**, 303–306.
- Granell, A., Harris, N., Pisabarro, A. G. & Carbonell, J. (1992). Temporal and spatial expression of a thiolprotease gene during pea ovary senescence, and its regulation by gibberellin. *Plant J* **2**, 907–915.
- Guerrero, F. D., Jones, J. T. & Mullet, J. E. (1990). Turgor-responsive gene transcription and RNA levels increase rapidly when pea shoots are wilted. Sequence and expression of three inducible genes. *Plant Mol Biol* **15**, 11–26.
- Heussler, V. T. & Dobbelaere, D. A. E. (1994). Cloning of a protease gene family of *Fasciola hepatica* by the polymerase chain reaction. *Mol Biochem Parasitol* **64**, 11–23.
- Inaoka, T., Bilbe, G., Ishibashi, O., Tezuka, K., Kumegawa, M. & Kokubo, T. (1995). Molecular cloning of human cDNA for cathepsin K: novel cysteine proteinase predominantly expressed in bone. *Biochem Biophys Res Commun* **206**, 89–96.
- Irvine, J. W., Coombs, G. H. & North, M. J. (1993). Purification of cysteine proteinases from trichomonads using bacitracin Sepharose. *FEMS Microbiol Lett* **110**, 113–120.
- Lima, A. P. C., Tessier, D. C., Thomas, D. Y., Scharfstein, J., Storer, A. C. & Vernet, T. (1994). Identification of new cysteine protease gene isoforms in *Trypanosoma cruzi*. *Mol Biochem Parasitol* **67**, 333–338.
- Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441.
- Lockwood, B. C., North, M. J. & Coombs, G. H. (1984). *Trichomonas vaginalis*, *Tritrichomonas foetus* and *Trichomitus batrachorum*: comparative proteolytic activity. *Exp Parasitol* **58**, 245–253.
- Lockwood, B. C., North, M. J., Scott, K. I., Bremner, A. F. & Coombs, G. H. (1987). The use of a highly sensitive electrophoretic method to compare the proteinases of trichomonads. *Mol Biochem Parasitol* **24**, 89–95.

- Lockwood, B. C., North, M. J. & Coombs, G. H. (1988). The release of hydrolases from *Trichomonas vaginalis* and *Tritrichomonas foetus*. *Mol Biochem Parasitol* **30**, 135–142.
- McLaughlin, J. & Müller, M. (1979). Purification and characterization of a low molecular weight thiol proteinase from the flagellate protozoon *Tritrichomonas foetus*. *J Biol Chem* **254**, 1526–1533.
- Mallinson, D. J., Lockwood, B. C., Coombs, G. H. & North, M. J. (1994). Identification and molecular cloning of four cysteine proteinase genes from the pathogenic protozoon *Trichomonas vaginalis*. *Microbiology* **140**, 2725–2735.
- Matsumoto, I., Watanabe, H., Abe, K., Arai, S. & Emori, Y. (1995). A putative digestive cysteine proteinase from *Drosophila melanogaster* is predominantly expressed in the embryonic and larval midgut. *Eur J Biochem* **227**, 582–587.
- Mottram, J. C., North, M. J., Barry, J. D. & Coombs, G. H. (1989). A cysteine proteinase cDNA from *Trypanosoma brucei* predicts an enzyme with an unusual C-terminal extension. *FEBS Lett* **258**, 211–215.
- Müller, M. (1973). Biochemical cytology of trichomonad flagellates. I. Subcellular localization of hydrolases, dehydrogenases, and catalase of *Tritrichomonas foetus*. *J Cell Biol* **57**, 453–474.
- Neale, K. A. & Alderete, J. F. (1990). Analysis of the proteinases of representative *Trichomonas vaginalis* isolates. *Infect Immun* **58**, 157–162.
- North, M. J. (1991). Proteinases of trichomonads and *Giardia*. In *Biochemical Protozoology*, pp. 234–244. Edited by G. H. Coombs & M. J. North. London: Taylor & Francis.
- North, M. J. (1994). Cysteine endopeptidases of parasitic protozoa. *Methods Enzymol* **244**, 523–539.
- North, M. J., Robertson, C. D. & Coombs, G. H. (1990a). The specificity of trichomonad cysteine proteinases analysed using fluorogenic substrates and specific inhibitors. *Mol Biochem Parasitol* **39**, 183–194.
- North, M. J., Mottram, J. C. & Coombs, G. H. (1990b). Cysteine proteinases of parasitic protozoa. *Parasitol Today* **6**, 270–274.
- Pears, C. J., Mahbubani, H. M. & Williams, J. G. (1985). Characterization of two highly diverged but developmentally co-regulated cysteine proteinase genes in *Dictyostelium discoideum*. *Nucleic Acids Res* **13**, 8853–8866.
- Reed, S. L., Bouvier, J., Pollack, A. S., Engel, J. C., Brown, M., Hirata, K., Que, X., Eakin, A., Hagblom, P., Gillin, F. & McKerrow, J. H. (1993). Cloning of a virulence factor of *Entamoeba histolytica*: pathogenic strains possess a unique cysteine proteinase gene. *J Clin Invest* **91**, 1532–1540.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**, 5463–5467.
- Scott, D. A., North, M. J. & Coombs, G. H. (1995). *Trichomonas vaginalis*: amoeboid and flagellated forms synthesise similar proteinases. *Exp Parasitol* **80**, 345–348.
- Shi, G. P., Munger, J. S., Meara, J. P., Rich, D. H. & Chapman, H. A. (1992). Molecular cloning and expression of human alveolar macrophage cathepsin S, an elastolytic cysteine protease. *J Biol Chem* **267**, 7258–7262.
- Souza, A. E., Waugh, S., Coombs, G. H. & Mottram, J. C. (1992). Characterization of a multicopy gene for a major stage-specific cysteine proteinase of *Leishmania mexicana*. *FEBS Lett* **311**, 124–127.
- Talbot, J. A., Nielsen, K. & Corbeil, L. B. (1991). Cleavage of proteins of reproductive secretions by extracellular proteinases of *Tritrichomonas foetus*. *Can J Microbiol* **37**, 384–390.
- Thomford, J. W., Talbot, J. A., Ikeda, J. S. & Corbeil, L. B. (1996). Characterization of extracellular proteinases of *Tritrichomonas foetus*. *J Parasitol* (in press).
- Velasco, G., Ferrando, A. A., Puente, X. S., Sánchez, L. M. & López-Otín, C. (1994). Human cathepsin O. Molecular cloning from a breast carcinoma, production of the active enzyme in *Escherichia coli*, and expression analysis in human tissues. *J Biol Chem* **269**, 27136–27142.
- Yule, A., Skirrow, S. Z. & BonDurant, R. H. (1989). Bovine trichomoniasis. *Parasitol Today* **5**, 373–377.

Received 22 May 1995; revised 29 July 1995; accepted 14 August 1995.