

Evolutionary Relationships and Protein Domain Architecture in an Expanded Calpain Superfamily in Kinetoplastid Parasites

Klaus Ersfeld,¹ Helen Barraclough,^{2,*} Keith Gull³

¹ Department of Biological Sciences, University of Hull, Hull HU6 7RX, UK

² School of Biological Sciences, University of Manchester, Oxford Road, Manchester M13 9PT, UK

³ Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK

Received: 3 September 2004 / Accepted: 8 July 2005 [Reviewing Editor: Dr. John Oakeshott]

Abstract. Employing whole-genome analysis we have characterized a large family of genes coding for calpain-related proteins in three kinetoplastid parasites. We have defined a total of 18 calpain-like sequences in *Trypanosoma brucei*, 27 in *Leishmania major*, and 24 in *Trypanosoma cruzi*. Sequence characterization revealed a well-conserved protease domain in most proteins, although residues critical for catalytic activity were frequently altered. Many of the proteins contain a novel N-terminal sequence motif unique to kinetoplastids. Furthermore, 24 of the sequences contain N-terminal fatty acid acylation motifs indicating association of these proteins with intracellular membranes. This extended family of proteins also includes a group of sequences that completely lack a protease domain but is specifically related to other kinetoplastid calpain-related proteins by a highly conserved N-terminal domain and by genomic organization. All sequences lack the C-terminal calmodulin-related calcium-binding domain typical of most mammalian calpains. Our analysis emphasizes the highly modular structure of calpains and calpain-like proteins, suggesting that they are involved in diverse cellular functions. The discovery of this surprisingly large family of calpain-like proteins in lower eukaryotes that combines novel and conserved sequence modules contributes to our understanding of the evolution of this abundant protein family.

Key words: Calpain — Trypanosoma — Leishmania — Parasites — Genome — Multigene family — Protein domains — Microtubules — Cytoskeleton — Acylation

Introduction

Calpains (EC 3.4.22.17; clan CA, family C02) constitute a large family of calcium-dependent intracellular cysteine proteases (for a recent review see Goll et al. 2003). Typical calpains, such as m- and μ -calpain, have been characterized mainly in mammals. Although their precise functions are poorly understood, calpains are involved in cytoskeletal remodeling, signal transduction, and cell differentiation (Ono et al. 1998; Sato and Kawashima 2001). Mutations in some calpain genes are also linked with diseases such as type 2 diabetes and limb-girdle muscular dystrophy (Richard et al. 1995; Horikawa et al. 2000; Huang and Wang 2001; Taveau et al. 2003).

Calpains are heterodimeric proteins, consisting of a large subunit of approximately 80 kD and a small subunit of 28 kD. The large subunit is typically divided into four domains (Hosfield et al. 1999a; Strobl et al. 2000). The N-terminal domain I is only 18–20 amino acid residues in length and of unknown function. Domain II (330 amino acids) forms the catalytic core of the enzyme, characterized by the presence of the catalytic triad containing the amino acids cysteine, histidine, and asparagine. The sequence is highly

*Present address: Corbett Research, Centre for Immunology, St. Vincents Hospital, Cnr. Boundary and West Street, Sydney, NSW 2011, Australia

Correspondence to: Klaus Ersfeld; email: k.ersfeld@hull.ac.uk

conserved between different members of the calpain family (Berti and Storer 1995). In addition, domain II contains two calcium-binding sites that are essential for enzyme activity (Moldoveanu et al. 2002). Domain III (150 amino acids) is described as a linker between the catalytically active domain II and domain IV, transmitting and amplifying conformational changes between the two adjacent domains. Domain IV (170 amino acids) contains five EF-hand motifs able to bind calcium. This activity is essential for enzyme activity and also for the dimerization with the small subunit. The sequence of domain IV is related to calmodulin, leading to the hypothesis that calpains are modular and have evolved by fusion of a calmodulin-like gene with a protease precursor domain.

In addition to typical calpains that conform to the four-domain structure, unconventional calpains have been identified in mammals and in other eukaryotes (Dear et al. 1997; Sorimachi et al. 1997; Margis and Margis-Pinheiro 2003). These proteins are referred to as calpain-like proteins or calpain homologues (Ono et al. 1998). The major differences to conventional calpains are amino acid changes within the catalytic triad and the lack of an EF-hand-containing domain IV. It is unknown whether any of the atypical calpains have enzymatic/proteolytic activities. Calpain-like proteins have mainly been found in invertebrates and lower eukaryotes. Genetic evidence shows that many of these atypical calpains are, like conventional calpains, involved in signal transduction cascades, tissue differentiation, and sex determination. In *Caenorhabditis elegans*, the calpain-like protein TRA-3 is involved in a pathway leading to female worms (Barnes and Hodgkin 1996). TRA-3 lacks a typical domain IV. Interestingly, two homologues of TRA-3 in humans, CAPN5 and CAPN6, also lack domain IV and their tissue-specific expression pattern suggests involvement in sexual development (Dear et al. 1997; Dear and Boehm 1999). Calpain-like proteins have also been identified in the fungi *Emericella nidulans* (PalB) and *Saccharomyces cerevisiae* (Clp1p), where they function in adaptation of growth under alkaline conditions (Denison et al. 1995). A single gene for a calpain-like protein, also lacking a typical domain IV, has been found in maize, *Arabidopsis*, and other plants (Margis and Margis-Pinheiro 2003).

In mammals, calpains form a large gene family with currently 15 members in humans. In the nematode *C. elegans* 14 genes for calpain-like proteins, all lacking domain IV, have been identified. Four calpain-like proteins have been identified in *Drosophila*, and other arthropods, such as crustaceans, seem to have a similar number. Most fungal genomes sequenced have only a single gene for calpain-like proteins, the exception being *Neurospora crassa*, where three genes have been identified (data from <http://www.merops.ac.uk>). Little information is

available from protozoan organisms. The malaria parasite *Plasmodium falciparum*, whose genome sequencing project has recently been completed, has only a single gene for a calpain-like protein of unknown function (Wu et al. 2003).

Previously we have shown the presence of calpain-like proteins in the parasite *Trypanosoma brucei* (Hertz-Fowler et al. 2001). The protein CAP5.5 is expressed only during the insect stage of the life cycle and the differentiation process (Matthews and Gull 1994). Similar to other calpain-like proteins, CAP5.5 lacks a typical domain IV and has replaced the catalytically active amino acids cysteine and histidine with serine and tyrosine. In this study we present a systematic analysis of calpain-like proteins in the three kinetoplastid protozoa *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania major* (Order Kinetoplastida, Family Trypanosomatidae), representative of "ancient" organisms that diverged early from the eukaryotic evolutionary lineage (Stevens et al. 1999). We show the presence of a large and diverse family of calpain-like proteins in all three organisms, often organized in syntenic gene clusters. The data provide further evidence for the hypothesis that calpains evolved as a result of modular gene fusion events. The presence of numerous calpain-like proteins, exceeding the numbers found in most other organisms including vertebrates, and their unique protein architecture point to important and organism-specific functions for these proteins.

Methods

Database Mining and Sequence Analysis

The genome databases for predicted coding sequences of *T. brucei*, *T. cruzi*, and *L. major* at <http://www.genedb.org> were searched using tBLASTn with previously described calpain-like proteins of *T. brucei* as query sequences or by searching for the term "calpain*" using the databases' own annotations. Multiple alignments were done using ClustalW at <http://www.ebi.ac.uk> or, for smaller datasets, TCOFFEE at <http://www.ch.embnet.org>. After manual editing alignments were displayed and shaded with Boxshade at <http://www.ch.embnet.org> and further annotated in Adobe Illustrator. Assignment of sequences to the calpain (C02)-family of proteases was confirmed by searching the peptidase database at <http://www.merops.ac.uk> against all kinetoplastid calpains. Phylogenetic analysis was done from edited ClustalW alignments using the neighbor-joining algorithm with 1000 bootstrap trials as implemented in PAUP*4.0. Potential N-terminal myristoylation of proteins was predicted using MYR Predictor at <http://mendel.imp.univie.ac.at>. All proteins were scanned for motifs using InterPro at <http://www.ebi.ac.uk> or SMART at <http://smart.embl-heidelberg.de>. Analysis for coiled-coil protein domains was done using COILS at <http://www.ch.embnet.org> and secondary structure predictions were performed with PSA at <http://bmerc-www.bu.edu/psa>.

Nomenclature

To describe members of the calpain superfamily, we have adopted the nomenclature used by Goll et al. (2003) in their comprehensive

recent review of the calpain system. Typical or conventional calpains are closely related to mammalian m- and μ -calpain (also called Capn1 and Capn2). They are characterized by the presence of a defined four-domain structure supported by crystallographic data (Hosfield et al. 1999b; Strobl et al. 2000), where domain II contains the catalytic triad of cysteine, histidine, and arginine and domain IV contains multiple, functional calcium-binding EF-hand motifs. Atypical or unconventional calpains are described as calpain-like proteins and contain only a domain II consensus signature, with the catalytic triad not necessarily intact, and no EF-hand-containing domain IV.

Due to the absence of experimental evidence as to the function of any of the calpain-related proteins in kinetoplastids, the term orthologue is used here to describe putative functionally equivalent genes in different kinetoplastid species, based entirely on sequence similarity and the presence of shared motifs (such as acylation signals).

Sequences presented in this study were given systematic names based on their assignment to a particular chromosome and to the relative position within a gene cluster (Table 1). For example, TbCALP1.3 describes a *Trypanosoma brucei* calpain-like protein located on chromosome 1 in the third position of a calpain gene cluster. In *Trypanosoma cruzi* individual genes have not been assigned to chromosomes yet. Therefore, chromosome numbers were replaced by the preliminary prefix x and y for those calpain-like proteins that are part of clusters. "Orphan" calpain-like proteins in *T. cruzi* that could not be allocated to a particular cluster are named by using the last five or six digits of their GeneDB identity number. Since many GeneDB numbers for all three organisms are preliminary, we have also listed the first eight N-terminal amino acids of each protein to enable unambiguous identification using BLAST searches (Table 1).

Results

Sequence Discovery

The discovery in many organisms of atypical calpain-like proteins that lack similarity to the Ca^{2+} -binding, EF-hand-containing domain IV limits the definition of calpain-like proteins to the presence of significant similarities within the protease domain II. Using domain II of the previously identified calpain-like protein CAP5.5 of *Trypanosoma brucei* as a bait, we identified 12 calpain-related sequences in *T. brucei*, 17 sequences in *Leishmania major* and 15 sequences in *T. cruzi* (Table 1 and Fig. 1). However, when we analyzed the full-length sequences it became clear that the N-terminal sequences outside domain II of many of these sequences showed significant similarities to each other. This sequence element, corresponding to the location of domain I in conventional calpains but significantly longer (~100 aa), was subsequently found not only at the N-terminus of kinetoplastid calpain-like proteins, but also as the core element of a large number of short open reading frames, with an average length of approximately 200 amino acids. Six short sequences were found in *T. brucei*, ten in *L. major* and nine in *T. cruzi*. These short sequences have no similarities to any other sequences currently in the databases. They are exclusively present either as short "stand-alone" genes or in conjunction with

domain II-containing calpain-like proteins, representing the domain I equivalent. This unique correlation between calpain-like proteins and the short sequences was the rationale to include the short sequences in this analysis. However, in order to clearly distinguish the short, domain I-only sequences (that lack the typical calpain domain II) and the multidomain sequences (that contain the typical calpain domain II), we termed the short sequences SKCRPs (small kinetoplastid calpain-related proteins) and the multidomain proteins CALPs (calpain-like proteins). In summary, we classified a total of 18 sequences in *T. brucei* (12 CALPs, 6 SKCRPs), 27 in *L. major* (17 CALPs, 10 SKCRPs), and 24 in *T. cruzi* (15 CALPs, 9 SKCRPs). Specific details of the nomenclature are described under Methods.

Sequence Characterization

Domain I

The domain I equivalent of kinetoplastid calpain-like proteins can be grouped into two categories. The first group is characterized by the presence of this sequence domain both at the N-terminus of most domain II-containing calpain-like proteins (CALPs) and in the short calpain-like proteins lacking domain II (SKCRPs). The sequence showed no similarities to other proteins and therefore there is no indication as to its function. Comparison between the different kinetoplastids revealed a high degree of sequence conservation (Suppl. 1). On average, sequence identity was approximately 50%, and similarity 70%. Putative orthologues between species, defined here on the basis of sequence and motif similarity, showed values of up to 60% and 80%, respectively. The sequence is characterized by the presence of mainly three highly conserved motifs: glycine, leucine, leucine, phenylalanine, or tyrosine (GLLF/Y) toward the N-terminus; tryptophane, alanine, phenylalanine, asparagine, aspartate, and threonine (WAFYNDT) in the center; and valine, tyrosine, proline, any, glutamate, threonine, glutamate (VYPxETE) toward the C-terminus. We have termed this sequence element domain I^K (K for kinetoplastids).

The second group comprises N-terminal sequence domains that are heterogeneous in both intra- and interspecies comparisons (data not shown) and occurs only in proteins classified as kinetoplastid CALPs. Unlike domain I^K this heterogeneous domain I is not found within any of the short sequences of SKCRPs. Similarities are only significant between putative orthologues. These sequences are also not related to any motifs present in the databases. The exception was the orthologous pair TbCALP6.1 and TcCALP507.70. The domain I equivalent of these two sequences showed a significant degree of homology to the N-terminus of the regulatory

Table 1. Summary of calpain-like proteins in *Trypanosoma brucei*, *Leishmania major*, and *Trypanosoma cruzi*

| Name | GeneDB ID | Chromosome | N-terminus | Catalytic triad | Length (aa) | Domains | Group |
|---------------------------|------------------------|------------|------------|-----------------|-------------|-----------------|-------|
| <i>Trypanosoma brucei</i> | | | | | | | |
| TbCALP1.1 | Tb927.1.2100 | 1 | MSEFELMS | SHN | 1146 | I,II,III,C | 1 |
| TbCALP1.2 | Tb927.1.2110 | 1 | MSSRNGSL | SHN | 682 | I,II,III,C | 1 |
| TbCALP1.3 | Tb927.1.2120 | 1 | MGCGASST | YHN | 732 | I,II,III,C | 1 |
| TbSKCRP1.4 | Tb927.1.2150 | 1 | MGLFLSKQ | | 130 | I | 3 |
| TbSKCRP1.5 | Tb927.1.2160 | 1 | MGCQSTV | | 140 | I | 3 |
| TbSKCRP1.6 | Tb927.1.2230 | 1 | MGCGGSKT | | 123 | I | 3 |
| TbSKCRP1.7 | Tb927.1.2260 | 1 | MGCGGSTT | | 130 | I | 3 |
| TbCALP4.1CAP5.5 | Tb04.1D20.740 | 4 | MGCGGSKV | SYN | 853 | I,II,III,C | 1 |
| TbCALP4.2 | Tb04.1D20.770 | 4 | MTSSGGLL | SSN | 784 | I,II,III,C | 1 |
| TbCALP6.1 | Tb06.2N9.670 | 6 | MVVSLSGY | SHS | 1010 | I,II,III,C | 2 |
| TbSKCRP7.1 | Tb07.5F10.560 | 7 | MSDITYEN | | 115 | I | 3 |
| TbCALP9.1 | Tb09.160.5550 | 9 | MMQASADG | CYN | 856 | I,II,III,C | 1 |
| TbCALP10.1 | Tb10.389.0470 | 10 | MQVPTDQL | KT- | 678 | I,II,III,C | 2 |
| TbCALP10.2 | Tb10.70.5950 | 10 | MATLDDL | -HH | 1594 | I,II,III,C | 2 |
| TbSKCRP10.3 | Tb10.70.5060 | 10 | MTELVPEA | | 484 | I | 3 |
| TbCALP11.1 | Tb11.47.0035 | 11 | MNYQAITE | CQN/CQN/CQN | 6145 | 3 × II, 2 × III | 4 |
| TbCALP11.2 | Tb11.47.0036 | 11 | MNEHSHEL | YCC | 1389 | I,II,III,C | 5 |
| TbCALP11.3 | Tb11.02.2240 | 11 | MRHIQRIS | CHN | 1053 | I,II,III,C | 2 |
| <i>Leishmania major</i> | | | | | | | |
| LmCALP4.1 | LmjF04.0450 | 4 | MSYPPEV | CHN | 854 | I,II,III,C | 1 |
| LmSKCRP14.1 | LmjF14.0850 | 14 | MSEIKYEN | | 115 | I | 3 |
| LmSKCRP17.1 | LmjF17.0870 | 17 | MTTADDVC | | 444 | I | 3 |
| LmCALP18.1 | LmjF18.1060 | 18 | MHIKRNKN | -IA | 774 | I,II,III,C | 2 |
| LmCALP20.1 | LmjF20.1180 | 20 | MSQDQLND | SCN | 977 | I,II,III,C | 1 |
| LmCALP20.2 | LmjF20.1190 | 20 | MTAMNGKV | SHN | 687 | I,II,III,C | 1 |
| LmCALP20.3 | LmjF20.1200 | 20 | MFDEFRED | SYN | 743 | I,II,III,C | 1 |
| LmCALP20.4 | LmjF20.1210 | 20 | MAVAEHT | -MK | 414 | I,II,III,C | 1 |
| LmSKCRP20.5 | LmjF20.1220 | 20 | MGCFNSTD | | 146 | I | 3 |
| LmSKCRP20.6 | LmjF20.1230 | 20 | MGRISKE | | 133 | I | 3 |
| LmSKCRP20.7 | LmjF20.1250 | 20 | MPIESYLV | | 236 | I | 3 |
| LmSKCRP20.8 | LmjF20.1280 | 20 | MGNKQSHA | | 152 | I | 3 |
| LmSKCRP20.9 | LmjF20.1300 | 20 | MGGSNSTN | | 136 | I | 3 |
| LmSKCRP20.10 | LmjF20.1310 | 20 | MGCGASSE | | 131 | I | 3 |
| LmCALP25.1 | LmjF25.1480 | 25 | MGNGCCGC | CHN | 705 | I,II,III,C | 1 |
| LmSKCRP25.2 | LmjF25.1750 | 25 | MGCGYSTY | | 142 | I | 3 |
| LmCALP27.1 | LmjF27.0490 | 27 | MAASSIHE | CQN/CQN/-CN | 4553 | 3 × II, 3 × III | 4 |
| LmCALP27.2 | LmjF27.0500 | 27 | MSVEVPD | CQN/CGN/-CN | 6164 | 3 × II, 3 × III | 4 |
| LmCALP27.3 | LmjF27.0510 | 27 | MSSPFTEV | CQN/CRN/-CN | 5358 | 3 × II, 3 × III | 4 |
| LmCALP31.1 | LmjF31.0390 | 31 | MYVCITSA | THN | 855 | I,II,III,C | 1 |
| LmCALP31.2 | LmjF31.0400 | 31 | MSVTKAAI | HHN | 932 | I,II,III,C | 1 |
| LmCALP31.3 | LmjF31.0410 | 31 | MSSAQDEF | HIS | 754 | I,II,III,C | 1 |
| LmCALP31.4 | LmjF31.0430 | 31 | MDGNAIRS | VMC | 938 | I,II,III,C | 1 |
| LmCALP31.5 | LmjF31.0440 | 31 | MGCCNCKE | SYN | 723 | I,II,III,C | 1 |
| LmCALP31.6 | LmjF31.0460 | 31 | MTAILLIL | CHN | 732 | I,II,III,C | 1 |
| LmCALP33.1 | LmjF33.2010 | 33 | MPSSVPAE | CHN | 1379 | I,II,III,C | 2 |
| LmSKCRP34.1 | LmjF34.0190 | 34 | MTEEVAP | | 491 | I | 3 |
| <i>Trypanosoma cruzi</i> | | | | | | | |
| TcSKCRPx.1 | Tc00.1047053509003.30 | X | MGCGASSK | | 131 | I | 3 |
| TcSKCRPx.2 | Tc00.1047053506563.70 | X | MGCGASSR | | 133 | I | 3 |
| TcSKCRPx.3 | Tc00.1047053506563.90 | X | MGCGPSTQ | | 209 | I | 3 |
| TcSKCRPx.4 | Tc00.1047053506563.110 | X | MGCGSSKS | | 147 | I | 3 |
| TcSKCRPx.5 | Tc00.1047053506563.120 | X | MGGGASLA | | 169 | I | 3 |
| TcSKCRPx.6 | Tc00.1047053506563.130 | X | MPLQIFFS | | 380 | I | 3 |
| TcSKCRPx.7 | Tc00.1047053506563.170 | X | MGSSSSKL | | 130 | I | 3 |
| TcSKCRPx.8 | Tc00.1047053506563.180 | X | MGSSTSKH | | 140 | I | 3 |
| TcCALPx.9 | Tc00.1047053506563.190 | X | MGNNGSST | QNN | 858 | I,II,III,C | 1 |
| TcCALPx.10 | Tc00.1047053506563.200 | X | MGCGASTT | SYN | 729 | I,II,III,C | 1 |
| TcCALPx.11 | Tc00.1047053506563.210 | X | MASYGPKD | SHN | 692 | I,II,III,C | 1 |
| TcCALPx.12 | Tc00.1047053508999.190 | X | MSAENQWE | AHN | 837 | I,II,III,C | 1 |
| TcCALPy.1 | Tc00.1047053509237.120 | Y | MGLGFFCG | SSG | 715 | I,II,III,C | 1 |
| TcCALPy.2 | Tc00.1047053508555.60 | Y | MGCCASKE | NYN | 806 | I,II,III,C | 1 |
| TcCALPy.3 | Tc00.1047053508555.70 | Y | MRRPHHTA | HHN | 1231 | I,II,III,C | 1 |

Continued

Table 1. Continued

| Name | GeneDB ID | Chromosome | N-terminus | Catalytic triad | Length (aa) | Domains | Group |
|---------------|------------------------|------------|------------|-----------------|-------------|-----------------|-------|
| TcCALP855.40 | Tc00.1047053503855.40 | ? | MPFLSDRA | CHN | 850 | I,II,III,C | 1 |
| TcCALP227.130 | Tc00.1047053506227.130 | ? | MSVEALTC | CHN | 1067 | I,II,III,C | 2 |
| TcCALP269.70 | Tc00.1047053511269.70 | ? | MAFFFCFL | KMC | 770 | I,II,C | 2 |
| TcCALP507.70 | Tc00.1047053511507.70 | ? | MTTYFAGL | SHS | 1030 | I,II,III,C | 2 |
| TcSKCRP407.50 | Tc00.1047053506407.50 | ? | MTDLVDEI | | 488 | I | 3 |
| TcCALP721.30 | Tc00.1047053506721.30 | ? | MAVAVSAE | LCN/CQN/CQN | 4571 | 3 × II, 3 × III | 4 |
| TcCALP441.10 | Tc00.1047053511441.10 | ? | MNARAQEL | Y– | 1256 | I,II | 5 |
| TcCALP121.170 | Tc00.1047053510121.170 | ? | MTTLEDLY | –HH | 1592 | I,II,III,C | 2 |
| TcCALP329.10 | Tc00.1047053511329.10 | ? | MASLAESR | –CN | 454 | I,II,IIIC | 1 |

Note. Given are the systematic name, GeneDB identity number of the corresponding GeneDB parasite database at <http://www.genedb.org>, chromosomal location, N-terminal eight-amino-acid residue (potentially acylated residues are underlined), amino acid residues that align with the catalytic triad in human μ -calpain (accession no. AAH08751), number of amino acids, domain structure, and classification (see Fig. 3 and text for details).

subunit of cAMP-dependent kinases identified in *T. brucei* and *T. cruzi* (Suppl. 2). The domain I of this second group was labeled I^H (H for heterogeneous).

Domain II

With the exception of the short calpain-like proteins (SKCRPs) which contain only the domain I^K equivalent, all other sequences were identified on the basis of their similarity to the catalytic domain II of conventional calpains (Fig. 1). Some of the sequences contained three copies of domain II. These sequences are discussed separately below. A typical domain II contains a number of motifs that are unique to calpain proteases and distinguish this protease family from related cysteine proteases, such as papains (Ono et al. 1998). Critical amino acid residues are located close to the amino acids of the catalytic triad. Aspartate, preceding the active cysteine, and proline, next to the active asparagine, are indicative of calpains. In the majority of sequences these amino acids are conserved, but in a significant number they are replaced by other amino acids. A further motif unique to calpains is the sequence lysine, alanine, tyrosine, alanine, lysine (KAYAK-motif) in the center of the domain. Its particular function is not known but it represents a motif that was highly conserved in all kinetoplastid calpain-like proteins. The overall identity of domain II of kinetoplastids compared with conventional domain II of a catalytically active calpain is approximately 25% (45% similarity), but the alignment showed local clusters of much higher agreement. A complete conservation of the catalytic triad cysteine, histidine, arginine (C,H,N) was observed in only 7 of a total of 44 domain II sequences. Similar deviations from the classical C,H,N motif have been observed in calpains from other organisms, including humans. The absence of amino acid residues essential for catalytic activity and the moderate overall degree of sequence identity suggest that most calpain-like proteins do not act as cysteine proteases.

The proteolytic activity of conventional calpains is regulated by binding of Ca²⁺ to the EF-hand located within domain IV. Recent crystallographic studies, have, however, revealed that also domain II of conventional calpains is able to bind calcium via an EF-hand-independent mechanism (Moldoveanu et al. 2002, 2004). The coordinated binding of two Ca²⁺ ions in domain II is essential to align the catalytic amino acids within the active site. Five amino acid residues that are critical for binding of calcium within domain II have been identified in mammalian calpains. Analysis of the corresponding positions in calpain-like proteins of kinetoplastids revealed that these amino acid residues are partially conserved in some sequences (Suppl. 3). In one of the sequences (LmCALP25.1) all five residues were present, indicating that this protein may bind calcium. Two of the sequences in this group, including LmCALP25.1, also showed complete conservation of the catalytic triad residues.

Domain III

It is thought that the conformational changes induced by the binding of Ca²⁺ to the EF hands of domain IV of conventional calpains are transmitted and amplified to the catalytic domain II via domain III, thereby regulating enzymatic activity (Strobl et al. 2000). Although none of the kinetoplastid sequences has EF-hand domains IV, the majority of sequences that contained domain II showed some degree of similarity (15–23% identity, 25–35% similarity) to domain III of conventional calpains (Suppl. 4, A). Moreover, secondary structure analysis predicted a high probability of β -strand conformation in this region, a feature consistent with the presence of an eight-stranded antiparallel β -sandwich in domain III of typical calpains (Suppl. 4, B) (Hosfield et al. 1999a; Strobl et al. 2000; Reverter et al. 2001). A possible explanation for the presence of domain III in trypanosomatid calpain-like proteins could be the

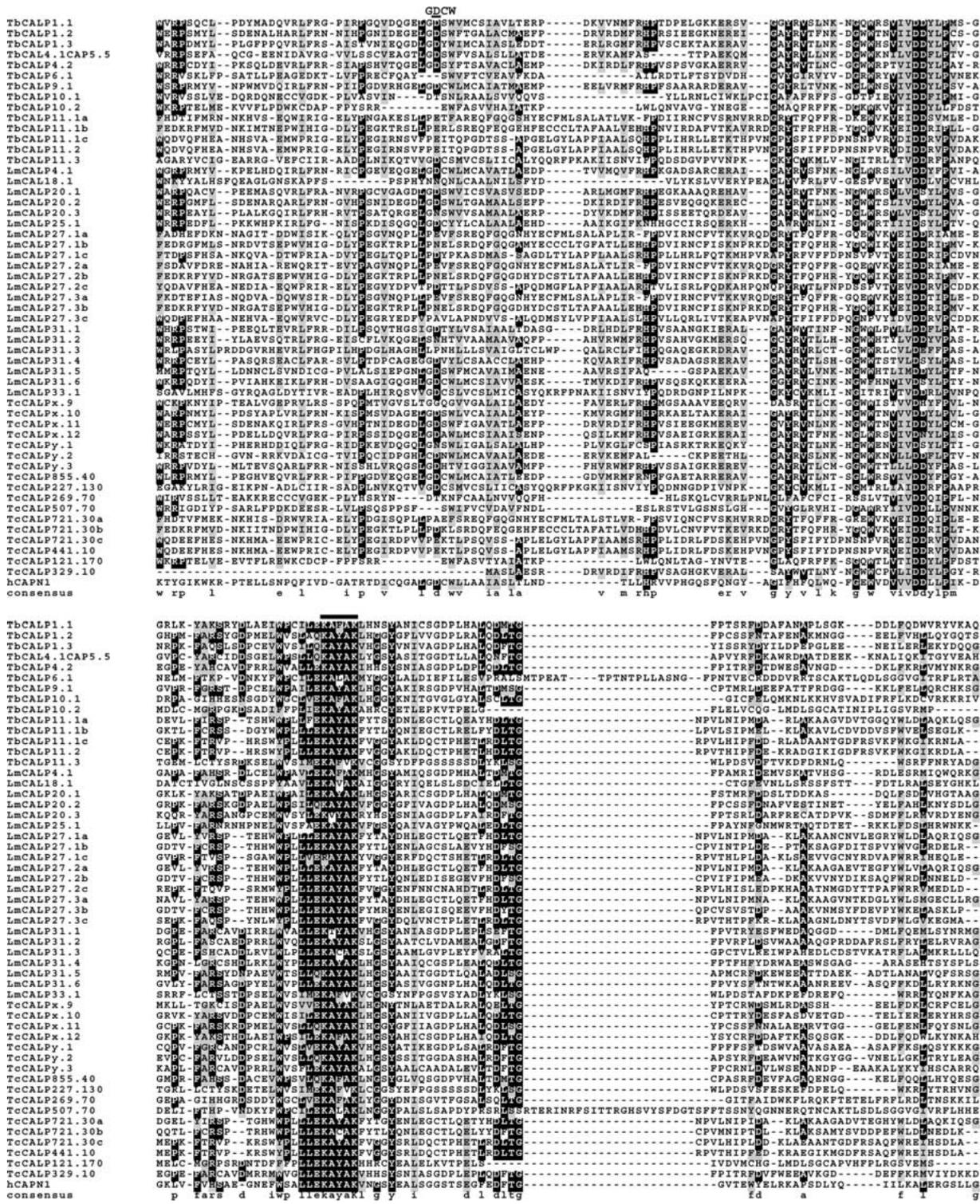


Fig. 1. Multiple alignment of the protease domain (domain II) of kinetoplastid calpain-like proteins. Identical residues are highlighted in black; similar residues, in gray. Sequence elements typical for calpains are given above the aligned sequences, with residues that are conserved in typical calpains underlined. The conserved KAYAK motif is indicated by a bar above the sequences. The

corresponding sequence of human CAPN1 (or μ -calpain; accession no. AAH08751) is included for comparison. Letters a, b, and c after some of the sequence names refer to the first, second, and third repeats of domain II occurring in five calpain-like proteins (Group 4 and 5 CALPs; see Fig. 2 for details).

utilization of the domain III transmitter/amplifier function in Ca^{2+} -independent regulatory mechanisms.

Domain IV
The C-terminal sequences of kinetoplastid calpain-like proteins did not show any similarities to the

```

                                     H                               RNPWG
TbCALP1.1  YQIILSTPGKNPRKASG-----ECHTARRYTSVGLTGHAVVLDAAFFPE-----YELRIVKRNANAGR--GSEQNDGSDGDEK--W
TbCALP1.2  YQIILVTPTRGALK-----LQGAS-----VEAYEAGLLPGHVVTVLSVMPPFE-----YNLRLLLRNANSHKRNANAKRGEK--W
TbCALP1.3  FKIIILSTHNAKTD-----NPNVDETYREVGLLAGHAAVAVKAVRYPAS-----ENIADLQIRNPNCR--ETEKKGGWANGNAE--W
TbCAL4.1CAP5.5  NPVILSVP-----TGRKAATAVAN-----GLREGYSVAMSVHNFPFE-----ENITLDMKRNPNPEP--AMPKGGWRGSDK--W
TbCALP4.2  YITVLTVPFPAALKAQAP-SDTSPTTEARKPRVTCKEGLQAHSGSGLVQVCKCFEE-----MGLPDLQVRRNPST--GEEKGGWRGSDK--W
TbCALP6.1  EFKQDEWASMLDMRSRGAIMVGTDPEN-----ALPGLTETAHARITHTVQYVGG-----IKHVMSSGSP--QQKGGWRGSDK--W
TbCALP9.1  YQIILTAGGSRVVTDKNEATTQSNRTDAKSLMDEGAPVGYATLQVRAVHGFTQA-----LDLQVTRVQVQK--CANRGGWRGSDK--W
TbCAL10.1  AFHANPEIS-----RAFRCLESLIPFPFGLFAVAEAPNAD-----SGYTCVVFVRSLLNDKPIEYLAAPHLVKGDIYE--W
TbCAL10.2  QEQQDVIWMEMKNNAKPNVLCSLLVVGRGDSNGASERREGILVDHIYVPLDARFLEG-----YRIVKRNHWGQPE-EINLCSKWRGSDK--W
TbCALP11.1a  QFVGSLLTR-----ETDLESGVQSQEQQGLVLEIFSLTG-----TSSVSDIVHILVHNPED-EEFTYKQPLNS-KDTE--W
TbCALP11.1b  DCACAAR-----PGFDDTGLSNGQEQAVLDVVVLTGDGE-----ASSISDLVVKLHNPFL--ETEYKQPMNA-SDAA--W
TbCALP11.1c  -AGDVITCI-----SNHESVQGLHSQCSALVDSVIETVHE-----SNDVSDIVIKLHNCY--DNPTYDGLPCN-DDDN--W
TbCALP11.2  -AGDVIMAM-----TNGSVPLGHSQCSALVDSVIETVHE-----SNDVSDIVIKLHNCY--DSPTYDGLPCN-DDDN--W
TbCALP11.3  SLIITVSTP-----PLSEAEERRELDASEHAAVLDMMREMSG-----VRRVQVRRNPVSR--RANSGRFLNDRKPE--W
LmCALP4.1  YQAILTTPGKAPA-IPAEAQSSPDFSDEMEAPAGTFLPGHAAVLDVKEFKQK-----GQIRICDRNPNVH--GSGTTRAAWDALE--W
LmCAL18.1  VQACIPIHTS-----TEGTFEFGVCCGGIHLAETDACCQKMG-----SHYARIEQVAPESG-RALKAFPIEMYTEVNV--W
LmCALP20.1  YTVICSTPGRGPHDK-----DQELCEYAVQVYVTCATLTERAKHTES-----KDLQVNRNPNVGR--SVEGSGRGNDDAK--W
LmCALP20.2  YQLVFTPTPRAIMTARG-----AISCGSPVQAEVYERSLLGHTYFVLRVGYFEE-----HDLRQLQIRNPNCR--SSSDNCPNKRDN--W
LmCALP20.3  HLVLSTPFGSG-----DPRASHTYKKEKLVGYAAVLEARAEASPTLRRKK-----VPPLRLRLRNPNAE--ATKSKGSDGSSM--W
LmCALP25.1  YLICATP--SNGALK--MAGKQLSANGLFEKAGRIGHAAVLDVRRHFL-----HRLCLLRNPNVGR--HVEGSGRGNDDAK--W
LmCALP27.1a  ERFVASALTK-----DVEVETGLQREQQGVLEIFSLHG-----TSALDDIVHILHNPED-DEFLSGLPLN-NDSA--W
LmCALP27.1b  TTVAAQC-----GDAEESGLAPHQYVAVDDELLATSYPP-----RTLS-EVIVQVRRNPNV--EIFFTQPMSPDADPR--W
LmCALP27.1c  -RGDVPVAV-----SGDLVPPGILPHQYVAVDVIETVPG-----SNDPSDVIKLNHNCYH--DAPYRQPLAM-GDSD--W
LmCALP27.2a  QFIASVLTG-----DIELEHTGLQREQQGVLEIFSLTG-----TSSINDIVHILHNPED-EYLYKQPLNS-KDMT--W
LmCALP27.2b  QTLAALAL-----GEABQGLHHGSAVDFIETRNAI-----NLTPADVVKHNPDE--DSPTPCMHR-DDSS--W
LmCALP27.2c  -RGDVPVAV-----ANGSVPLGHPQCSALDDVVEYKPC-----TSDPLDVIKLNHNCY--DAPHYKPLRK-GDSM--W
LmCALP27.3a  DHVLSALTT-----DADLESGLQREQQAVLEIFSLTG-----SSALDDIVHILHNPED-DEFLSGLPLN-KDSK--W
LmCALP27.3b  DLCTTALAS-----GREAEVQGLKERYAAVLDVVIANGE-----RVR-DMVVKHNPDE--GSPYTPMNA-KDDA--W
LmCALP27.3c  -RGDVPVAV-----SNAEVPNGHLEKRYAAVLDVVIANGE-----RVR-DMVVKHNPDE--GSPYTPMNA-KDDA--W
LmCALP31.1  HLOVLTCPSDGGEAFG-----NANVVSANPELEATYKGLLHNGVAVDQTYEFD-----LDLRLVLRNPNVGN--GSEOTGAWKNDKR--W
LmCALP31.2  CTVLVSPPPPTDEQSSAT-VDMSRKRSSRMPVKAQVLPFGHIFLRDAVAYEA-----LDLRLVLRNPNVGN--ETKNIKLYKWKYITW--W
LmCALP31.3  RTAEMPSVVCVQLPHFRG--TARTQHIGDRLVLYVEDTDPNIAIVVGLTGLDGGQ-----CVARRQTAQRKRS--AESLDDLRRRAGKS--W
LmCALP31.4  EHFPTSLSEKVVHQRYGLV--CLCLPDDGPTAQAAQGVVFGMSAVRVLVRHGS-----VRLDQICQPLK--LCSRQVWCAQSGAR--W
LmCALP31.5  ACVVLSTPC--HMSSEY-LGRSQASDAAFRARKYKGVHVGITLSERVVVE-----COTLKFVRRNPNV--SSKTPGYSYSGRE--W
LmCALP31.6  YLESISTPDTHTSYMN--AGSGMANAVLEARYKAAGLTHGMSVAKVQRQVVI-----PRVRLKLRNPNVGR--GDEMLCPWKNSTK--W
LmCALP33.1  ALFTVSTP-----VKMPAADAAEGLAAGHAAVDMQEQVK-----ERVMLLRNPNVGR--QESRQVRRRDTSEAS--W
TcCALPx.9  FIVVLDNRPLNTSG-----KSSLQRGLGEGNAVIVVSMKMPQLPNSNGVSGQLQQQIHRVVRNPNVGR--ARDVGDQNLWPPA--W
TcCALPx.10  YIISINTPVRDPPQ-----DEEKEKYKEGLMGMVYVVRKVRHFKS-----EISLQIRNPNVGR--DVEKGGWANDSK--W
TcCALPx.11  YQVLFVAPTRTLNRG-----AMDGVS-----ESTYARVGLRLHVGAVVKKLFFPE-----YNLRVQVRRNPNVGRDGAIVNGPWKKGDRK--W
TcCALPx.12  YQNMSTPGKDPNDKGS--DPLKLVKLSVGLSGLSGLHATVTDQTYPAD-----HELRLVLRNPNVGR--TABNGWVCGDDEK--W
TcCALPy.1  YLLETSTPVIIGAYTNI-YDMSGELDNGMKER-YRSGLELGLSGLVVDVRRFRF-----PDVLDKIRVQVCEC-GLKTPGAWSEASEK--W
TcCALPy.2  NIVLNFEP-----AGKSTKQNAKP-----GQITGLTCLKEVRFQFN-----FGITMLKIQNPEP--TEPTGAWCAGSD--W
TcCALPy.3  FLNVYTPSPKPSAPL-MNLICPSTRPVGSGEVSFPLPFGVETLETYED-----LDLRLVLRNPNVGR--SAAVLRFPWKRSMK--W
TcCALP25.40  FQIILTAGKAPAPISGTNKIASLPDEPELERILGGTFLPGHATVLDLRYPPKE-----GDQLQIRNPNVGR--CEBESGDSSESSE--W
TcCALP27.130  SVLITASTG-----TLSEEEKERKAAASHAAVDMIREVDG-----ERDQLKRNPNVGR--QSSGSKVSDVDRREKS--W
TcCALP269.70  SARLIPRNS-----MAMLSAKERLEPYMSVVAATSAHRDE-----KPHPSYVQVLSQPKLECETDFIYSYKMESVSGL--W
TcCALP507.70  QFQLDSEWLTMSDVHSAGAMMVAVSRSANG-----TLPIEFSHARVQLQQLSG-----ARIEIUSSTK--MQRKGGWSDSVQ--W
TcCALP721.30a  QFVAVSLTK-----DMETESGIGREQDTGVLEIFSMTG-----TSSVNDIVHILHNPED-EEFTYKQPLNS-KDSH--W
TcCALP721.30b  NCAYGATAR-----SGIGSNIGIQEDQTVLISVISTRNSV-----NPELSDIVVHNPVAV--EAVTGPMMN-EDIR--W
TcCALP721.30c  -KGDVITCM-----SNVDAGDGIHPLCSALFAVETVKE-----SNDPADIVIKLNHNCYF--DEFFYSGLPLN-KDGP--W
TcCALP441.10  -KGDITAM--SNKHVPPGDIHPLCSALFAVETVKE-----SNDPADIVIKLNHNCYF--DEFFYSGLPLN-KDGP--W
TcCALP412.170  AQGNLWKKLNAIQDVLPTFLLRGDSASERSLISDHLFAVRRFVGS-----QVYVGLVGRHWGQPE-EINLCSKWRGSDK--W
TcCALP329.10  PITMLVTPR-NDDSGGQP-AHGSINGAATVNERNCVREGRLHCGLTVLRVYFED-----FDLRLQIRNPNVGR--GEEPTGAWKNDKR--W
hCAPN1  CSDDISSVLD-----MEAITFKVLDLGHAAVLDGKQVNYRG-----QVVSIRMRNPNVGR--EVEPTGAWSDSSE--W
consensus  iil t                               vgl g y y v l                               lvklnpww          w g w

```

Fig. 1. Continued

EF-hand-containing domain IV of conventional calpains or to the so-called domain T and PBH domain that is the equivalent of domain IV in some atypical calpains in a number of organisms (Sorimachi et al. 1997). To distinguish the kinetoplastid sequences from other domain IV equivalents, we have termed it domain C (for C-terminus). Except for putative orthologous genes, no similarities were obvious among the three organisms, nor did the sequences contain any recognizable functional motifs.

Sequences with an Unusual Domain Composition

Most of the domain II-containing sequences described here have a length of approximately 700 amino acids. However, a few sequences are considerably longer. TbCALP11.1, LmCALP27.1, 2, and 3, and TcCALP721.30a are between 4500 and 6200 amino acids in length. This is due to the presence of three copies of the protease domain (domain II followed by domain III) and to the separation of the second and third protease domains by long arrays of near-perfect tandem repeats of 65–68 amino acid residues unit length. Very similar repeats are also found in two shorter sequences, TbCALP11.2 and TcCALP441.10 (Fig. 2). These two sequences lacked the first and second copies of the protease domains. The repeats of all seven sequences are homologues to each other (Suppl. 5). Secondary structure analysis predicted a high probability of the repetitive arrays to form a

coiled-coil structure. Database searches revealed that short fragments containing some of the repeats of TbCALP11.2, LmCALP27.2, and TcCALP441.10 had been published as antigens GM6 in *T. brucei*, Lcr1 in *Leishmania chagasi*, and FRA in *T. cruzi*, respectively (Lafaille et al. 1989; Muller et al. 1992; Wilson et al. 1995). It was shown in these studies that GM6 is associated with the cellular microtubule structures and FRA with the flagellar cytoskeleton.

Protein Modification by Fatty Acids

The N-termini of a significant number of both long and short sequences, outside the domain I^K consensus sequence, share the common feature of possessing an acylation modification motif. This motif, consisting of a subterminal glycine followed by a cysteine, is also present in CAP5.5, where dual myristoylation and palmitoylation has been experimentally demonstrated (Hertz-Fowler et al. 2001). The presence of glycine and cysteine is necessary but not sufficient for acylation to occur. The enzymes catalyzing the modification with fatty acids require an extended sequence motif that is usually contained within the first 10–15 N-terminal amino acids (Maurer-Stroh et al. 2002). Using in silico analysis of the N-terminal 20 amino acids of all sequences that possess the glycine-cysteine, we showed that acylation is highly

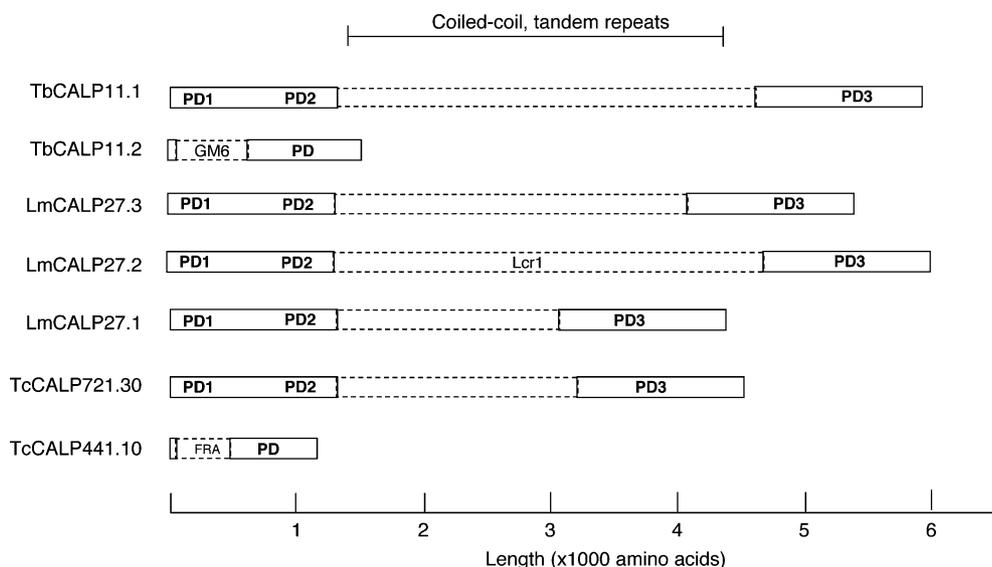


Fig. 2. Domain structure of calpain-like proteins with an internal 65- to 68-amino-acid repeat motif in kinetoplastids. TbCALP11.2 and TcCALP441.10 have previously been identified as cytoskeleton-associated proteins GM6 and FRA. Solid boxes indicate the

position of the protease domains (PD). Dashed boxes correspond to the position of the repeats. The locations of the previously published sequence fragments GM6, Lcr1, and FRA are indicated.

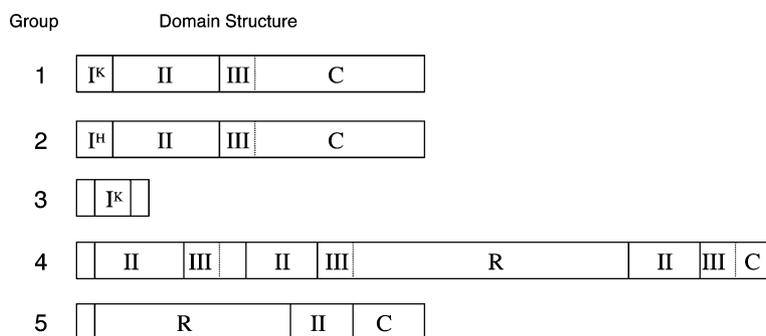


Fig. 3. Classification of calpain-like proteins in kinetoplastids according to their domain structures. I^K, kinetoplastid-specific domain I; I^H, heterogeneous domain I; R, repetitive sequence domains; C, C-terminal domain.

likely, with the exception of TcCALPy.1 (Table 1, column 4). Furthermore, the acylation motif was present only on calpain-like protein sequences that possessed the canonical kinetoplastid domain I^K sequence and not in domain I^H-containing sequences.

Classification of Kinetoplastid Calpain-like Proteins

According to their domain structure and sequence composition, we have categorized the calpain-like proteins identified in the three kinetoplastid species into five groups (Fig. 3). Group 1 includes sequences that resemble conventional calpains most closely. The proteins can be dissected into distinct domains. A kinetoplastid-specific domain I (I^K) is followed by a typical calpain domain II. It is followed in most sequences by a clearly recognizable domain III homologue. The C-terminal domain IV (domain C) does not contain EF-hand motifs. Except for orthologous sequences, domain C is not conserved between sequences.

Group 2 proteins are similar in domain structure to Group 1 sequences, but domain I (I^H) is heterogeneous and unrelated to domain I^K found in Groups 1 and 3.

Group 3 sequences contain a domain I^K that is highly similar to domain I of Group 1^K sequences. N- and C-terminal extensions are short and show no similarities to other sequences. One member of this group, TbSKCALP10.3, contains four degenerate, tandem repeats of domain I. Groups 1 and 3 sequences are also related by the presence of a dual or single N-terminal acylation motif on a number of sequences.

Group 4 proteins contain three repeats of domains II and III. The second and third domain copies are separated by varying numbers of tandem repeats of 70 amino acids unit length, most likely forming a coiled-coil structure.

Group 5 represents sequences with N-terminal repeats similar to Group 3 repeats, but only with single C-terminal calpain domains II and III.

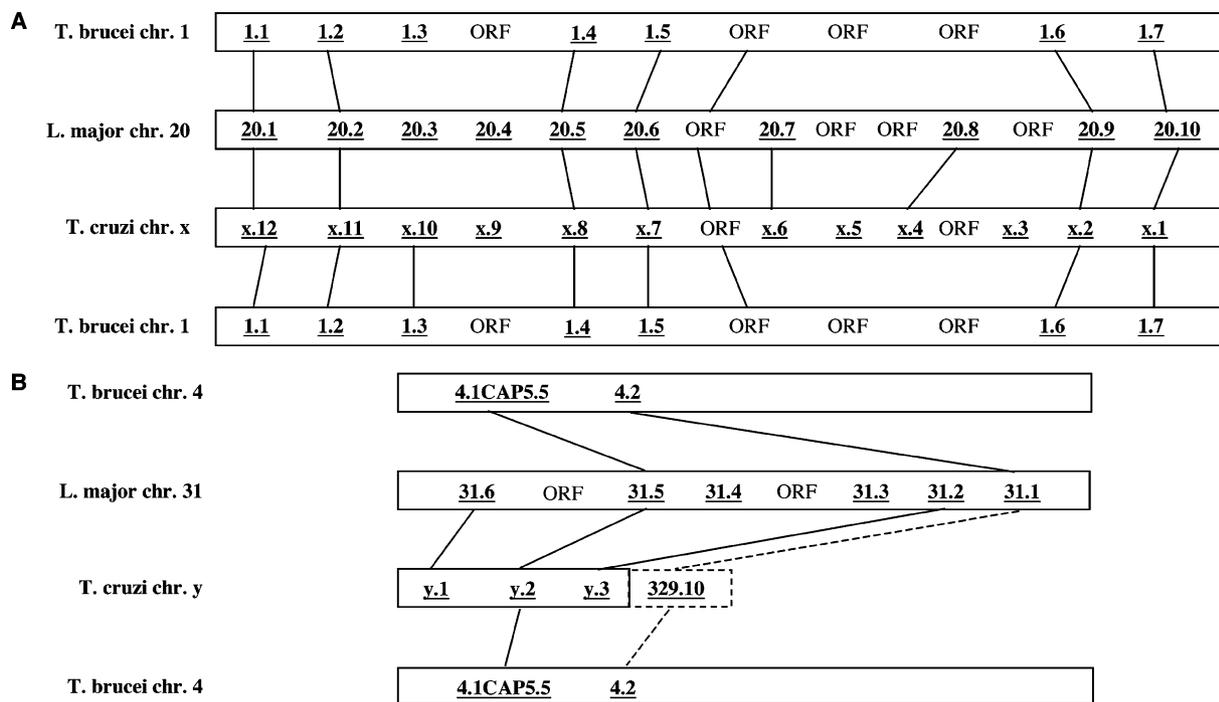


Fig. 4. Arrangement of two chromosomal clusters (**A**, **B**) containing genes for calpain-like proteins. Connection lines between genes on chromosomes of different species are based on the degree of similarity of whole-sequence cross-species alignments. The

hatched box as part of *T. cruzi* chromosome y indicates a putative assignment of sequence TcCALP329.10 to this position. Calpain-like genes are in boldface and underlined. ORF, open reading frames unrelated to calpain-like genes.

Genome Organization

A typical feature of gene organization in trypanosomes and *Leishmania* is the presence of gene clusters containing several copies of identical or paralogous genes as tandem repeats (Myler et al. 1999; El-Sayed et al. 2003; Hall et al. 2003; Worthey et al. 2003). Often, syntenic groups are conserved between different trypanosomatid species (Bringaud et al. 1998).

Many of the genes coding for the calpain-like proteins described in this study were found to be clustered on particular chromosomes (Fig. 4). The largest cluster of calpain-like proteins in *T. brucei* was localized on chromosome 1 and contains seven genes. The corresponding gene cluster in *L. major* was on chromosome 20, with 10 genes, and in *T. cruzi* on a single cluster that has not yet been assigned to a particular chromosome (labeled x), containing 12 genes (Fig. 4A). All three clusters contained both long, domain II-containing genes for calpain-like proteins in Group 1 and short, domain I^K-only genes in Group 3. Genes for both groups were not interspersed and open reading frames unrelated to calpain-like proteins were also present within all three gene clusters. Interspecies similarity between sequences of equivalent positions was often greater than intraspecies similarities of adjacent genes. For example, TbCALP1.4 is, across the entire sequence, more similar to LmCALP20.5 (44% identity) and TcCALPx.8

(48%) than to adjacent TbCALP1.5 (28%). However, whereas TbCALP1.4, TbCALP1.5, and TcCALPx.8 contained a dual N-terminal acylation motif, LmCALP20.5 contained only a single acylation motif.

A smaller cluster of related calpain-like sequences was localized on *T. brucei* chromosome 4, on *L. major* chromosome 31, and on an unassigned cluster in *T. cruzi* (labeled y) (Fig. 4B). All members of these clusters are domain II-containing calpain-like genes of Group 1. In addition, the genes coding for Groups 4 and 5-type calpain-related proteins are found adjacent to each other in all three species (Table 1).

Comparative Sequence Analysis

The abundance of calpain-like proteins in all three kinetoplastid species suggested complex patterns of gene evolution. Some genes will have evolved in a similar fashion in all three species with a high probability of the presence of orthologous pairs and triplets. Conversely, the identification of divergent sequences within species could indicate a species-specific function.

To address questions concerning the evolutionary ontogeny of the large group of calpain-like proteins in kinetoplastids a phylogenetic analysis was done comparing domain II sequences (Fig. 5). These were the only sequences that were conserved between all calpain-like proteins (CALPs), except for domain I-only sequences (SKCRP, Group 3), and could also be

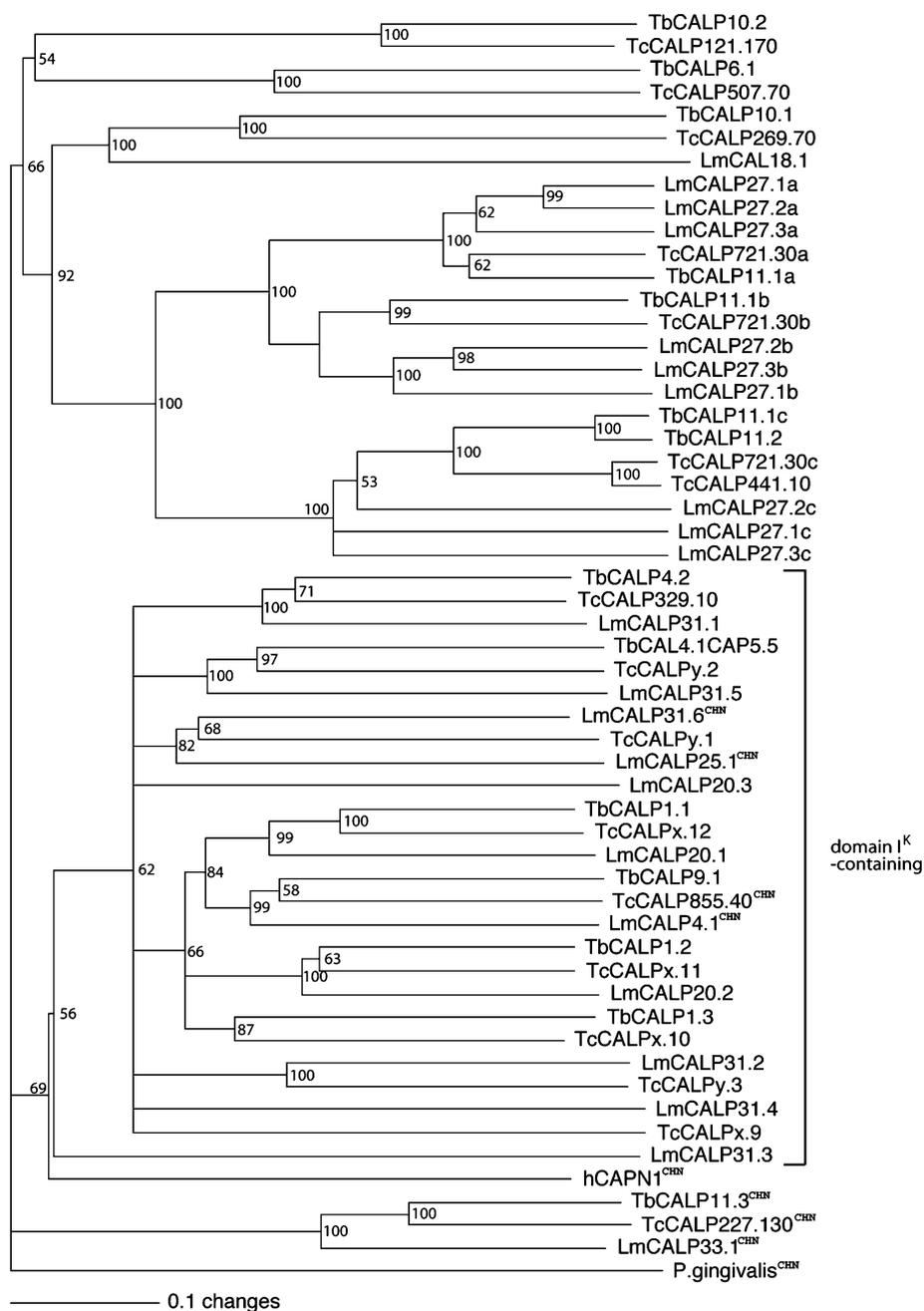


Fig. 5. Phylogenetic tree representation of the protease domain (domain II) sequences. Protein sequences were aligned with ClustalW and manually edited, and a neighbor-joining tree was constructed using PAUP*4.0. Bootstrap support values of 1000 trials are indicated. The tree is rooted with the sequence of a calpain-related protein in the prokaryote *Porphyromonas gingivalis* (accession no. P25806). The protease domain of human CAPN1 (accession no. AAH08751) is also included. Sequences that contain the kinetoplastid-specific domain I^K are indicated. All other se-

quences contain a heterogeneous N-terminal domain (I^H; see text). It should be noted that neither domain I^K nor I^H was used in the alignment to create the tree. Letters a, b, and c after some of the sequence names refer to the first, second, and third repeats of domain II occurring in five calpain-like proteins (Groups 4 and 5 CALPs; see Fig. 2 for details). Sequences containing the conserved catalytic triad residues cysteine, histidine, and asparagine are labeled with a superscript index (^{CHN}).

compared and rooted against outgroup proteins such as the evolutionarily distantly related calpain-like cysteine protease Tpr from the prokaryote *Porphyromonas gingivalis* (Bourgeau et al. 1992).

An interesting relationship was revealed for the multiple copies of domain II found in Group 4 calpain-like proteins (Fig. 2). A given domain II was

more closely related to the equivalent domain in the same position in a different protein of this group from the same or another species than to one of the other two copies of domain II within the same protein (Fig. 5).

A clear segregation was observed between calpain-like proteins that contained a kinetoplastid-specific

domain I^K and those that contained the unrelated domain I^H. The N-terminus of the latter group was, as described above, heterogeneous and contained little similarities between nonorthologous sequences.

Discussion

Characterization of Novel Members of the Calpain Superfamily

With the availability of the completed genomes of the kinetoplastid parasites *Trypanosoma brucei*, *T. cruzi*, and *Leishmania major*, it is now possible to compare and analyze proteins at a genomewide level. Although the three species are related, their parasitic lifestyles are very different from each other and comparative genome analysis helps to identify proteins that have parasite-specific functions and contribute to the understanding of distinct modes of pathogenicity (Cox et al. 1998; Beverley 2003). Furthermore, kinetoplastids are organisms that have diverged early in eukaryotic evolution, and studying protein families that are also found in higher eukaryotes contributes to the understanding of their evolutionary dynamics (de Meeus and Renaud 2002).

Based on our earlier work on the identification of a calpain-like protein in *T. brucei* (Hertz-Fowler et al. 2001), we have identified a large and diverse group of proteins in all three parasites that are members of the calpain superfamily. The majority of the sequences were identified due to the presence of a domain that has significant homology with the catalytic domain II of calpain cysteine proteases. The absence of amino acid residues critical for catalytic activity in most of the sequences makes it unlikely that the proteins function as cysteine proteases, although the absence or nonequivalence of catalytic residues is not always an indication of catalytic inactivity (for an extensive discussion of functional protein evolution see Todd et al. 2001). In typical calpains, the catalytic triad is composed of the amino acids cysteine (C), histidine (H), and asparagine (N). Deviations from the classical C,H,N catalytic triad pattern is observed not only in kinetoplastids, but also in a number of calpains of higher eukaryotes. For example, in the human calpain CAPN6 the triad is changed to K,Y,N (Dear et al. 1997).

Recent systematic surveys have investigated the relationship between a range of enzymes and their nonenzyme homologues (Todd et al. 2002; Pils and Schultz 2004). The presence of inactive enzyme homologues is widespread in many enzyme families and they have often acquired functions in regulatory networks, co-evolving with increasing cellular and organismal complexity (Bartlett et al. 2003). A mechanistic explanation could be the shift from

substrate catalysis to substrate binding only (Dev-edjiev et al. 1997; Lamb et al. 2000). It has been argued that nonenzyme proteins are derived from catalytic precursors, because the majority of members of a particular enzyme family are active (Todd et al. 2002). In kinetoplastids, however, the number of calpain-like proteins with a nonstandard catalytic domain far outnumbers the few proteins with the classical C,H,N triad. On the other hand, the few C,H,N-containing proteins are present in the two major branches of the phylogenetic tree (Fig. 5), consistent with the possibility that the other proteins are derived from an active precursor (assuming that the absence of a Ca²⁺-binding domain IV in all of the kinetoplastid proteins is, in evolutionary terms, not essential for activity). In most cases where the amino acids of the typical catalytic triad have not been preserved, the substituting amino acid, in both kinetoplastids (this study) and other organisms (for alignments see Sorimachi and Suzuki 2001), requires only a single mutation in their respective codons and could therefore be the consequence of a statistical rather than a functional preference.

Typical calpains are calcium dependent in their activity. Two domains that are able to bind calcium ions have been identified in these proteins. The C-terminus of calpains (domain V) contains five EF-hand motifs that are able to bind calcium and a further two non-EF-hand binding sites were identified by protein crystallography within the proteolytic domain II (Hosfield et al. 1999a; Strobl et al. 2000). None of the kinetoplastid calpain-like proteins possesses a typical domain IV with EF-hand motifs. This finding is not surprising, as it has been shown that calpains with EF-hand motifs are found only within the animal kingdom. It has been hypothesized that this functional module was added to an ancestor protease by gene fusion with a calmodulin-like gene (Ohno et al. 1984; Emori et al. 1986). Also essential for function is the binding of two additional calcium ions at a site located within the protease domain II (Moldoveanu et al. 2002, 2004). Five amino acid residues have been identified in mammalian CAPN1 (or μ -calpain) that are required for calcium coordination (Moldoveanu et al. 2004). In kinetoplastids, these amino acids are found only in LmCALP25.1. LmCALP25.1 is also one of the few sequences that show an intact C,H,N-catalytic triad. This suggests that this protein has the highest probability of all sequences analyzed in this study of possessing a cysteine protease activity and binding calcium.

In addition to sequences that can easily be identified by their similarity to domain II of conventional calpains, we describe a novel class of calpain-like proteins. These short sequences of between 120 and 160 amino acids have no common features with typical calpains or calpain-like proteins. However, the core 100-amino-

acid residues are very similar to the sequence domain that constitutes the N-terminal domain I^K of the majority of the domain II-containing kinetoplastid calpain-like proteins. A phylogenetic analysis, based only on a comparison of domain I^K-sequences, is unable to resolve Group 1 (domain II-containing) and Group 3 (domain II absent) sequences into two distinct branches (data not shown). This sequence element is not found in any other context in kinetoplastids and is entirely absent from the genomes of other organisms that are included in BLAST-searchable translated DNA or protein databases.

The second group of N-terminal sequences, termed domain I^H, is entirely different from domain I^K. The sequences are dissimilar when compared to each other and have no similarities to other sequences in the databases, with the exception of TbCALP6.1 and TcCALP507.70. The N-terminus of these two calpain-like proteins shows a significant similarity to the N-terminus of the regulatory subunit of two cAMP-dependent kinases that have been identified in *T. brucei* and *T. cruzi*. In kinases, this domain is not involved in cAMP-binding but is essential for kinase dimerization and for interaction with A-kinase anchoring proteins (AKAPs), which recruit kinases to intracellular target sites such as the cytoskeleton or membranes (Newlon et al. 1999; Diviani and Scott 2001; Newlon et al. 2001).

Another common feature of many of the short and long, domain II-containing calpain-like proteins is the presence of a single or dual N-terminal acylation motif. This motif is found only in sequences containing domain I^K (Groups 1 and 3) and is excluded from all other calpain-like proteins. In TbCALP4.1CAP5.5, a protein localized to the microtubule cytoskeleton of the cell body, modification of the protein by the addition of myristate and palmitate has been experimentally verified (Hertz-Fowler et al. 2001). Recently, it has also been shown that the domain I-only sequence SMP-1 (identical to LmCALP20.10) is also dually acylated and localized to the flagellar membrane in *Leishmania* (Tull et al. 2004). Ablation of acylation by mutating the gene abolishes exclusive flagellar localization and most of the protein remains in the cytosol. Since both TbCALP4.1CAP5.5 and LmCALP20.10 are dually acylated, but are targeted to different cellular locations, acylation by itself cannot be the only factor that determines protein targeting. Acylation is most likely involved in relatively unspecific association with cellular membranes (Resh 1999), but the retention in the cell body (TbCALP4.1CAP5.5) or the selective transport to the flagellum (LmCALP20.10/SMP-1) will be regulated by additional motifs within the sequences. It will be informative to see whether acylated proteins of the calpain family in kinetoplastids are associated with membranes other than the flagellar and cell body membrane. Also, the functional differences between

proteins containing single myristoylation and dual myristoylation/palmitoylation need to be established. In other systems, it has been shown that dual acylation is necessary and sufficient for membrane targeting, whereas myristoylation on its own is necessary but not sufficient, and additional signals are required for membrane binding (Boutin 1997; Resh 1999).

Calpain-like proteins without acylation motifs are not necessarily excluded from membrane association. One of the major targets of conventional mammalian calpains is proteins of the cytoskeleton, and they are involved in modulation of cytoskeletal dynamics at the cytoskeleton-membrane interphase (Goll et al. 2003). However, despite their localization in association with membranes, none of the conventional calpains is acylated. It is interesting to note, however, that some of the established in vivo substrates of mammalian calpains, the cytoskeletal proteins vinculin, spectrin, ankyrin, and band 4.1, are themselves acylated and membrane-associated (Staufenbiel and Lazarides 1986; Burn and Burger 1987; Maretzki et al. 1990; Mariani et al. 1993; Bhatt et al. 2002).

An additional group of calpain-like proteins with its own distinctive feature are seven sequences that are characterized by the presence of internal repeats of 65–68 amino acids that are homologous between all seven sequences. Two of these sequences, TbCALP11.2 and TcCALP441.10, have previously been characterized as cytoskeleton-associated proteins GM6 and FRA, respectively (Lafaille et al. 1989; Muller et al. 1992). At the time of these publications, only fragments of the repeat were identified by cDNA sequencing and their calpain-like properties were not recognized. Both sequences have short, but similar N-termini, followed by 10 copies of the repeat and a C-terminal calpain domain II. Despite their overall similarity in sequence and domain structure, the localization of both proteins is different. GM6 is located on fibers which connect the microtubules of the membrane skeleton with the flagellum, whereas FRA (flagellar repetitive antigen) is localized inside the flagellum (Lafaille et al. 1989; Muller et al. 1992). The presence of repeats in the other five calpain-like proteins of this group, but also the observation that a number of cytoskeleton-associated proteins in trypanosomatids have an internal repeat structure (Gull 1999), suggests that these repeats mediate interaction with microtubules and that the presence of calpain domain II adds an additional element of functionality to these proteins.

Genomic Organization of a New Gene Family

The presence of 18 new calpain-like proteins in *T. brucei*, 24 in *T. cruzi*, and 27 in *L. major* constitutes the largest number of calpain/calpain-like proteins in

any organism where such proteins or their genes have been found. Repetitive gene families in kinetoplastids are not unusual and are a characteristic feature of their genome organization (Myler et al. 1999; El-Sayed et al. 2003; Hall et al. 2003; Worthey et al. 2003). However, in most cases identical or near-identical copies with presumably redundant functionality form the basis of the presence of multicopy gene clusters. In the case of calpain-like proteins, there is, despite their large number, no evidence of identical copies of a particular gene. Even the genes organized in large gene clusters on chromosome 1 in *T. brucei*, chromosome 20 in *L. major*, and chromosome x in *T. cruzi* code for proteins that are sufficiently divergent to make distinct functions likely. This is also emphasized by the observation that TbCALP4.1CAP5.5 is differentially expressed only in the insect stage of the life cycle (Matthews and Gull 1994), whereas the immediately adjacent gene coding for TbCALP4.2 is expressed in both bloodstream and insect stages of the life cycle of *T. brucei* (Hertz-Fowler et al. 2001). Also, domain I^K-only sequences (Group 3) are found on the same clusters as domain II-containing genes (Group 1). Both types of genes are, however, never mixed within a cluster but are always segregated into two adjoining subclusters.

Possible Functions of Calpain-like Proteins in Kinetoplastids

As yet there are no data on the specific functions of any of the calpain-like proteins in the three species. However, a few common themes emerge from a number of observations. First, the few calpain-like proteins characterized so far at a cellular level are associated with the cytoskeleton or membranes as indicated by biochemical studies and immunolocalizations (Lafaille et al. 1989; Muller et al. 1992; Hertz-Fowler et al. 2001; Tull et al. 2004). Second, a number of calpain-like proteins are differentially expressed during the life cycle of the parasites. TbCALP4.1CAP5.5 is expressed in the procyclic insect stage, but expression already commences during the differentiation process from bloodstream forms. Depletion of the protein in procyclic cells by RNA interference is lethal (K. Gull and S. Vaughan, unpublished). Array-based transcriptome analysis data showed that TbCALP1.1 is upregulated in the bloodstream form of *T. brucei* (El-Sayed et al. 2000). A similar analysis in *L. major* revealed that LmCALP20.2 is upregulated in the promastigote insect stage and LmCALP20.1, coded by the adjacent gene, is upregulated in the subsequent metacyclic insect stage (Saxena et al. 2003). The flagellar calpain-like protein LmCALP20.10/SMP-1 is detectable only in promastigote stages of *Leishmania* containing a well-developed flagellum, and not in amastigotes, which contain only a highly truncated flagellum (Tull et al. 2004). Third, the presence of acylation motifs in many

of the proteins, the internal repetitive structure of some of the proteins, and the presence of a cAMP-kinase anchoring protein (AKAP) interaction motif on two proteins indicate that these proteins are associated with membranes or the cytoskeleton. Also, domain III in conventional calpains is structurally similar to C₂ domain, a lipid-binding domain found in a number of membrane-associated proteins (Rizo and Sudhof 1998; Hosfield et al. 1999a). The predicted structural similarity of domain III of kinetoplastid calpain-like proteins with a conventional calpain domain III indicates a role in protein-membrane interaction.

Taken together, compartmentalization to membranes or cytoskeleton, either in the cell body or in the flagellum, and, at least for some proteins, life cycle-specific expression may demarcate the search for specific functions. Given the elaborate nature of the kinetoplastid cytoskeleton, including the flagellum, and its role in life cycle differentiation, motility, and cell division, it is possible that the establishment of a large family of calpain-like proteins was an evolutionary response to use these structures as scaffolds for diverse cellular functions.

Evolution of Calpain-like Proteins in Kinetoplastids

The presence of a large number of calpain-like proteins in all three trypanosomatids analyzed in this study strongly suggests that this gene family has evolved by multiple events of gene duplication. Interspecies comparison showed that many of the sequences are closely related and are likely to present orthologues. Related sequences are also often found on calpain gene clusters with a similar interspecies organization. These observations indicate that on the whole the development of the family calpain-like protein coding genes was completed prior to speciation into *T. brucei*, *T. cruzi*, and *L. major*. The larger number of genes in the latter two kinetoplastids suggests that additional gene duplications have occurred after divergence from *T. brucei*. Phylogenetic analysis of the protease domain II separated the calpain-like proteins into two clusters, one containing domain I^K and the second containing domain I^H sequences. This clear segregation indicates that the acquisition of domains I^K and I^H by a domain II-containing gene occurred after the initial duplication of a domain II gene. These duplicated genes then acquired independently various domain I gene modules.

The presence of calpain-like proteins with three copies of domain II has to be a relatively late gene segment duplication in an ancestral species because the first, second, and third domain IIs of each of these proteins are more closely related to the corresponding domains of similar proteins in the same or one of the other species than they are to each other.

Similar events of calpain gene duplications and subsequent functional diversification have occurred

independently in other species (Jekely and Friedrich 1999). The largest numbers are found in humans and *C. elegans*, with 15 and 13 different genes, respectively (Goll et al. 2003). Other organisms have far fewer genes coding for calpain-like proteins. With few exceptions, most organisms outside the animal kingdom have only a single calpain gene (Sorimachi and Suzuki 2001; Goll et al. 2003). In the protozoan parasite *Plasmodium falciparum*, only a single calpain-like protein gene was identified (Wu et al. 2003). Other cysteine proteases are abundant in lower eukaryotes and parasitic protozoa (Sajid and McKerrow 2002; Mottram et al. 2003). The existence of such a large number of calpain-like proteins in kinetoplastids is therefore very unusual for lower eukaryotes and indicates trypanosomatid-specific functions of this gene family. Common with other calpain-like proteins identified in lower eukaryotes is, however, the absence of the Ca^{2+} -binding EF-hand motif of a conventional domain IV. The acquisition of this domain is considered a late event in calpain evolution that has occurred during animal evolution but not in other eukaryotic branches (Ohno et al. 1984).

The diversities and similarities in domain structure of calpains and calpain-like proteins in kinetoplastids and other organisms have reinforced the concept of the modular, multifunctional nature of this calpain superfamily. It will be interesting to study the significance of the divergence of the catalytic domain, address the question whether any of these proteins have proteolytic or a different enzymatic activity, and identify the relevant protein interaction networks. Also, the presence of the kinetoplastid-unique domain I^K indicates a kinetoplastid-specific function.

Supplementary Data

A list of complete sequences, alignments, and analysis of subsets of sequences discussed in this paper are available as supplementary data.

Acknowledgments. We thank David Lunt (Hull University) for constructive discussions concerning phylogenetic analysis and Robin Allaby (Astra Zeneca) for help during the initial phase of this project. This work was funded by a Wellcome Trust Programme Grant to K.G. and a University of Hull Start-Up Grant to K.E. Kinetoplastid genomic data were accessed via <http://www.genedb.org/>. Most of the genomic data were provided by the Wellcome Trust Sanger Institute and The Institute for Genomic Research (TIGR), which are supported by the Wellcome Trust and the National Institutes of Health, USA (Grant AI043062), respectively.

References

Barnes TM, Hodgkin J (1996) The tra-3 sex determination gene of *Caenorhabditis elegans* encodes a member of the calpain regulatory protease family. *EMBO J* 15:4477–4484

Bartlett GJ, Borkakoti N, Thornton JM (2003) Catalysing new reactions during evolution: economy of residues and mechanism. *J Mol Biol* 331:829–860

Berti PJ, Storer AC (1995) Alignment/phylogeny of the papain superfamily of cysteine proteases. *J Mol Biol* 246:273–283

Beverley SM (2003) Protozoomics: trypanosomatid parasite genetics comes of age. *Nat Rev Genet* 4:11–19

Bhatt A, Kaverina I, Otey C, Huttenlocher A (2002) Regulation of focal complex composition and disassembly by the calcium-dependent protease calpain. *J Cell Sci* 115:3415–3425

Bourgeau G, Lapointe H, Peloquin P, Mayrand D (1992) Cloning, expression, and sequencing of a protease gene (tpr) from *Porphyromonas gingivalis* W83 in *Escherichia coli*. *Infect Immun* 60:3186–3192

Boutin JA (1997) Myristoylation. *Cell Signal* 9:15–35

Bringaud F, Vedrenne C, Cuvillier A, Parzy D, Baltz D, Tetaud E, Pays E, Venegas J, Merlin G, Baltz T (1998) Conserved organization of genes in trypanosomatids. *Mol Biochem Parasitol* 94:249–264

Burn P, Burger MM (1987) The cytoskeletal protein vinculin contains transformation-sensitive, covalently bound lipid. *Science* 235:476–479

Cox FEG, Kreier JP, Wakelin D (1998) Parasitology. In: Collier L, Balows A, Sussman M (eds) *Topley & Wilson's microbiology and microbial infections*. Arnold, London

de Meeus T, Renaud F (2002) Parasites within the new phylogeny of eukaryotes. *Trends Parasitol* 18:247–251

Dear N, Matena K, Vingron M, Boehm T (1997) A new subfamily of vertebrate calpains lacking a calmodulin-like domain: implications for calpain regulation and evolution. *Genomics* 45:175–184

Dear TN, Boehm T (1999) Diverse mRNA expression patterns of the mouse calpain genes Capn5, Capn6 and Capn11 during development. *Mech Dev* 89:201–209

Denison SH, Orejas M, Arst HN Jr (1995) Signaling of ambient pH in *Aspergillus* involves a cysteine protease. *J Biol Chem* 270:28519–28522

Devedjiev Y, Popov A, Atanasov B, Bartunik HD (1997) X-ray structure at 1.76 Å resolution of a polypeptide phospholipase A2 inhibitor. *J Mol Biol* 266:160–172

Diviani D, Scott JD (2001) AKAP signaling complexes at the cytoskeleton. *J Cell Sci* 114:1431–1437

El-Sayed NM, Hegde P, Quackenbush J, Melville SE, Donelson JE (2000) The African trypanosome genome. *Int J Parasitol* 30:329–345

El-Sayed NM, Ghedin E, Song J, MacLeod A, Bringaud F, Larkin C, Wanless D, Peterson J, Hou L, Taylor S, Tweedie A, Biteau N, Khalak HG, Lin X, Mason T, Hannick L, Caler E, Blandin G, Bartholomeu D, Simpson AJ, Kaul S, Zhao H, Pai G, Van Aken S, Utterback T, Haas B, Koo HL, Umayam L, Suh B, Gerrard C, Leech V, Qi R, Zhou S, Schwartz D, Feldblyum T, Salzberg S, Tait A, Turner CM, Ullu E, White O, Melville S, Adams MD, Fraser CM, Donelson JE (2003) The sequence and analysis of *Trypanosoma brucei* chromosome II. *Nucleic Acids Res* 31:4856–4863

Emori Y, Ohno S, Tobita M, Suzuki K (1986) Gene structure of calcium-dependent protease retains the ancestral organization of the calcium-binding protein gene. *FEBS Lett* 194:249–252

Goll DE, Thompson VF, Li H, Wei W, Cong J (2003) The calpain system. *Physiol Rev* 83:731–801

Gull K (1999) The cytoskeleton of trypanosomatid parasites. *Annu Rev Microbiol* 53:629–655

Hall N, Berriman M, Lennard NJ, Harris BR, Hertz-Fowler C, Bart-Delabesse EN, Gerrard CS, Atkin RJ, Barron AJ, Bowman S, Bray-Allen SP, Bringaud F, Clark LN, Corton CH, Cronin A, Davies R, Doggett J, Fraser A, Gruter E, Hall S, Harper AD, Kay MP, Leech V, Mayes R, Price C, Quail MA, Rabinowitz E, Reitter C, Rutherford K, Sasse J, Sharp S, Shownkeen R, MacLeod A, Taylor S, Tweedie A, Turner CM, Tait A, Gull K, Barrell B, Melville SE (2003) The DNA sequence of chromosome I of an African trypanosome: gene

- content, chromosome organisation, recombination and polymorphism. *Nucleic Acids Res* 31:4864–4873
- Hertz-Fowler C, Ersfeld K, Gull K (2001) CAP5.5, a life-cycle-regulated, cytoskeleton-associated protein is a member of a novel family of calpain-related proteins in *Trypanosoma brucei*. *Mol Biochem Parasitol* 116:25–34
- Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PE, del Bosque-Plata L, Oda Y, Yoshiuchi I, Colilla S, Polonsky KS, Wei S, Concannon P, Iwasaki N, Schulze J, Baier LJ, Bogardus C, Groop L, Boerwinkle E, Hanis CL, Bell GI (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat Genet* 26:163–175
- Hosfield CM, Elce JS, Davies PL, Jia Z (1999a) Crystal structure of calpain reveals the structural basis for Ca(2+)-dependent protease activity and a novel mode of enzyme activation. *EMBO J* 18:6880–6889
- Hosfield CM, Ye Q, Arthur JS, Hegadorn C, Croall DE, Elce JS, Jia Z (1999b) Crystallization and X-ray crystallographic analysis of m-calpain, a Ca²⁺-dependent protease. *Acta Crystallogr D Biol Crystallogr* 55(8):1484–1486
- Huang Y, Wang KK (2001) The calpain family and human disease. *Trends Mol Med* 7:355–362
- Jekely G, Friedrich P (1999) The evolution of the calpain family as reflected in paralogous chromosome regions. *J Mol Evol* 49:272–281
- Lafaille JJ, Linss J, Krieger MA, Souto-Padron T, de Souza W, Goldenberg S (1989) Structure and expression of two *Trypanosoma cruzi* genes encoding antigenic proteins bearing repetitive epitopes. *Mol Biochem Parasitol* 35:127–136
- Lamb AL, Torres AS, O'Halloran TV, Rosenzweig AC (2000) Heterodimer formation between superoxide dismutase and its copper chaperone. *Biochemistry* 39:14720–14727
- Maretzki D, Mariani M, Lutz HU (1990) Fatty acid acylation of membrane skeletal proteins in human erythrocytes. *FEBS Lett* 259:305–310
- Margis R, Margis-Pinheiro M (2003) Phytocalpains: orthologous calcium-dependent cysteine proteinases. *Trends Plant Sci* 8:58–62
- Mariani M, Maretzki D, Lutz HU (1993) A tightly membrane-associated subpopulation of spectrin is 3H-palmitoylated. *J Biol Chem* 268:12996–13001
- Matthews KR, Gull K (1994) Evidence for an interplay between cell cycle progression and the initiation of differentiation between life cycle forms of African trypanosomes. *J Cell Biol* 125:1147–1156
- Maurer-Stroh S, Eisenhaber B, Eisenhaber F (2002) N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. *J Mol Biol* 317:541–557
- Moldoveanu T, Hosfield CM, Lim D, Elce JS, Jia Z, Davies PL (2002) A Ca(2+) switch aligns the active site of calpain. *Cell* 108:649–660
- Moldoveanu T, Jia Z, Davies PL (2004) Calpain activation by cooperative Ca²⁺ binding at two non-EF-hand sites. *J Biol Chem* 279:6106–6114
- Mottram JC, Helms MJ, Coombs GH, Sajid M (2003) Clan CD cysteine peptidases of parasitic protozoa. *Trends Parasitol* 19:182–187
- Muller N, Hemphill A, Imboden M, Duvallat G, Dwinger RH, Seebeck T (1992) Identification and characterization of two repetitive non-variable antigens from African trypanosomes which are recognized early during infection. *Parasitology* 104(1):111–120
- Myler PJ, Audleman L, deVos T, Hixson G, Kiser P, Lemley C, Magness C, Rickel E, Sisk E, Sunkin S, Swartzell S, Westlake T, Bastien P, Fu G, Ivens A, Stuart K (1999) *Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proc Natl Acad Sci USA* 96:2902–2906
- Newlon MG, Roy M, Morikis D, Hausken ZE, Coghlan V, Scott JD, Jennings PA (1999) The molecular basis for protein kinase A anchoring revealed by solution NMR. *Nat Struct Biol* 6:222–227
- Newlon MG, Roy M, Morikis D, Carr DW, Westphal R, Scott JD, Jennings PA (2001) A novel mechanism of PKA anchoring revealed by solution structures of anchoring complexes. *EMBO J* 20:1651–1662
- Ohno S, Emori Y, Imajoh S, Kawasaki H, Kisaragi M, Suzuki K (1984) Evolutionary origin of a calcium-dependent protease by fusion of genes for a thiol protease and a calcium-binding protein? *Nature* 312:566–570
- Ono Y, Sorimachi H, Suzuki K (1998) Structure and physiology of calpain, an enigmatic protease. *Biochem Biophys Res Commun* 245:289–294
- Pils B, Schultz J (2004) Inactive enzyme-homologues find new function in regulatory processes. *J Mol Biol* 340:399–404
- Resh MD (1999) Fatty acylation of proteins: new insights into membrane targeting of myristoylated and palmitoylated proteins. *Biochim Biophys Acta* 1451:1–16
- Reverter D, Strobl S, Fernandez-Catalan C, Sorimachi H, Suzuki K, Bode W (2001) Structural basis for possible calcium-induced activation mechanisms of calpains. *Biol Chem* 382:753–766
- Richard I, Broux O, Allamand V, et al. (1995) Mutations in the proteolytic enzyme calpain 3 cause limb-girdle muscular dystrophy type 2A. *Cell* 81:27–40
- Rizo J, Sudhof TC (1998) C2-domains, structure and function of a universal Ca²⁺-binding domain. *J Biol Chem* 273:15879–15882
- Sajid M, McKerrow JH (2002) Cysteine proteases of parasitic organisms. *Mol Biochem Parasitol* 120:1–21
- Sato K, Kawashima S (2001) Calpain function in the modulation of signal transduction molecules. *Biol Chem* 382:743–751
- Saxena A, Worthey EA, Yan S, Leland A, Stuart KD, Myler PJ (2003) Evaluation of differential gene expression in *Leishmania major* Friedlin procyclics and metacyclics using DNA microarray analysis. *Mol Biochem Parasitol* 129:103–114
- Sorimachi H, Suzuki K (2001) The structure of calpain. *J Biochem (Tokyo)* 129:653–664
- Sorimachi H, Ishiura S, Suzuki K (1997) Structure and physiological function of calpains. *Biochem J* 328(3):721–732
- Staufenbiel M, Lazarides E (1986) Ankyrin is fatty acid acylated in erythrocytes. *Proc Natl Acad Sci USA* 83:318–322
- Stevens JR, Noyes HA, Dover GA, Gibson WC (1999) The ancient and divergent origins of the human pathogenic trypanosomes, *Trypanosoma brucei* and *T. cruzi*. *Parasitology* 118(1):107–116
- Strobl S, Fernandez-Catalan C, Braun M, Huber R, Masumoto H, Nakagawa K, Irie A, Sorimachi H, Bourenkow G, Bartunik H, Suzuki K, Bode W (2000) The crystal structure of calcium-free human m-calpain suggests an electrostatic switch mechanism for activation by calcium. *Proc Natl Acad Sci USA* 97:588–592
- Taveau M, Bourg N, Sillon G, Roudaut C, Bartoli M, Richard I (2003) Calpain 3 is activated through autolysis within the active site and lyses sarcomeric and sarcolemmal components. *Mol Cell Biol* 23:9127–9135
- Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* 307:1113–1143
- Todd AE, Orengo CA, Thornton JM (2002) Sequence and structural differences between enzyme and nonenzyme homologs. *Structure (Cambr)* 10:1435–1451
- Tull D, Vince JE, Callaghan JM, Naderer T, Spurck T, McFadden GI, Currie G, Ferguson K, Bacic A, McConville MJ (2004) SMP-1, a member of a new family of small myristoylated proteins in kinetoplastid parasites, is targeted to the flagellum membrane in *Leishmania*. *Mol Biol Cell* 15:4775–4786
- Wilson ME, Young BM, Andersen KP, Weinstock JV, Metwali A, Ali KM, Donelson JE (1995) A recombinant *Leishmania chagasi* antigen that stimulates cellular immune responses in infected mice. *Infect Immun* 63:2062–2069

Worthey EA, Martinez-Calvillo S, Schnauffer A, Aggarwal G, Cawthra J, Fazelinia G, Fong C, Fu G, Hassebrock M, Hixson G, Ivens AC, Kiser P, Marsolini F, Rickell E, Salavati R, Sisk E, Sunkin SM, Stuart KD, Myler PJ (2003) *Leishmania major* chromosome 3 contains two long convergent polycistronic gene

clusters separated by a tRNA gene. *Nucleic Acids Res* 31:4201–4210

Wu Y, Wang X, Liu X, Wang Y (2003) Data-mining approaches reveal hidden families of proteases in the genome of malaria parasite. *Genome Res* 13:601–616