Biol. Chem., Vol. 390, pp. 1185–1197, November 2009 • Copyright © by Walter de Gruyter • Berlin • New York. DOI 10.1515/BC.2009.124

CA/C1 peptidases of the malaria parasites *Plasmodium falciparum* and *P. berghei* and their mammalian hosts – a bioinformatical analysis*

Ke Xiao[1,2,3], Franz Jehle[3], Christoph Peters[3], Thomas Reinheckel[3], R. Heiner Schirmer[2] and Thomas Dandekar[1,2,3,]**

[1] Lehrstuhl für Bioinformatik, Universität Würzburg, Biozentrum, D-97074 Würzburg, Germany
[2] Biochemiezentrum der Universität Heidelberg, Im Neuenheimer Feld 504, D-69120 Heidelberg, Germany
[3] Institut für Molekulare Medizin und Zellforschung, Fachbereich Medizin, Universität Freiburg, D-79104 Freiburg, Germany

**Corresponding author
e-mail: dandekar@biozentrum.uni-wuerzburg.de

Abstract

In genome-wide screens we studied CA/C1 peptidases of malaria-causing plasmodia and their hosts (man and mouse). For *Plasmodium falciparum* and *P. berghei*, several new CA/C1 peptidase genes encoding proteases of the L- and B-family with specific promoter modules were identified. In addition, two new human CA/C1 peptidase loci and one new mouse gene locus were found; otherwise, the sets of CA/C1 peptidase genes in man and mouse seem to be complete now. In each species studied there is a multitude of CA/C1 peptidases with lysosomal localization signals and partial functional overlap according to similar but subfamily-specific structures. Individual target structures in plasmodia include residues specifically different in CA/C1 peptidase subsite 2. This is of medical interest considering CA/C1 peptidase inhibition for chemotherapy in malaria, malignancies and other diseases. Promoter structures and mRNA regulation differ widely among CA/C1 peptidase subfamilies and between mammals and plasmodia. We characterized promoter modules conserved in mouse and man for the CA/C1 peptidase families B and L (with the L-like subfamily, F-like subfamily and mouse-specific J-like subfamily). RNA motif searches revealed conserved regulatory elements such as GAIT elements; plasmodial CA/C1 peptidase mRNA elements include ARE elements and mammalian mRNAs contain 15-lox DICE elements.

Keywords: cathepsin; expression; genome screen; malaria; papain-like proteases; protein structure.

*Electronic supplementary material to this article DOI 10.1515/BC.2009.124SUP is available from the journal online site at www.reference-global.com/toc/bchm/390/11.

Introduction

The cysteine cathepsins form a large ubiquitous family of proteases also known as papain-like proteases. Several family members, such as some SERA proteins, contain a serine residue at the active site instead of the canonical cysteine, and there are family members without apparent protease activity (McCoubrie et al., 2007). In this paper, we prefer the modern and more accurate term cysteine 'peptidases of clan CA family C1' (CA/C1 cysteine peptidases). In most organisms, CA/C1 peptidases are parts of protein digesting machineries but they can also mediate highly selective cleavage of specific protein substrates. There are two major families: the L-family and the B-family named after their representative CA/C1 peptidases, which were originally called cathepsins L and cathepsins B, respectively.

Human CA/C1 peptidases are relevant targets to treat osteoporosis, rheumatoid arthritis, osteoarthritis, bronchial asthma as well as cancer (Mohamed and Sloane, 2006). The cathepsin K inhibitor AAE-581 (balicatib) has recently passed phase II clinical trials as an anti-osteoporosis drug (Vasiljeva et al., 2007).

Plasmodial CA/C1 peptidases such as the falcipains are potential drug candidates against malaria (Singh et al., 2006). In the malaria parasite, the digestion of hemoglobin in the food vacuole, the processing of numerous proteins and the egress of the parasites from mammalian hepatocytes and erythrocytes rely on CA/C1 peptidases (Sijwali et al., 2006).

A census of plasmodial CA/C1 peptidases is therefore of biological and medical interest. There is functional overlap both within the CA/C1 peptidases family and of CA/C1 peptidases with unrelated proteases. To better understand this, both the multitude of CA/C1 peptidases (B-family, L-family and subfamilies) and their regulation of activity have to be taken into account. In their 2003 study, Puente et al. compared all human and mouse proteases available at that time (Puente et al., 2003). Subsequently Rossi et al. (2004) studied the human genome draft sequence applying the program TBlastN. In addition to the known cysteine cathepsins, their scan identified three pseudogenes, closely related to cathepsin L on chromosome 10 as well as two remote homologs, tubulointerstitial protein antigen and tubulointerstitial protein antigen-related protein. mRNA expression profiles for 10 known human cysteine cathepsins showed varying expression levels in 46 different human tissues and cell lines. No expression of any of the three cathepsin L-like pseudogenes was found which indicates that they are probably proteolytically inactive CA/C1 homologs. Based on these results, Rossi et al. concluded that all human cysteine cathepsins were known.

Building on these findings and starting from the set of known CA/C1 peptidases in man, mouse, *P. falciparum* and *P. berghei*, the following studies are presented here.

i. We have identified several incomplete CA/C1 peptidase sequences in all four genomes as well as two new complete human and one complete murine CA/C1 peptidase and several new plasmodial CA/C1 peptidases.
ii. The CA/C1 peptidases of the four genomes were subsequently compared in their geometric structures. Notably, the structural differences between peptidases from mammals and plasmodia include the S2 subsites.
iii. Promoter module structure as well as regulation of gene expression was found to be subfamily-specific and there are specific plasmodial promoter elements. Regulatory elements in CA/C1 peptidase mRNAs are conserved differently between CA/C1 peptidases from plasmodia and their mammalian hosts.

Results

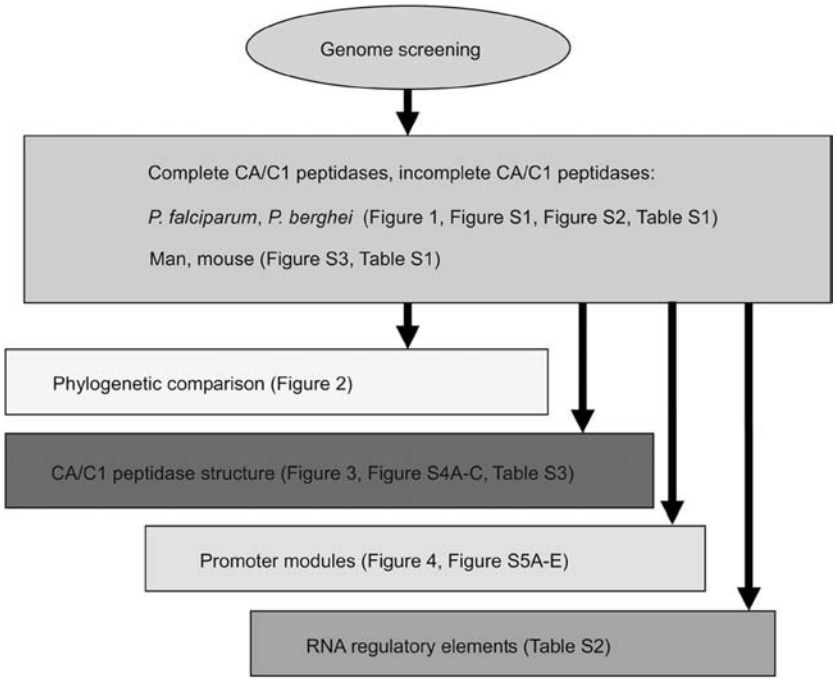
Scheme 1 summarizes the results including the supplementary material. Starting from genome screening including genomic shorter sequences, we consider phylogenetic distribution in all four organisms, peptidase structures including the S2 subsite, subfamily-specific promoter modules, as well as conserved regulatory elements in CA/C1 peptidase mRNA. At http://cac1peptidases.bioapps.biozentrum.uni-wuerzburg. de, we provide navigation within the scheme to individual Figures and Tables of the results (mini-database) and also provide direct links to several full-grown databases that we found useful for studying CA/C1 peptidases in different organisms.

Genome-wide screens for CA/C1 peptidases

Genome-wide screens were done on the genomes of the malaria parasites *P. falciparum*, *P. berghei* and their mammalian host genomes (man, mouse) applying sensitive sequence analysis methods (see the materials and methods section). The analysis revealed a number of new complete CA/C1 peptidases in plasmodia and only two new genome loci in man, as well as one locus in the mouse genome encoding additional CA/C1 peptidases. Table S1 in the supplementary material lists all identified 59 CA/C1 peptidases (sorted according to L-family with subfamilies and B-family, as well as within each subfamily for the species *P. falciparum*, *P. berghei*, mouse, and man). We include the newly identified complete CA/C1 peptidases (yellow background in Table S1) and identified shorter CA/C1 peptidases-like genome-encoded sequences (gray background in Table S1). We found nine new complete (two in *P. falciparum*, four in *P. berghei*, one in mouse, and two in man) and six incomplete CA/C1 peptidases (three in mouse and three in man), as well as two known SERA proteins (supplementary Figure S1).

New CA/C1 peptidase loci in the genomes of *P. falciparum* and *P. berghei*

The sequence analysis of the two parasite genomes confirms the previously known falcipains (L-like CA/C1 peptidases falcipain 1, 2, 2′ and 3; Sijwali et al., 2006) and



Scheme 1 Summary scheme of all results. Starting from the genome screening including shorter genomic sequences, we consider phylogenetic distribution in all four organisms, peptidase structures including the S2 subsite, subfamily-specific promotor modules, as well as conserved regulatory elements in CA/C1 peptidase mRNA.

the CA/C1 peptidases SERA6 and SERA7 (McCoubrie et al., 2007) but now includes the genome locus for falcipain 1 from *P. berghei* (Table 1). More importantly, we identified several new genome loci encoding complete CA/C1 peptidase sequences (Table 1) as well as incomplete sequences (supplementary Table S1). For the W-like CA/C1 peptidase in *P. berghei*, only an automatic prediction of 'hypothetical protein, putative cysteine protease' was previously known. However, domain composition and sequence analysis (see materials and methods section) support that this locus is probably encoding a W-like cathepsin. For all these plasmodia CA/C1 peptidases, evidence for expression was checked (Figure 1) as well as intactness of the catalytic triad and of the reading frames (supplementary Figure S2). The potential overlap in catalytic specificity and function should also be taken into account for efforts in drug design against plasmodia (Goh and Sim, 2005; Sijwali et al., 2006). Some plasmodial chromosomes harbor several CA/C1 peptidases. There are falcipains in chromosome 11 in *P. falciparum* (Table S1). Furthermore, there is a cluster of eight SERAs on chromosome 2 in *P. falciparum* (Miller et al., 2002).

All CA/C1 peptidase loci of the human and murine genomes

Despite extensive screens, in addition to the previously characterized CA/C1 peptidases we only found shorter cathepsin fragments in mouse and man (Table S1; Figure S3), as well as two complete L-like CA/C1 peptidase loci in man and one in mouse encoding complete reading frames with no stop codons and with evidence for expression from expressed sequence tag (EST) data (Figure S3). These yield Table 1 as the complete list of the CA/C1 peptidases in mouse and man (shorter sequences in Table S1). Probably there are no further hidden CA/C1 peptidases in the human genome and the nomenclature of the human CA/C1 peptidases can be finalized. We detected CTSL3 as a new gene locus. Its gene name is cathepsin L-family member 3, an alias hCTSL-s and the Gene identifier GID 392360. Looking at EST evidence we observe that it is also expressed but based on EST evidence it is not certain whether the complete sequence is expressed.

The second locus, CTSLL4 (alias hCTSLL) gene from man, was not analyzed further or named before, but the

Table 1 Complete CA/C1 peptidase sequences that were newly identified[a].

Name	Chromosome[d]	Location	Covering ESTs	Length (aa)	M_w (kDa)
Plasmodium falciparum					
Cathepsin C putative[b]	Chr. 12	NC_004316 1968493-1971057	BM276389 CA856661 CA856571 BM274201	504	59.5
PFD0230c[b]	Chr. 4	NC_004318 266922-269829	BI815922 BQ596335 BI815232 BU496508	939	110.5
PFB0335c[c]	Chr. 2	NC_000910 298900-301733	AJ555580 BI816005 BQ451595	893	104
PFB0330c[c]	Chr. 2	NC_000910 294276-297394	BU497091 BU497271 BU498055 AU088388	946	109.6
Plasmodium berghei					
Falcipain-1 (*P. berghei*)	Unclear	CAAI01000515 3892-5446	BF298730 BB980790 BB979022 BB972286	517	60.0
PB000352.01.0 (S-like)	Unclear	CAAI01006605 420-3104	BF298768 BB978429 BB971003	829	96.2
PB000233.03.0 (W-like)	Unclear	CAAI01002787 416-1951	BM160328 BP114560	454	53.0
PB000856.03.0	Unclear	CAAI01003269 1273-3314	BP114222 BP113912 BP113908	680	78.6
Human					
hctsl-like (CTSLL4)	Chr. 10q22.3	NC_000010 52135146-52137882	DN831176 DN831175 BG219223 BG202425 BG201222 BG217032 BD119660 CN412434 BD119661 AX984802	332	37.4
hctsl-similar (CTSL3)	Chr. 9q22.1	AC_000052 60958324-60972276	CQ733879 AA367954 AX914832 BD050365 CV423299 AW885791	297	31.9
Mouse					
mCTSL-like (mCTSLL)	Chr.13	NC_000079 60808536-60811038	BY735637 BY756020 BY709516 CA550116	330	36.6

[a]Including chromosomal location (if known) and covering EST evidence, as well as predicted protein length and molecular mass.
[b]Putative cathepsin C and PFD0230c are *Plasmodium falciparum* cysteine proteases different from known SERA proteins because they are located on Chr. 12 and Chr. 4, respectively. For more details, see the supplementary material (Figure S1).
[c]PFB0335c and PFB0330c are identical to SERA6 and SERA7 of *P. falciparum*, according to multiple alignment comparison results (supplementary Figure S1).
[d]Abbreviated by 'Chr.'.

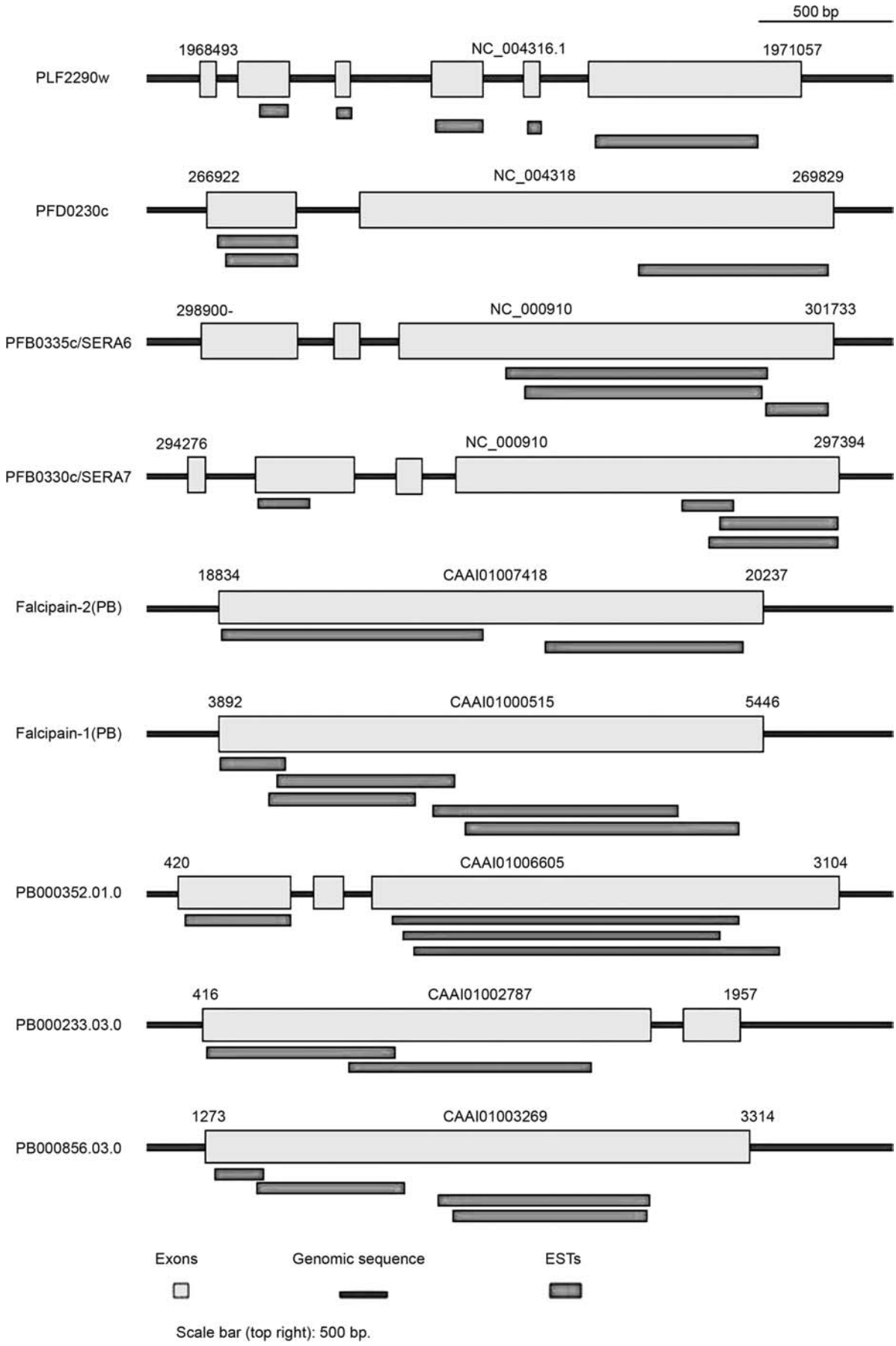


Figure 1 Map of the predicted putative new CA/C1 peptidases from *P. falciparum* and *P. berghei.*
The genome loci of the identified plasmodial CA/C1 peptidases are shown. Light grey rectangles on the DNA strand (black) indicate predicted exons from the genome sequence. Shown below are in dark grey found corresponding ESTs. A scale bar indicates 500 base pairs.

gene locus was already integrated in ENSEMBL (http://www.ensembl.org/Homo_sapiens/geneview?gene=OTTHUMG00000018237;db=vega) as part of the chromosome 10 sequencing effort (Deloukas et al., 2004).

Regarding the additional murine complete cathepsin locus mCTSLL found by the screen, this is confirmed by a RIKEN cDNA clone (Carninci et al., 2005) stressing its expression and is in agreement with the hypothetical protein locus given for the RIKEN cDNA2310051M13 gene. We detected further short L-like genomic sequences in the mouse genome, but only MER26800 has covering ESTs indicating expression. These ESTs are clearly different from those covering the cathepsin L-like genome loci including the mCTSL-like locus (Table S1; Figure S3).

The four new murine L-like loci described here all cluster on chromosome 13 as do several J-like murine CA/C1 peptidases (Deussing et al., 2002) preferentially expressed in placenta. Human clusters of CA/C1 peptidases are on chromosomes 9 and 10 (Table S1).

Phylogenetic comparison

To obtain a comparative overview we next compared all available sequences in a phylogenetic tree (Figure 2). Asterisks (*) denote the newly identified CA/C1 peptidases. The observed branching pattern is statistically well supported in its major branches by high bootstrap values (Figure 2; well above 500 and in most cases close to 1000 for 1000 bootstrap trials). Phylogenetic analysis of all known complete CA/C1 peptidases and the new sequences given here show that there are again the L- and B-families in plasmodia but in the L-family there is no J-like subfamily nor a F-like subfamily with the exception of the cathepsin W-like genome loci in *P. berghei*. The L-family in man has an L-like and an F-like subfamily (Figure 2, bottom), whereas in mouse the L-family has three subfamilies: L-like, F-like, and J-like family. The latter is expressed mainly in placenta (Figure 2, middle), murine CA/C1 peptidases mctsj, mctsm, and R, -1, -2, -3, and -6 (species specific trees are available on request from the authors). The families are thus limited, new complete CA/C1 peptidases were only found for the cathepsin L-family (L-like subfamily in all four genomes; F-like subfamily with cathepsin W in plasmodia) or the B-family. Genomic fragments were either L-family (man) or B-family (mouse, plasmodia).

Besides the cysteine cathepsins there are also serine repeat antigens (SERAs), a family of secreted 'cysteine-like' proteases of Plasmodium parasites. The human malaria parasite *Plasmodium falciparum* possesses six 'serine-type' (SERA1–SERA5 and SERA9) and three 'cysteine-type' (SERA6–SERA8) SERAs. Also, these are functionally redundant; however, SERA5 seems to be particularly important for blood stage survival and presents a potential valuable target for therapeutic intervention (McCoubrie et al., 2007). The relationship of these SERA proteins, in particular the cysteine-type proteins, to the newly discovered B- and L-like CA/C1 peptidases was analyzed by sequence alignment and phylogenetic analysis (Table 1 and Figure S1). The serine-type SERA-like proteins are related but are different in their sequence to the plasmodial CA/C1 peptidases described here. However, PFB0335c_PF* is identical to SERA6, PFB0330c_PF* is identical to SERA7 (Table 1; both marked by +). *P. falciparum* CA/C1 peptidase cathepsin C′ and CA/C1 peptidase PFD0230C are identified cysteine proteases and they are different from the SERA family. Furthermore, cathepsin C′ is located at chromosome 12, PFD0230C is located at chromosome 4, but all known SERAs are from chromosomes 2 or 9.

Structure analysis

Regarding differences in structure or catalytic activity, the mammalian L- and B-family established previously remain the same when including the predicted structures according to the new genome-derived sequences (Figure S4). Specific differences include the B-exclusion loop in B-like CA/C1 peptidases (Turk et al., 2003). Cathepsin B inhibitors are a potential strategy against metastasis (Vasiljeva et al., 2006; Caglic et al., 2009). However, the identified short expressed CA/C1 peptidase sequences from the genome-wide searches are incomplete in their catalytic residues and could thus have an inhibitory effect or function. In plasmodia, all established CA/C1 peptidases are again L-family (e.g., the falcipains) or B-family members. Expression as complete proteins and, hence, catalytic activity for these plasmodial CA/C1 peptidases is predicted from EST results (in each case the expressed sequence is complete and has all residues necessary for catalysis conserved).

The structures of the F-like subfamily as part of the L-family of mouse and man are still more diverse and this applies also for the W-like CA/C1 peptidases in *P. berghei* (Figure 3 and Figure S4). Furthermore, several of the plasmodial sequences have no cathepsin template or can only in part be homology modeled (see the materials and methods section). PB000590 and PB000888 had no template with enough similarity (at least 30%) to derive structure models. Regarding PFD0230c, the residue range 484–570 shares similarity of 31% to pancreatic α amylase (pdb code 1k3b, chain B, residues 207–302), so here only a part of the final structure can be predicted by homology modeling and these sequences are predicted to fold differently from the known CA/C1 peptidase crystal structures.

Detailed homology models for CA/C1 peptidases highlight specific residues which occur only in this individual CA/C1 peptidase. This is shown for all newly identified complete CA/C1 peptidases and SERA7 (Figure 3; ball and stick representation). Supplementary Figure S4 shows crystal structures or homology models for all complete CA/C1 peptidases including SERA6, as well as for all identified incomplete CA/C1 peptidases. The panel is sorted according to families and subfamilies.

As a first basis for drug targeting the plasmodial CA/C1 peptidases with preference but not the mammalian peptidases, we investigated the *substrate binding site S2*. The alignment (Figure S4B) shows that there are some notable traits, e.g., residue 133 of the S2 subsite is often a serine or a glutamine in the plasmodial peptidases, and the phylogenetic tree on the S2 subsites (Figure S4C) shows that there is a tendency (three related branches) of most plasmodial S2 subsites to cluster together except falcipain 2 from *P. berghei* (the latter has an uncertain position and low bootstrap support).

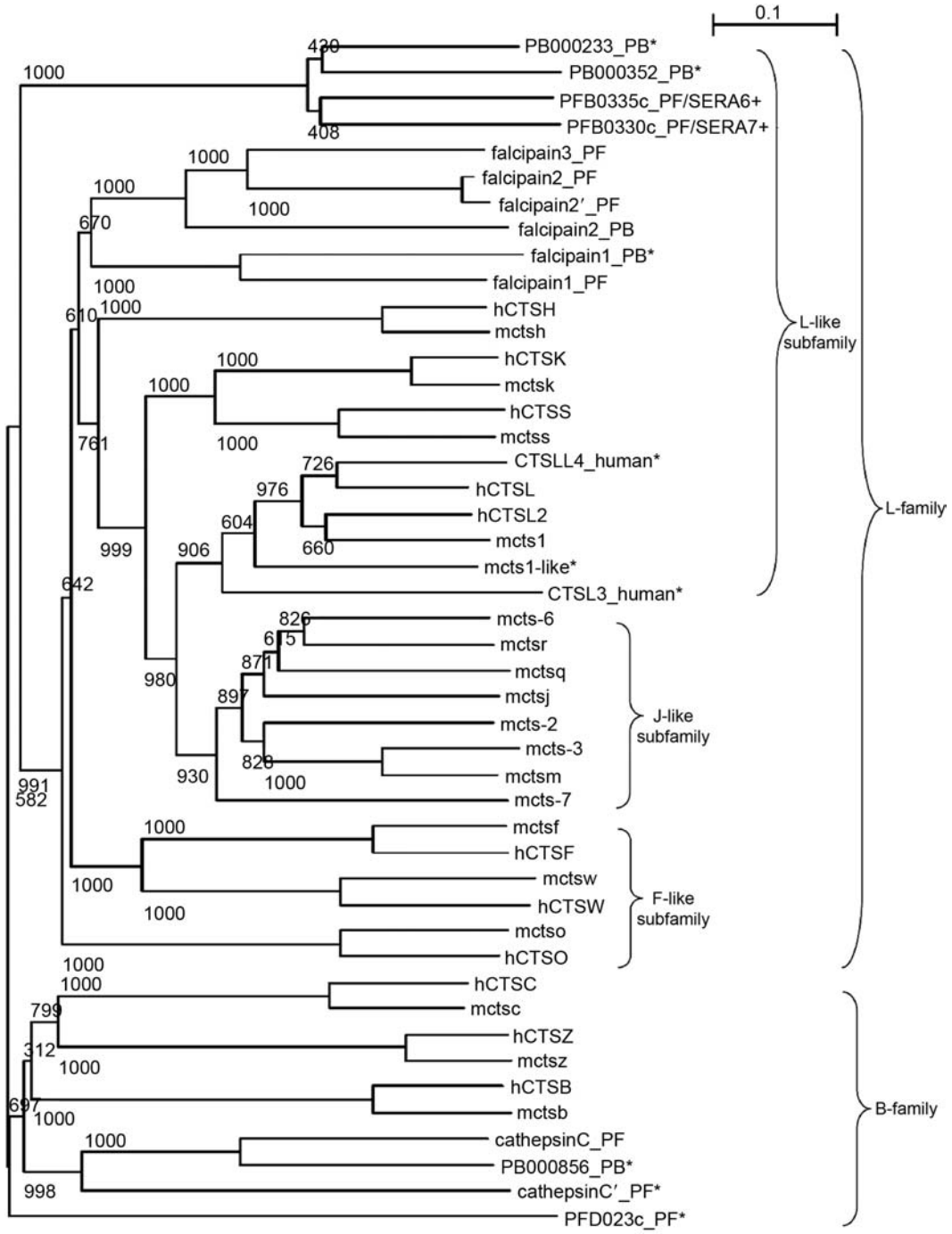


Figure 2 Comparison of the encoded protein sequences derived from the newly found cathepsin gene loci in four genomes. Predicted protein sequences are compared within a tree of all known complete human cathepsin protein sequences (winged brackets embrace the L-family and B-family) and include L-subfamilies (L-like, F-like, and murine J-like subfamily; a detailed tree is given in the supplementary Figure S3). Novel sequences are marked with an asterisk. The two *P. falciparum* sequences identical to SERA proteins are marked with $^{+}$. Significant bootstrap values (1000 trials) are indicated. Detailed organism-specific trees are available on request from the authors.

The resulting specific S2 subsite cleft topology is shown in the ball and stick representation for plasmodial CA/C1 peptidases PB000856 and PFB0335c/SERA6 (Figure 3, bottom). This can, in principle, be targeted by suitable modified peptidase inhibitors, which nevertheless will require a lot of further considerations and experimental work. Another interesting feature is the C-terminal hemoglobin binding site in falcipain 2 (Pandey et al., 2005). Deletion (or pharmacological targeting) of this site prevents falcipain 2 to digest hemoglobin and does not impair its general peptidase activity but impairs its interaction with the prodomain, its natural inhibitor. Our sequence data (Figure S2) show that this falcipain 2 site is in fact only shared by falcipain 2′ and no other CA/C1 peptidase from the four organisms studied. This provides a way to target only these falcipains. Both are expressed

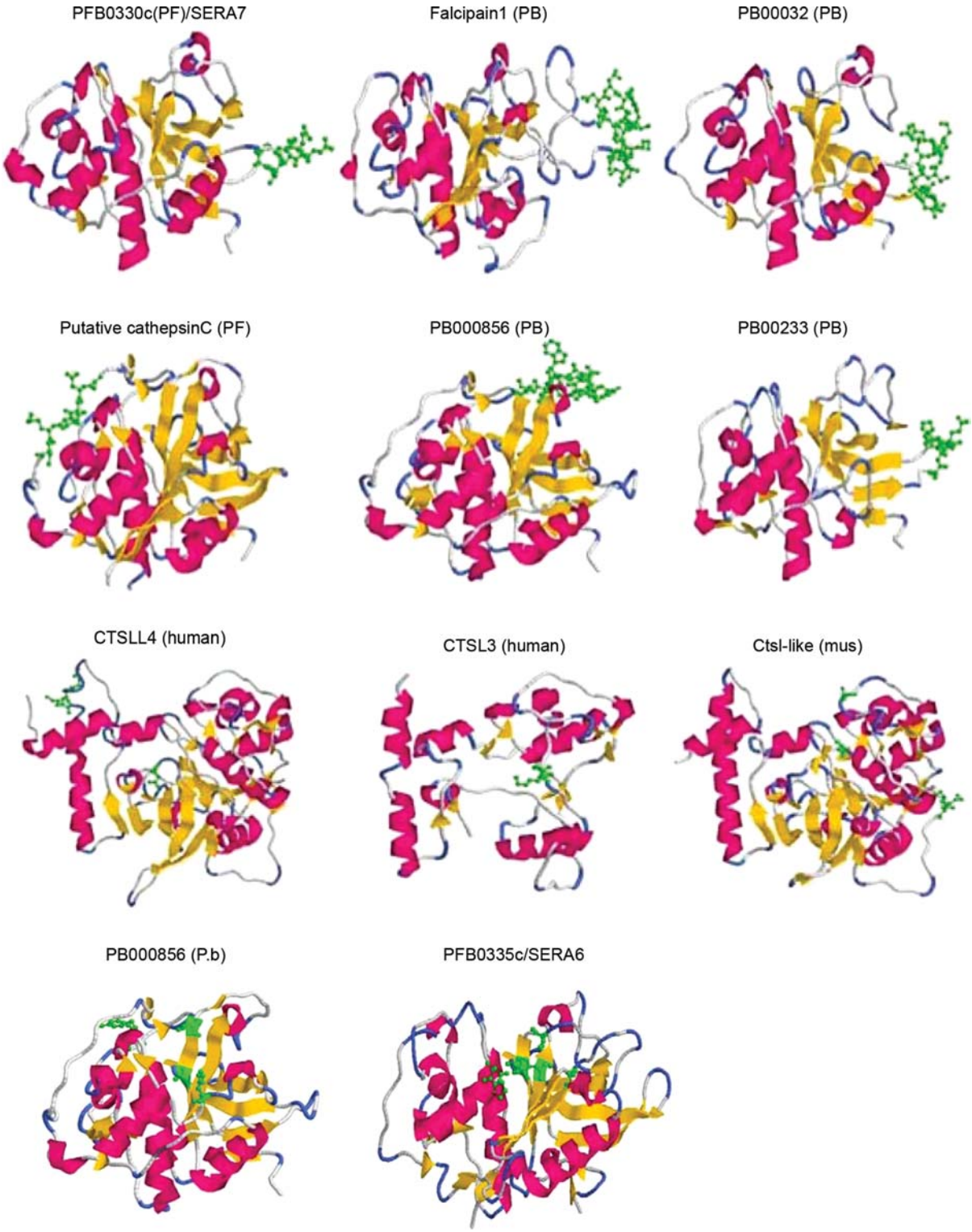


Figure 3 Structures of the newly identified CA/C1 peptidases.
The panel shows homology models (helices in red, strands in yellow, loop regions in gray) calculated and drawn for all identified complete CA/C1 peptidases. Specific residues unique for the specific structure are labeled in green and shown in a ball and stick representation of the side chain. The first two rows show six plasmodial CA/C1 peptidases (left: *P. falciparum*, middle and right: *P. berghei*). The next row shows predicted structures for the two newly identified complete human cathepsins and (right) the murine complete cathepsin. The bottom models examine the S2 subsite, highlighting residues (listed in supplementary Figure S4B) specific for Plasmodia (shown for *P. berghei* PB000856 and *P. falciparum* SERA6).

in different plasmodial stages (see discussion section) so that specific targeting of falcipain 2 or falcipain 2′ is possible. Similar mechanisms might also target other CA/C1 peptidases to more specific biological substrates despite their general activity. We noted at least individual peptidase-specific differences in the C-terminal region. Different structure-activity relationship studies have been carried out on several malaria and human papain family

enzymes (Selzer et al., 1999; Krasky et al., 2007; Dude et al., 2008; Kerr et al., 2009) and this could be used to bolster such efforts.

Finally, we also looked at localization signals in the different CA/C1 peptidases applying the LocTree support vector machines (see materials and methods section). As expected, most of the CA/C1 peptidases compared are predicted to be clearly lysosomal across orthologous enzymes from different organisms (Table S3). This is also true most plasmodial peptidases. In addition, they have corresponding prosite and pfam signatures (motifs listed in Table S3 known to be associated with this localization). For instance, pfam motif PF08246 is specific for most L-family CA/C1 peptidases. However, four plasmodial CA/C1 peptidases including SERA6 and SERA7 are differently localized and predicted by the LocTree algorithm to be secreted.

CA/C1 peptidase gene promotor structures and regulatory RNA elements in mRNA

Transcription factor binding sites (TFBS; see details in supplementary material) assemble in specific combinations to subfamily-specific promoter modules. Applying genomatix software (see materials and methods section), we found that plasmodia CA/C1 peptidase promoters are different from mammalian peptidase promoters as shown by shared modules in *P. falciparum* for falcipain 2, falcipain 3 and cathepsin C (Figure 4; Figure S5), which are partly conserved in *Plasmodium berghei* CA/C1 peptidase promoters (*P. berghei* falcipain 2 and PB000352 CA/C1 peptidase gene).

Supporting known expression differences for different CA/C1 peptidase families (Deussing et al., 2002; Turk et al., 2002; Sijwali et al., 2006; Vasiljeva et al., 2006), we can show for L-family and B-family CA/C1 peptidases that the different promoters have specific transcription factor binding sites conserved in subfamilies and clearly different compared to the plasmodia promoters (Figure S5 and legend; see materials and methods section). Of pathological interest is a desmin element, also called HMTB element, found in the L-like promoters and known to be a muscle- and heart-specific regulator and implicated in regulatory processes in heart insufficiency (Petermann et al., 2006). It can also be implicated in pathological processes, such as heart insufficiency, after cathepsin L knockout in mice (Petermann et al., 2006). In contrast, the B-family promotor structure leads, e.g., for cathepsin B to pronounced expression in macrophages as well as in tumor tissue. This has profound effects on cell connectivity and promotes metastasis (Vasiljeva et al., 2006). Key C1/CA peptidase promoter modules are furthermore conserved in mammals. Thus, the human and murine L-like subfamily promoter elements AP2F, NKXH, and SP1F are also found in the promoter sequences of cathepsin L of chimpanzee (*Pan troglodytes*) and rat (*Rattus norvegicus*). The desmin element is conserved in man (CTSL, CTSLL4, and CTSL2), chimpanzee (CTSL and CTSL2) and mouse CTSL. Regarding the B-family promoter modules, the SP1F element, EBOX and ETSF are again confirmed in cathepsin B promoters from chimpanzee and rat. Finally, within the L-family we observe F-like subfamily-specific

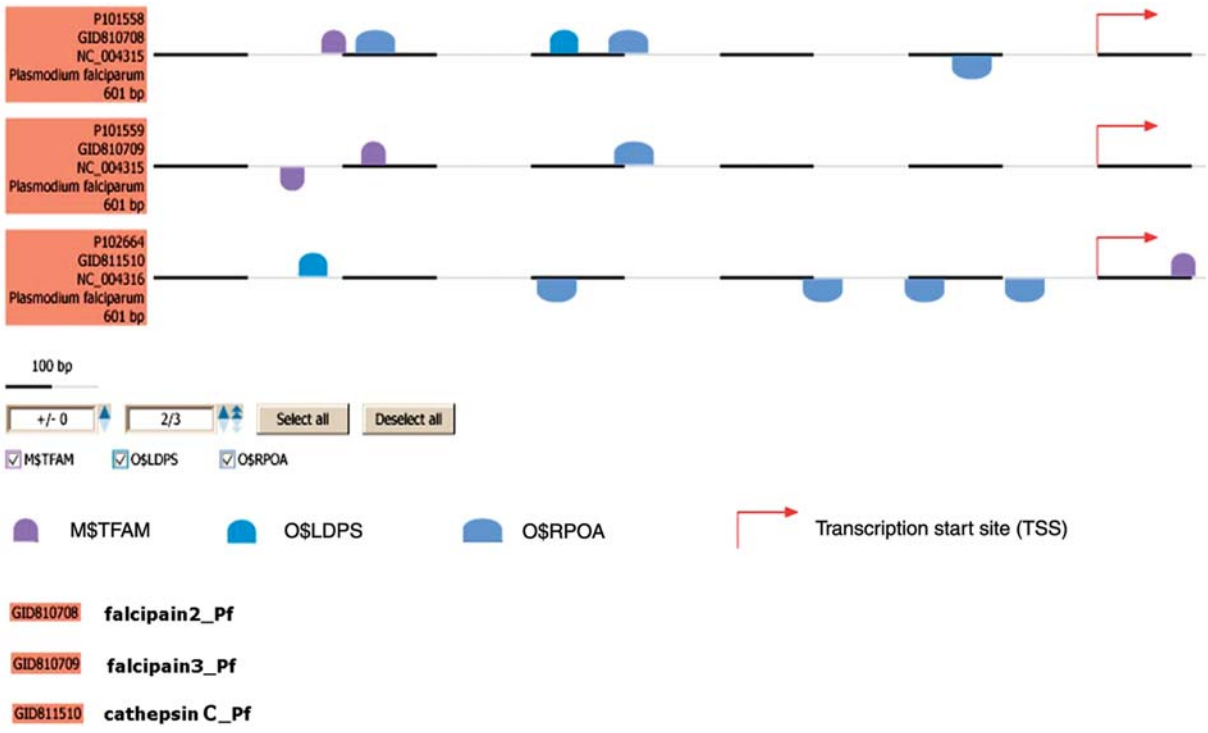


Figure 4 Promoter structures in plasmodial CA/C1 peptidases.
Common transcription factor binding sites (TFBS, differently colored half circles) shared by falcipain 2, falcipain 3, and cathepsin C of *Plasmodium falciparum* are TFAM (Choi et al., 2002), LDPS (Chang and Gay, 2001) and RPOA (Hall and Milcarek, 1989), respectively; details are given in the supplementary material (text and Figure S5).

promoter elements conserved in cathepsin F promoters from rat and chimpanzee.

The identified subfamily-specific promoter modules are also partly conserved in the shorter genome encoded mammalian CA/C1 peptidases without complete catalytic triad (see supplementary Figure S5). Additional specific promoter binding elements add further to each specific cathepsin or cathepsin subgroup promotor structure but are not so commonly shared between cathepsin promoters. For instance, human cathepsin L2 has specific elements (Figure S5B) and hence its expression is confined to skin (*stratum corneum*), thymus, testis and cornea and its expression levels inversely correlate with skin color (Chen et al., 2006). As new data and experiments characterize the promoter structure further, more detail will be added to this analysis. Available data nevertheless show the presence of promoter elements that are conserved in a subfamily-specific way, whereas additional promotor elements fine-tune expression of individual CA/C1 peptidases and suggest new targeting strategies (see discussion section).

After DNA elements for regulation of gene transcription we also looked at regulatory elements at the mRNA level. Different RNA elements are used in the cathepsin mRNAs (Table S2). GU-rich RNA element and GAIT element occur across species in nearly all CA/C1 peptidases, the k-Box occurs often. ARE elements are present in most plasmodial UTRs but not in mammalian cathepsin mRNAs. Brd Box and GY Box and in particular 15-lipoxygenase differentiation control element elements (15-lox DICE) occur only in the mammalian hosts.

The conservation of these specific RNA elements supports their functional or organism specific relevance, whereas additional experimental data are required to measure details of the regulation.

Discussion

New CA/C1 peptidases in plasmodia genomes

In previous work, we have identified new proteins and function by detailed comparative sequence analysis of genomes in prokaryotes (Dandekar et al., 2000; Gaudermann et al., 2006). The present study applies similar bioinformatic techniques for a comparative genome analysis in plasmodia and their mammalian hosts. Genome data are complemented by direct data on gene expression from EST databanks, detailed analyses of sequence and structure, as well as of promoter and RNA elements. Nevertheless, all these data can certainly be further extended by biochemical studies and will be developed in the future regarding individual CA/C1 peptidases and their specific functions and regulation.

To the well known falcipain family in plasmodia only one genome locus encoding *P. berghei* falcipain 1 was added. However, the genomic screens pointed to several new genome loci encoding and expressing complete B-like plasmodial (*P. berghei* and *P. falciparum*) and W-like CA/C1 peptidases (*P. berghei*) to be considered in antimalarial strategies based on cysteine protease inhibitors.

A recent study (Sunil et al., 2008) revealed that a ~1-kb sequence upstream of the translational start site is sufficient for the functional transcriptional activity of the falcipain-1 gene, whereas falcipain-2, -2′ and -3 genes that exist within the 12-kb stretch on chromosome 11 require ~2 kb upstream sequences for the expression of reporter luciferase activity. Falcipain-1, 2, and 3 promoters exhibited maximum activity in the throphozoite stage. In contrast, falcipain-2′ showed an expression maximum in the schizont stage. EMSA analysis elucidated binding of distinct nuclear factors to specific sequences within the 5′ upstream regions of falcipain genes. These authors found parasite specific sequence elements such as poly(dA) poly(dT) tracts, CCAAT boxes, as well as a single 7-bp G-rich sequence, (A/G)NGGGG(C/A) in the 5′ upstream regulatory regions of these genes. However, their experimental data suggest that expression of plasmodial CA/C1 peptidases is regulated at the transcriptional level indicate for different transcription factor complexes; our data define specific transcription factor binding sites involved in this.

Intraerythrocytic malaria parasites use host hemoglobin as a major nutrient source. Aspartic proteases (plasmepsins) and cysteine proteases (falcipains) function together in the early steps of the hemoglobin degradation pathway. There is extensive functional redundancy within and between these protease families. Plasmepsins are synthesized as integral membrane proenzymes that are activated by cleavage from the membrane. This cleavage is mediated by a maturase activity for which identification has been elusive. Plasmepsin processing occurs primarily via the falcipains; however, if falcipain activity is blocked, autoprocessing might take place, serving as a weaker alternative activation system (Drew et al., 2008).

Also egress of *Plasmodium* blood stage merozoites, liver stage merozoites and mosquito midgut sporozoites relies on protease activity. Involved are the cytoskeleton-degrading malarial proteases falcipain-2 and plasmepsin II, plus the family of putative papain-like proteases called SERA. Recent experiments have shown that activation of the SERA proteases might be triggered by regulated secretion of a subtilisin-like serine protease called SUB1 (Blackman, 2008).

Regarding pharmacological targeting, the situation is complex with protease multitude and overlap of function in specific families and subfamilies. However, the overlap in function was shown to be only partial for several falcipains (Sijwali et al., 2006). Our data help to improve drug strategies targeting plasmodial proteases. For instance, falcipain 2′ is a promising target in *P. falciparum* schizonts when combining the facts that this enzyme is the predominant falcipain in schizonts and that there are structural differences in subsite S2 structures among CA/C1 peptidases. The targeting strategies can be further improved by combining protease inhibitors with other drugs such as methylene blue (Akoachere et al., 2005).

The list of CA/C1 peptidases in the host genomes is probably complete

The names of the new human CA/C1 peptidases and gene loci identified in this study in addition to Puente et al. (2003) were given and confirmed in accordance with the human gene nomenclature committee (HGNC) commission (in collaboration with Dr. Tam Pederson Sned-

don, HGNC). Our extensive screening revealed only two additional genome loci for complete L-like CA/C1 peptidase sequences. One locus (hCTSLL4) is part of chromosome 10, sequenced previously by Deloukas et al. (2004). The newly discovered locus hCTSL3 (alias hCTSLS) is complete in sequence and is expressed (Figure S3). The new locus in mouse (also complete and expressed) has complementary evidence from the RIKEN cDNA consortium (Carninci et al., 2005). In summary, we expect that there are no further complete CA/C1 peptidase genes hidden in the host genomes and that we also identified most of the shorter fragments.

Comparative analysis between human and mouse proteases indicates a high percentage (82%) of mouse genes having a strict ortholog in the human genome (Puente et al., 2003). Shorter pseudogenes and the two new complete human gene loci are L-like (Figures S2 and S3; Table S1). Several shorter genome encoded cathepsin-like fragments identified in man are all N-terminally truncated and B-like (Figure S2). Incomplete and complete L-like sequences are found in the mouse genome (Figure S3; Table S1). These identified sequences are clearly different from previously reported sequences. A subset is confirmed by EST data, compatible with expression. A modulating function in protein degradation was investigated and suggested for propeptides by Wiederanders et al. (2003).

Differences in structure

Structure models were investigated to better delineate differences among the CA/C1 peptidases, in particular it would be interesting to find plasmodia-specific structure elements which could be targeted by drugs. We can clearly show that regarding structure of the complete CA/C1 peptidases there are only the two previously known families. This is true both for the previously identified and the new genome encoded CA/C1 peptidases in man and mouse, as well as for the two plasmodia investigated. However, regarding genome loci encoding shorter and incomplete CA/C1 peptidases, some of them differ sufficiently in sequence so that other templates are preferred for homology modeling. Targeting such shorter, probably, protease inhibitory sequences could provide a new strategy for anti-plasmodial drug design. Furthermore, there are specific differences in individual complete CA/C1 peptidases which can be targeted, including plasmodia-specific structural features in the S2 subsite (Figure 3 and Figure S4). Individual CA/C1 peptidases and subfamilies have sufficient structural and functional differences to suggest a strategy simultaneously attacking several plasmodial CA/C1 peptidases (Krasky et al., 2007; Dude et al., 2008; Kerr et al., 2009) while largely sparing the human CA/C1 peptidases (Palermo and Joyce, 2008). Such studies should be conducted in parallel to serine protease inhibition (Rupp et al., 2008). Moreover, anti-osteoporosis drugs such as cathepsin K inhibitor AAE-581 (balicatib; Vasiljeva et al., 2007) can now be improved further on the basis of our data by targeting the specific S2 subsite structure of cathepsin K (Figure S4B). Most CA/C1 peptidases have lysosomal localization signals; however, for four plasmodial peptidases we predict secretion signals.

Differences in elements for CA/C1 peptidase expression regulation

We finally looked at differences regarding regulation, either at the gene level or at the RNA level. Conserved promotor elements in gene loci were compared using genomatix software (Werner, 2003). The conservation of subfamily-specific promotors in mammals is contrasted by different promoter elements for CA/C1 peptidases from plasmodia. Important promotor elements were further confirmed looking for conservation in other mammalian organisms and their CA/C1 peptidases (chimpanzee and rat). These promoter differences are directly mirrored by differences in expression time and tissue preferences of the respective cathepsins. These results suggest that targeting strategies against falcipains and human cathepsins, respectively, should not only be based on differences of the enzyme structures, e.g., at subsite 2 but also on the biology and expression chronology of these proteins.

Finally, mRNA elements have to the best of our knowledge not yet been analyzed systematically in CA/C1 peptidases. Several elements show a striking conservation, e.g., the GAIT element, whereas others often occur in plasmodia but not in the host (ARE) or only in the host but not in plasmodia (15-lox DICE, Brd-Box, GY Box). This shows differences for regulatory elements on the level of CA/C1 peptidase mRNAs and supports specific regulation for different CA/C1 peptidases and for different organisms at the translational level. However, to understand the detailed regulatory consequences of these features will require further experimental studies on the individual mRNAs for CA/C1 peptidases.

Conclusion

We have established a complete list for CA/C1 peptidases in plasmodia and their host genomes (man and mouse) that also includes shorter sequences. Newly identified plasmodial CA/C1 peptidases cover four *P. berghei* and two *P. falciparum* gene loci (L-like, B-like), as well as two new human CA/C1 peptidase loci and one new mouse gene locus. This list is considered complete for the CA/C1 peptidase family of all four organisms with the possible exception of very short and degenerate sequences. The repertoire of CA/C1 peptidase families is limited (L- and B-family) but it includes structural differences, such as in subsite 2. In all studied species, we observe a multitude of CA/C1 peptidases with functional overlap but yet clear differences in regulation, e.g., on the basis of promoter structure and mRNA regulation. Pharmacological targeting of plasmodial CA/C1 peptidases should thus exploit a combination of structure and regulation differences, e.g., by targeting subsites of falcipain 2′ which is the major expressed falcipain in the blood schizontal stage of all known plasmodia species.

Materials and methods

Initial sequence set and databanks used

The initial set of sequences used as queries in the human genome were all known cathepsin sequences as of December

2006. For the sequence searches in other organisms in addition to the CA/C1 peptidases known from mouse, *P. falciparum* and *P. berghei* were used. The following databanks provided the starting sequences as direct CA/C1 peptidase protein entries: Swiss-Prot databank (Boeckmann et al., 2003) and MEROPS peptidase databank (Rawlings et al., 2002) were used as systematic protein databanks to provide a clear and direct outlook on the biological features of already known CA/C1 peptidases. The high throughput genomics division of Genbank (Benson et al., 2002) and the mouse genome database (Blake et al., 2001) were explored to examine the human and mouse complete genomic sequences, as well as to also obtain the known murine CA/C1 peptidases. Furthermore, the Ensembl databank system (Hubbard et al., 2005) provided human and mouse genome data, both finished and draft sequences. Extensive screening of the plasmodia genomes relied on PlasmoDB, GenBank, TIGR databank, and EBI databank. PlasmoDB (release 5.3; http://plasmodb.org/plasmo/) provided the list of known plasmodial CA/C1 peptidases (e.g., falcipains). A further recommended source is the mammalian degradome database (Wu et al., 2003; http://degradome.uniovi.es) because all mouse and human cysteine proteases are listed here. Web pointers to all these resources are given at http://cac1peptidases.bioapps.biozentrum.uni-wuerzburg.de.

Sequence analysis

The cathepsin starting sequences (see above) were used to search for new and undetected cathepsin protein sequences in organism genome data by iterative sequence alignment searches (Altschul et al., 1997) with a conservative threshold (expected value of chance hit less than 10^{-6}) to detect similarity and relationships between different CA/C1 peptidases and other proteins. Significant sequence similarities were further validated (Yeh et al., 2001; Gaudermann et al., 2006) by back searches in the database applying the same conservative threshold (expected value of chance hit less than 10^{-6}). Furthermore, we analyzed and noted the chromosomal location of each gene locus identified.

Predictions from genomic sequences for CA/C1 peptidases were translated and analyzed in detail. Domain and motif searches applied the simple modular architecture research tool (SMART; Letunic et al., 2004) and the database of protein domains, families and functional sites (PROSITE; Hulo et al., 2004). Completeness of critical residues for function and catalysis was tested using protein specific PROSITE signatures (no PROSITE motif mismatches were allowed in the comparison). Several sequences were shorter (Table 1) and incomplete in this regard. The typical CA/C1 peptidase domain composition was verified using the simple modular architecture research tool (expected value of chance hit less than 10^{-3}).

To test for expression of the genome sequence, Gene2EST software (Gemünd et al., 2001) enabled efficient retrieval of ESTs matching large genomic DNA queries (setting: expected value of chance hit less than 10^{-6}). This helped us to integrate EST data from the human genome project as well as dbEST data from other organisms to further identify and characterize potential novel CA/C1 peptidase genes: a sequence similarity match from the EST data was considered positive if at least 80% of the sequence could be mapped into the investigated genome with at least 95% similarity. In human EST mapping, we always considered only the high quality matches ('gold' and 'experimentally verified 5′ complete transcript').

Phylogenetic analysis and alignments

To compare and correctly place the newly identified CA/C1 peptidase sequences and to analyze the resulting cathepsin subfamilies, phylogenetic analysis was conducted. To calculate multiple alignments and phylogenetic trees, the programs ClustalW (Thompson et al., 1994), ClustalX and Njplot (Perrière and Gouy, 1996) were applied. This phylogenetic analysis helped to identify such similarities and differences, also among individual CA/C1 peptidases. Seaview (Galtier et al., 1996) provided an editor for manual corrections of the automatic multiple alignment (e.g., to preserve conservation of key catalytic residues in the obtained alignment of cathepsin sequences). Bootstrap values were calculated as a standard measure to indicate how often the branching pattern of the phylogenetic tree was supported if the alignment columns were randomly shuffled.

Structure analysis and homology models

Next, structure analysis was done as plasmodial CA/C1 peptidases are potential antibiotica targets, and human CA/C1 peptidases can be targeted by anti-cancer drugs against metastasis. Furthermore, residues specific for an individual CA/C1 peptidase available from the complete alignment of all CA/C1 peptidases (Figure S2) can be mapped on the three-dimensional coordinates and structures (Figure 3). In this way, strategies to target, e.g., plasmodial CA/C1 peptidases or falcipains can be improved (structure coordinates are available from the authors). For several CA/C1 peptidases, high-resolution crystal structures are available. For all remaining peptidases, suitable templates for homology modeling had to be searched. All complete CA/C1 peptidases identified could easily be modeled in this way relying on known cathepsin structures. The shorter and incomplete sequences could also often be modeled on cathepsin crystal structures as templates, but sometimes alternative templates would have been possible (see results) indicating different or less defined structure for these shorter sequences. For homo logy models, suite Swiss-Model software (Guex and Peitsch, 1997) was applied to obtain predictions. Only *bona fide* structure templates (more than 50% similarity, more than 80% model coverage) were used. Structure predictions were also analyzed using AnDOM software (Schmidt et al., 2002) to identify and analyze structural domain composition, applying the Dali 3-D server (Holm and Sander, 1993) for comparison and PHD software (Rost et al., 2004) to analyze accessibility and secondary structure, with additional expert evaluation. Rasmol (Sayle and Milner-White, 1995) and PdbViewer (Guex and Peitsch, 1997) were used for visualization and inspection of structures (including labeling of residues specific for individual CA/C1 peptidases).

Localization signals were detected (Table S3) applying the LocTree support vector machines (Nair and Rost, 2005). Briefly, these predict the subcellular localization of proteins, and DNA-binding propensity for nuclear proteins, by incorporating a hierarchical ontology of localization classes modeled onto biological processing pathways. Biological similarities are incorporated from the description of cellular components provided by the gene ontology consortium (Nair and Rost, 2005).

Promoter analysis and regulatory RNA elements

Following analysis of the sequence, sequence phylogeny and structure of the known, as well as the newly identified CA/C1 peptidases, we looked at the regulation of the genome loci in more detail, both on the DNA and RNA level. In particular, we analyzed whether there would be conserved transcription factor binding sites and spacings, so-called transcription factor modules. Furthermore, we wanted to understand subfamily-specific expression differences. For these promoter analyses, genomatix suite software was applied. In this suite of different promotor analysis software, the ElDorado tool (Werner, 2003) examined gene associated promoter regions. Chip2Promoter and multiple alignments were applied to identify common transcription factor binding sites (TFBS). Consecutive TFBS which

were significantly conserved led to the definition of promoter modules. To identify these, upstream promoter sequences (transcription start until 500 base pairs upstream of it, looking at both strands) were scanned for common transcription factor binding sites. As default, an optimized matrix threshold (maximum 3 hits per 10 000 base pairs of non-regulatory test sequence) was applied, the alignment cut-off was always n-1/n where n represents the total number of aligned sequences. Including this user defined threshold, genomatix software calculates significant matches of modules and transcription factors as described by Werner (2003). Further details on the method are available at http://www.genomatix.de/online_help/help_matinspector/matinspector_alg.html.

mRNAs encoded by the CA/C1 peptidase gene loci were predicted according to genome information, and all EST data available and re-testing the results applying standard gene prediction software (Burge and Karlin, 1997). mRNA sequences were then screened and analyzed applying UTRscan software (Pesole and Liuni, 1999) and Transterm (Jacobs et al., 2006) to identify regulatory elements. Next, conservation of the identified regulatory RNA elements across species and in different subfamilies was compared.

Acknowledgments

We thank Deutsche Forschungsgemeinschaft (SFB544/B2, SFB630/C6, Da 208/10-1) and Bundesministerium für Bildung und Forschung (Hepatosys 03 130 74A, Funcrypta 0313838B) for support.

References

Akoachere, M., Buchholz, K., Fischer, E., Burhenne, J., Haefeli, W.E., Schirmer, R.H., and Becker, K. (2005). *In vitro* assessment of methylene blue on chloroquine-sensitive and -resistant *Plasmodium falciparum* strains reveals synergistic action with artemisinins. Antimicrob. Agents Chemother. *49*, 4592–4597.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L. (2002). GenBank. Nucleic Acids Res. *30*, 17–20.

Blackman, M.J. (2008). Malarial proteases and host cell egress: an 'emerging' cascade. Cell. Microbiol. *10*, 1925–1934.

Blake, J.A., Eppig, J.T., Richardson, J.E., Bult, C.J., and Kadin, J.A. (2001). The mouse genome database (MGD): integration nexus for the laboratory mouse. Nucleic Acids Res. *29*, 91–94.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. *31*, 365–370.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. *268*, 78–94.

Caglic, D., Kosec, G., Bojic, L., Reinheckel, T., Turk, V., and Turk, B. (2009). Murine and human cathepsin B exhibit similar properties: possible implications for drug discovery. Biol. Chem. *390*, 175–179.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

Chang, L.J. and Gay, E.E. (2001). The molecular genetics of lentiviral vectors – current and future perspectives. Curr. Gene Ther. *1*, 237–251.

Chen, N., Seiberg, M., and Lin, C.B. (2006). Cathepsin L2 levels inversely correlate with skin color. J. Invest. Dermatol. *126*, 2345–2347.

Choi, Y.S., Lee, H.K., and Pak, Y.K. (2002). Characterization of the 5′-flanking region of the rat gene for mitochondrial transcription factor A (Tfam). Biochim. Biophys. Acta *1574*, 200–204.

Dandekar, T., Huynen, M., Regula, J.T., Ueberle, B., Zimmermann, C.U., Andrade, M.A., Doerks, T., Sanchez-Pulido, L., Snel, B., Suyama, M., et al. (2000). Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. Nucleic Acids Res. *28*, 3278–3288.

Deloukas, P., Earthrowl, M.E., Grafham, D.V., Rubenfield, M., French, L., Steward, C.A., Sims, S.K., Jones, M.C., Searle, S., Scott, C., et al. (2004). The DNA sequence and comparative analysis of human chromosome 10. Nature *429*, 375–381.

Deussing, J., Kouadio, M., Rehman, S., Werber, I., Schwinde, A., and Peters, C. (2002). Identification and characterization of a dense cluster of placenta-specific cysteine peptidase genes and related genes on mouse chromosome 13. Genomics *79*, 225–240.

Drew, M.E., Banerjee, R., Uffman, E.W., Gilbertson, S., Rosenthal, P.J., and Goldberg, D.E. (2008). *Plasmodium* food vacuole plasmepsins are activated by falcipains. J. Biol. Chem. *283*, 12870–12876.

Dude, M.A., Kaeppler, U., Herb, M., Schiller, M., Schulz, F, Vedder, B., Heppner, S., Pradel, G., Gut, J., Rosenthal, P.J., et al. (2008). Synthesis and evaluation of non-peptidic cysteine protease inhibitors of *P. falciparum* derived from etacrynic acid. Molecules *14*, 19–35.

Galtier, N., Gouy, M., and Gautier, C. (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. *12*, 543–548.

Gaudermann, P., Vogl, I., Zientz, E., Silva, F.J., Moya, A., Gross, R., and Dandekar, T. (2006). Analysis of and function predictions for previously conserved hypothetical or putative proteins in *Blochmannia floridanus*. BMC Microbiol. *6*, 1.

Gemünd, C., Ramu, C., Altenberg-Greulich, B., and Gibson, T.J. (2001). Gene2EST: a BLAST2 server for searching expressed sequence tag (EST) databases with eukaryotic gene-sized queries. Nucleic Acids Res. *29*, 1272–1277.

Goh, L.L. and Sim, T.S. (2005). Characterization of amino acid variation at strategic positions in parasite and human proteases for selective inhibition of falcipains in *Plasmodium falciparum*. Biochem. Biophys. Res. Commun. *335*, 762–770.

Guex, N. and Peitsch, M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis *18*, 2714–2723.

Hall, B. and Milcarek, C. (1989). Sequence and polyadenylation site determination of the murine immunoglobulin gamma 2a membrane 3′ untranslated region. Mol. Immunol. *26*, 819–826.

Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. J. Mol. Biol. *233*, 123–138.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., et al. (2005). Ensembl 2005. Nucleic Acids Res. *33*, D447–D453.

Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P., and Bairoch, A. (2004). Recent improvements to the PROSITE database. Nucleic Acids Res. *32*, D134–D137.

Jacobs, G.H., Stockwell, P.A., Tate, W.P., and Brown, C.M. (2006). Transterm – extended search facilities and improved integration with other databases. Nucleic Acids Res. *34*, D37–D40.

Kerr, I.D., Lee, J.H., Pandey, K.C., Harrison, A., Sajid, M., Rosenthal, P.J., and Brinen, L.S. (2009). Structures of falcipain-2 and falcipain-3 bound to small molecule inhibitors: implications for substrate specificity. J. Med. Chem. *52*, 852–857.

Krasky, A., Rohwer, A., Schroeder, J., and Selzer, P.M. (2007). A combined bioinformatics and chemoinformatics approach for the development of new antiparasitic drugs. Genomics *89*, 36–43.

Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P., and Bork, P. (2004). SMART 4.0: towards genomic data integration. Nucleic Acids Res. *32*, D142–D144.

McCoubrie, J.E., Miller, S.K., Sargeant, T., Good, R.T., Hodder, A.N., Speed, T.P., de Koning-Ward, T.F., and Crabb, B.S. (2007). Evidence for a common role for the serine-type *Plasmodium falciparum* serine repeat antigen proteases: implications for vaccine and drug design. Infect. Immun. *75*, 5565–5574.

Miller, S.K., Good, R.T., Drew, D.R., Delorenzi, M., Sanders, P.R., Hodder, A.N., Speed, T.P., Cowman, A.F., de Koning-Ward, T.F., and Crabb, B.S. (2002). A subset of *Plasmodium falciparum* SERA genes are expressed and appear to play an important role in the erythrocytic cycle. J. Biol. Chem. *277*, 47524–47532.

Mohamed, M.M. and Sloane, B.F. (2006). Cysteine cathepsins: multifunctional enzymes in cancer. Nat. Rev. Cancer *6*, 764–775.

Nair, R. and Rost, B. (2005). Mimicking cellular sorting improves prediction of subcellular localization. J Mol Biol. *348*, 85–100.

Palermo, C. and Joyce, J.A. (2008). Cysteine cathepsin proteases as pharmacological targets in cancer. Trends Pharmacol. Sci. *29*, 22–28.

Pandey, K.C., Wang, S.X., Sijwali, P.S., Lau, L.A., McKerow, J.H., and Rosenthal, P.J. (2005). The *Plasmodium falciparum* cysteine protease falcipain-2 captures its substrate, hemoglobin, via a unique motif. Proc. Natl. Acad. Sci. USA *102*, 9138–9143.

Perrière, G. and Gouy, M. (1996). WWW-query: an on-line retrieval system for biological sequence banks. Biochimie *78*, 364–369.

Pesole, G. and Liuni, S. (1999). Internet resources for the functional analysis of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Trends Genet. *15*, 378.

Petermann, I., Mayer, C., Stypmann, J., Biniossek, M.L., Tobin, D.J., Engelen, M.A., Dandekar, T., Grune, T., Schild, L., Peters, C., et al. (2006). Lysosomal, cytoskeletal, and metabolic alterations in cardiomyopathy of cathepsin L knockout mice. FASEB J. *20*, 1266–1268.

Puente, X.S., Sanchez, L.M., Overall, C.M., and Lopez-Otin, C. (2003). Human and mouse proteases: a comparative genomic approach. Nat. Rev. Genet. *4*, 544–558.

Rawlings, N.D., O'Brien, E., and Barrett, A.J. (2002). MEROPS: the protease database. Nucleic Acids Res. *30*, 343–346.

Rossi, A., Deveraux, Q., Turk, B., and Sali, A. (2004). Comprehensive search for cysteine cathepsins in the human genome. Biol. Chem. *385*, 363–372.

Rost, B., Yachdav, G., and Liu, J. (2004). The PredictProtein server. Nucleic Acids Res. *32*, W321–W326.

Rupp, I., Bosse, R., Schirmeister, T., and Pradel, G. (2008). Effect of protease inhibitors on exflagellation in *Plasmodium falciparum*. Mol Biochem. Parasitol. *158*, 208–212.

Sayle, R.A. and Milner-White, E.J. (1995). RASMOL: biomolecular graphics for all. Trends Biochem. Sci. *20*, 374.

Schmidt, S., Bork, P., and Dandekar, T. (2002). A versatile structural domain analysis server using profile weight matrices. J. Chem. Inf. Comput. Sci. *42*, 405–407.

Selzer, P.M., Pingel, S., Hsieh, I., Ugele, B., Chan, V.J., Engel, J.C., Bogyo, M., Russell, D.G., Sakanari, J.A., and McKerrow, J.H. (1999). Cysteine protease inhibitors as chemotherapy: lessons from a parasite target. Proc. Natl. Acad. Sci. USA *96*, 11015–11022.

Sijwali, P.S., Koo, J., Singh, N., and Rosenthal, P.J. (2006). Gene disruptions demonstrate independent roles for the four falcipain cysteine proteases of *Plasmodium falciparum*. Mol. Biochem. Parasitol. *150*, 96–106.

Singh, N., Sijwali, P.S., Pandey, K.C., and Rosenthal, P.J. (2006). *Plasmodium falciparum*: biochemical characterization of the cysteine protease falcipain-2′. Exp. Parasitol. *112*, 187–192.

Sunil, S., Chauhan, V.S., and Malhotra, P. (2008). Distinct and stage specific nuclear factors regulate the expression of falcipains, *Plasmodium falciparum* cysteine proteases. BMC Mol Biol. *9*, 47.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.

Turk, V., Turk, B., Guncar, G., Turk, D., and Kos, J. (2002). Lysosomal cathepsins: structure, role in antigen processing and presentation, and cancer. Adv. Enzyme Regul. *42*, 285–303.

Turk, D., Turk, B., and Turk, V. (2003). Papain-like lysosomal cysteine proteases and their inhibitors: drug discovery targets? Biochem. Soc. Symp. *70*, 15–30.

Vasiljeva, O., Papazoglou, A., Kruger, A., Brodoefel, H., Korovin, M., Deussing, J., Augustin, N., Nielsen, B.S., Almholt, K., Bogyo, M., et al. (2006). Tumor cell-derived and macrophage-derived cathepsin B promotes progression and lung metastasis of mammary cancer. Cancer Res. *66*, 5242–5250.

Vasiljeva, O., Reinheckel, T., Peters, C., Turk, D., Turk, V., and Turk, B. (2007). Emerging roles of cysteine cathepsins in disease and their potential as drug targets. Curr. Pharm. Des. *13*, 387–403.

Werner, T. (2003). Promoters can contribute to the elucidation of protein function. Trends Biotechnol. *21*, 9–13.

Wiederanders, B., Kaulmann, G., and Schilling, K. (2003). Functions of propeptide parts in cysteine proteases. Curr. Protein Pept. Sci. *4*, 309–326.

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

Yeh, R.F., Lim, L.P., and Burge, C.B. (2001). Computational inference of homologous gene structures in the human genome. Genome Res. *11*, 803–816.

Received December 20, 2008; accepted June 17, 2009