Published in final edited form as:
Genetica. 2009 May ; 136(1): 189–209. doi:10.1007/s10709-008-9336-9.

Improved Prediction of Malaria Degradomes by Supervised Learning with SVM and Profile Kernel

Rui Kuang[1,¶,*], **Jianying Gu**[2,¶], **Hong Cai**[3], and **Yufeng Wang**[3,*]

[1]Department of Computer Science and Engineering, University of Minnesota, Twin Cities, Minneapolis, MN 55455, United States of America

[2]Department of Biology, College of Staten Island/City University of New York, Staten Island, New York 10314, United States of America

[3]Department of Biology, University of Texas at San Antonio, San Antonio, Texas 78249, United States of America

Abstract

The spread of drug resistance through malaria parasite populations calls for the development of new therapeutic strategies. However, the seemingly promising genomics-driven target identification paradigm is hampered by the weak annotation coverage. To identify potentially important yet uncharacterized proteins, we apply support vector machines using profile kernels, a supervised discriminative machine learning technique for remote homology detection, as a complement to the traditional alignment based algorithms. In this study, we focus on the prediction of proteases, which have long been considered attractive drug targets because of their indispensable roles in parasite development and infection. Our analysis demonstrates that an abundant and complex repertoire is conserved in five *Plasmodium* parasite species. Several putative proteases may be important components in networks that mediate cellular processes, including hemoglobin digestion, invasion, trafficking, cell cycle fate, and signal transduction. This catalog of proteases provides a short list of targets for functional characterization and rational inhibitor design.

Keywords

Malaria; protease; parasite; *Plasmodium*; SVM

1. Introduction

Malaria remains one of the most important life-threatening diseases. It afflicts approximately 300–500 million people a year, killing 1–2 million, mostly in the developing countries in the tropical or subtropical regions. The causative agents of malaria are a group of protozoan parasites in the genus *Plasmodium*. The rapid spread of the parasite populations resistant to the available antimalarial drugs underscores the pressing need for new drugs.

Genomics-based searches for new antimalarial targets hold considerable promise (Carlton et al. 2002; Gardner et al. 2002; Carlton 2003), but have been limited by a practical difficulty: our inability to assign a functional identity to a large fraction of the recognized open reading frames (ORFs) in the parasite genome. In the case of *Plasmodium falciparum* which causes

*To whom correspondence should be addressed. E-mail: E-mail: yufeng.wang@utsa.edu (YW), Phone: (210)458-6492, Fax: (210) 458-5658; E-mail: kuang@cs.umn.edu (RK), Phone: (612) 624-7820, Fax : (612) 625-0572.
¶These authors contributed equally to this work

the most severe form of malaria, about 60% of the 5,300 ORFs were annotated as "hypothetical" due to the lack of statistically significant sequence similarity to proteins with known function/structure (Gardner et al. 2002). An effective solution to circumvent this problem lies in the development of new algorithms that can capture subtle similarities between the unknown proteins and the annotated proteins in protein databases.

We propose to improve protease prediction among those uncharacterized *Plasmodium* proteins with a computational prediction approach that applies support vector machines (SVMs) using extended profile kernels for remote homology detection. SVMs are a family of machine learning algorithms for classification and regression problems (Vapnik 1998). A SVM classifier is a linear function that separates the training data into two classes and also maximizes the geometric margin between them in a feature space. Our binary classification problem is the classification of an uncharacterized protein sequence as a member or a non-member of a given protein family with a SVM classifier learned from the training proteins. The SVM-based classification of protein sequences uses negative sequences (proteins outside the protein family) as well as positive sequences (members of the protein family) to learn the difference between the two classes. This discriminative nature of SVMs distinguishes them from those alignment-based approaches that build models only with positive sequences (Karplus, Barrett, and Hughey 1998), and often results in better empirical classification performance. Another desirable property of SVM is that learning a SVM classifier only depends on the pairwise similarity between the examples; therefore, we can use any symmetric and positive-definite similarity functions, called kernels, to achieve better classification performance and faster computation. Recently, it has been shown that SVM-based kernel approaches are especially effective in remote homology detection (Jaakkola, Diekhans, and Haussler 2000; Liao and Noble 2003; Leslie et al. 2004; Kuang et al. 2005; Rangwala and Karypis 2005). Our previous work on profile kernels (Kuang et al. 2005) established the-state-of-the-art performance for remote homology detection. The profile kernel is a function that measures the similarity of two protein sequence profiles based on their representation in a high-dimensional vector space indexed by all *k*-mers (*k*-length subsequences of amino acids). We modify the original profile kernel, which is defined on a feature space indexed by subsequence of a fixed length, to include subsequences of length in a certain range as features. We found that the extended profile kernels achieve significant improvements in protein classifications of the SCOP benchmark dataset (Results are not shown) (Murzin et al. 1995).

In this proof of concept study, we attempt to combine powerful SVM classifiers and the traditional alignment based PSI-Blast algorithms to predict the protease complements (degradomes) in *Plasmodium*. The proteases were chosen because:

(1) they have been thought of as attractive drug targets. Firstly, proteases, the digestive enzymes that hydrolyze peptides, are essential for the parasite life cycle: for example, aspartic proteases (plasmepsins) (Coombs et al. 2001; Goldberg 2005; Ersmark, Samuelsson, and Hallberg 2006), cysteine proteases (falcipains) (Rosenthal et al. 2002; Rosenthal 2004) and metalloprotease (falcilysin) (Eggleson, Duffin, and Goldberg 1999; Murata and Goldberg 2003a; Murata and Goldberg 2003b) are actively involved in hemoglobin digestion for parasite nutrition; serine proteases (subtilisins) are important for red blood cell invasion (Withers-Martinez, Jean, and Blackman 2004); and, recently, proteases have been implicated in cell cycle progression and cell signaling (Baker, Wijetilaka, and Urban 2006; O'Donnell et al. 2006; Le Chat, Sinden, and Dessens 2007; Meslin et al. 2007). Secondly, it is feasible to design specific inhibitors for proteases if the mechanism of protease action is known or can be predicted. Various types of inhibitors have been shown to effectively block parasite growth or/and invasion (Sharma 2007). Emerging techniques in combinatorial high throughput screening and computational structured based drug design (SBDD) have made promising contributions to the recent progress in searching out and designing malarial protease inhibitors: combinatorial

libraries have been synthesized and screened for plasmepsins (Carroll et al. 1998; Haque et al. 1999; Kasam et al. 2007) and a group of inhibitors for falcipains has been identified as well (Li et al. 1996; Scheidt et al. 1998; Pandey et al. 2006). Thirdly, because of the remote evolutionary relatedness between the malaria parasite and the human host, the inhibitors designed based on malaria protease targets should have little or no adverse effect on the host.

(2) A large amount of relevant data is available for the protease family, which makes the application of kernel based machine learning feasible. Substantial knowledge has been accumulated and a specialized expert-curated database, MEROPS, is available for proteases; it includes a catalog of characterized and predicted proteases in over 3100 organisms (Rawlings et al. 2008).

Here we report a catalog of the proteases in five species of *Plasmodium*, including the two human malaria parasites *P. falciparum* and *P. vivax*, and the three parasites *P. yoelii yoelii*, *P. berghei*, and *P. chabaudi*, which serve as the rodent models. This catalog opens a new line of novel proteases or protease-regulated cellular processes for functional characterization.

2.Methods

(1) Data Preparation

The predicted ORFs of the five *Plasmodium* species were downloaded from the PlasmoDB database (http://www.plasmodb.org/, release 5.2). In this release, there are 5,411 ORFs in *P. falciparum* genome, 5,352 in *P. vivax* genome, 7,861 in *P. yoeli* genome, 12,235 in *P. berghei* genome, and 15,007 in *P. chabaudi* genome. A total of 47,499 known peptidase units and peptidase inhibitor units in the MEROPS database (http://merops.sanger.ac.uk/, release 7.4) were used as the target sequences for PSI-Blast search and SVM training.

In the PSI-Blast search using the unidentified ORFs against the MEROPS sequences, one-iteration and the default e-value threshold 0.0001 are chosen to avoid retrieving too many false positives. The training data for SVM remote homology classification are constructed from the MEROPS database and the annotated proteins in *P. falciparum*, *P. vivax* and *P. yoelii* genomes. In the MEROPS database peptidase units and peptidase inhibitors are organized into a hierarchy with three levels—clans, superfamilies and families from the root to the leaves. We randomly sampled 1,208 proteases from all the protease families with a sample size from each family proportional to the total number of proteases in the family. We combined the 1,208 selected proteases from MEROPS with the 91 known *P. falciparum* proteases, the 72 known *P. vivax* proteases and the 98 known *P. yoelii* proteases to form the positive training set. We manually selected 1,087 annotated *P. falciparum* proteins, 553 annotated *P. vivax* proteins and 507 annotated *P. yoelli* proteins that are clearly not functionally related to any protease as the negative set, under the assumption that the negative proteins from *Plasmodium* species will be more sensitive examples for detecting their remote homologs in the uncharacterized ORFs. The construction is designed to maximize the detection performance with comprehensive representation of the data, while keeping the data size tractable for learning by the careful selection of training examples.

For all the protein sequences in the training set and the ORFs, we computed the sequence profiles by searching against a non-redundant protein database using PSI-Blast with 5 iterations and the default e-value threshold 0.0001. The positional frequencies of amino acids in the profiles were smoothed using background frequencies. We used the smoothed emission probabilities in computing the profile kernels for SVM training.

(2) Support Vector Machines

Support vector machines are a family of machine learning algorithms for classification and regression problems (Vapnik 1998; Cristianini and Shawe-Taylor 2000). The SVM learning algorithm finds a linear classifier $f(x)=<w,x>+b$ ($w \epsilon R^n$, $b \epsilon R$) to discriminate examples between the positive and the negative classes with a "large margin". The learned linear classifier defines a decision boundary, the hyperplane $<w, x>+b=0$.A test example x will be classified as positive if $f(x)>0$,, negative otherwise. Empirically, most of the real datasets are not separable in a linear feature space for learning such a SVM. For these harder cases, a soft margin SVM (Cristianini and Shawe-Taylor 2000), which incorporates a trade-off between maximizing the geometric margin and minimizing margin violations on the training set, can be learned to handle the exceptions. One important property of the SVM learning problem is that in its dual optimization form, we can replace the inner product between $x, y, < x, y >$ by a kernel function $K(x, y)$; here, the kernel implicitly maps (possibly nonlinearly) the original input vector space to a feature space (or a Hilbert space) with some feature mapping Φ, i.e. the kernel K is defined with the mapping Φ and $K(x, y) =< \Phi(x), \Phi(y) >$. If Φ is a non-linear mapping from the original feature space, it will allow SVM to easily handle non-linear data by learning a linear classifier in the new feature space.

We used the publicly available SPIDER package (http://www.kyb.tuebingen.mpg.de/bs/people/spider/) to learn the binary classifiers in our experiments. Due to the computational cost of constructing the SVM classifiers, we only applied the SVM classification on three species *P. falciparum*, *P. vivax* and *P. yoelii*, which are of more interest in this study.

(3) Extended Profile Kernels

We chose to use profile kernels (Kuang et al. 2005) for SVM learning since they have been shown to be the state-of-the-art kernels for remote homology detection. Profile kernels are kernel functions for measuring the similarity between a pair of protein sequence profiles based on their representation in a high-dimensional feature space indexed by all k-mers (k-length subsequences of amino acids). For a sequence x and its sequence profile $P(x)$ (e.g. PSI-Blast profile), the positional mutation neighborhood at position j with threshold δ is defined to be the set of k-mers $\beta = b_1 b_2 \ldots b_k$ satisfying a likelihood inequality with respect to the corresponding block of the profile $P(x)$, as follows:

$$M_{(k,\delta)}(P(x[\,j+1:j+k\,]))=\{\beta=b_1 b_2 \ldots b_k : -\sum_{i=1}^{k} \log p_{j+1}(b_i) \le k\delta\}$$

Note that in the definition $p_{j+i}(b_i)$ denotes the emission probability of amino acid b_i at position $j+i$ in the profile $P(x)$. Let $\sum$ be the alphabet of amino acids, the profile feature mapping of profile kernels can be defined as $\Phi_{k,\delta}(P(x))=(\phi_\beta(P(x)))_{\beta \epsilon \sum^k}$, (where the dimension $\phi_\beta(P(x))$ is the number of occurrence of in the mutational neighborhood $M_{(k,\delta)}(P(x))$.

We extended the original profile kernels by considering a new feature space indexed by all subsequences of lengths in a range $[k_{min}, k_{max}]$, i.e. the feature space is indexed by all the k-mers with $k_{min} \le k \le k_{max}$. The assumption of this extension is that lengths of most meaningful subsequences (motifs) are within a certain range. By limiting the possible length of the subsequences, the new feature space can cover most of the motifs without involving mapping to a space of much higher dimensions. If we use the same threshold for computing the positional mutation neighborhoods of k-mers with $k_{min} \le k \le k_{max}$, the positional mutation neighborhood of the extended profiles kernel is simply an addition of all the profile kernels computed with the k-mers of length in $[k_{min}, k_{max}]$. Since profile kernels can be efficiently computed with a trie

data structure in linear time complexity in terms of input sequence length, the time complexity of computing the combined profile kernels is also linear in sequence length.

In our experiments, $k_{min} = 4$ and $k_{max} = 6$ are chosen as the range of the *k*-mers by a cross-validation on the SCOP bench mark dataset for remote homology detection (Kuang et al. 2005). The extended profile kernels are normalized, and the SVM parameters are chosen by the default setting as in the benchmark experiments described in (Kuang et al. 2005).

(4) Multiple alignment and phylogenetic analysis

Multiple alignments were generated using the T-coffee program (http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee_cgi/index.cgi) (Notredame, Higgins, and Heringa 2000), followed by manual inspection and editing. Graphic representations of the alignment and consensus sequences were deduced by the program BOXSHADE (http://www.ch.embnet.org/software/BOX_form.html). Phylogenetic trees were inferred by the neighbor-joining method using MEGA (http://www.megasoftware.net/) (Tamura et al. 2007). Unweighted Maximum Parsimony (as implemented in PAUP 4.0) and Maximum Likelihood (as implemented in PHYLIP) (Felsenstein 1981) were used to examine(Hall et al. 2005) the robustness of the inferred phylogeny. Bootstrap resampling with 1,000 pseudoreplicates was carried out to assess support for individual branches. Bootstrap values of < 50% were collapsed and treated as polytomies.

3. Results and Discussion

(1) Protease Prediction with PSI-Blast and PF-SVM

In our study, we applied both SVMs using profile kernels and PSI-Blast to identify the proteases in the three complete or near complete genomes of *P. falciparum*, *P. vivax*, and *P. yoelii yoelii*. For *P. berghei* and *P. chabaudi*, only PSI-Blast was used for three empirical reasons: (1) the sequencing of these two genomes is not complete yet; gene finding and annotation is still at an early stage; (2) very little is known about the proteolytic machinery in these genomes; (3) the numbers of the predicted ORFs (12,235 in *P. berghei* and 15,007 in *P. chabaudi* genome) in these genomes are relatively larger than those in the other three species due to the fragmented nature of the sequence data and incomplete annotation of these genomes (Hall et al. 2005). Thus, a much longer time is required for computing the extended profile kernels.

The positively classified ORFs by the PF-SVM and the ORFs with e-value less than 1E-5 in the PSI-Blast search were subjected to further analysis. The domain organization of the predicted proteases was revealed by Pfam search (Finn et al. 2008). To annotate each predicted protease, we used the known protease sequence or protease domain with the highest similarity as a reference. The catalytic type and protease family were predicted in accordance with the MEROPS classification system, and the enzyme was named in accordance with the SWISS-PROT peptidase nomenclature (http://www.expasy.ch/cgi-bin/lists?peptidas.txt) and the literature. A Gene Ontology (GO) analysis was performed to predict the biological function, cellular process, and cellular location of the putative proteases (Ashburner et al. 2000). For *P. falciparum*, mining of the published microarray and mass spectrometry proteomics data revealed the expression of the putative proteases at the mRNA and protein levels, respectively (Florens et al. 2002; Lasonder et al. 2002; Bozdech et al. 2003a; Bozdech et al. 2003b; Le Roch et al. 2003; Florens et al. 2004; Le Roch et al. 2004; Hall et al. 2005).

Among the candidates predicted by PSI-Blast and PF-SVM, we discovered 28 putative proteases in *P. falciparum*, 45 in *P. vivax* and 19 in *P. yoelii yoelii*, all of which were not reported as proteases in the MEROPS database (release 7.4). For the two less-studied genomes, our PSI-Blast search predicted 127 putative proteases in *P. berghei*, and 137 in *P. chabaudi*.

In Table 1 we report the new proteases that are discovered only by PSI-Blast or PF-SVM but not by both. Overall PSI-Blast identified more of the verified predictions because our major verification relies heavily on analyzing sequence motifs. Many predictions made by the PF-SVM are unknown cases without reliable supporting evidence. PF-SVM also discovered several candidates that were not detectable by PSI-Blast. For example, we identified one putative PPPDE protease (PFI0940c and its orthologs in other *Plasmodium* species). This novel protease family has a circularly permuted papain-like fold and was postulated to play a role in the deubiquitination pathway and cell cycle control (Iyer, Koonin, and Aravind 2004). We also predicted a putative zinc protease PF13_0260, which has a weak prosite motif that was missed by PSI-Blast detection. Another example is PF10_0317. It does not have a detectable peptidase domain, but it has a novel domain belonging to the Der1-like family (Pfam PF04511 with E=3.4e-17). The Der1 protein is thought to play an indispensable role in the degradation process associated with the endoplasmic reticulum (ER) (Knop et al. 1996). Although there is no direct evidence of its proteolytic activity, this family may be distantly related to the rhomboid protease family, indicating a function in cellular signaling.

The PF-SVM performs reasonably well in keeping the homologous candidates at the top of the rank list, although profile kernels measure the overall similarity between two sequences instead of relying on estimating the statistical significance of a good alignment. In Figure 1, we show the plotting of the number of detected true positives given a certain number of false positives (up to 50). This plotting of sensitivity and specificity is commonly used to measure classification performance of remote homology detection in benchmark experiments (Jaakkola, Diekhans, and Haussler 2000;Liao and Noble 2003;Leslie et al. 2004;Kuang et al. 2005;Rangwala and Karypis 2005). In the experiments with *P. falciparum* genome and *P. yoelii yoelii* genome, the PF-SVM is more sensitive in detecting true positives compared with PSI-Blast, while PSI-Blast performs better on the *P. vivax* genome. From the plots in Figure 1, it is clear that when few false positives are present in the predictions, PF-SVM significantly outperforms PSI-Blast by ranking more true positives at the top of the rank list. At a given threshold of 10 false positives, the PF-SVM detects 4 more proteases than PSI-Blast in the *P. falciparum* genome (12 vs 8), 1 more in the *P. vivax* genome (19 vs 18) and 6 more in the *P. yoelii yoelii* genome (9 vs 3). Overall, PF-SVM performs better on the *P. falciparum* genome than on the other two genomes compared with PSI-Blast. We postulate that this difference might be related to the validation criteria for evaluating the predictions. In our analysis, the false positives are putative and many of them are unknown cases that cannot be fully determined with enough supporting evidences. This lack of evidence is a more severe problem for evaluating the predictions of PF-SVM since unlike PSI-Blast, PF-SVM does not provide any sequence alignment for the analysis, and many more predictions of PF-SVM are possibly unknown cases. Thus, the plots are just one empirical measure and they might not truly reflect the performance of PF-SVM compared against PSI-Blast. Furthermore, the *P. falciparum* genome has been relatively well studied. Presumably the predictions on this genome have relatively more supporting evidences, compared with those on the *P. vivax* genome and the *P. yoelii yoelii* genome.

The PF-SVM missed 20 putative proteases with good alignment (with e-value less than 1E-20). Thirteen of the missed candidates fall into four MEROPS families, C14 (caspase family), C50 (separase family), C54 (Aut2 peptidase family) and C65 (otubain-1 family). To test if this resulted from insufficient sampling of the MEROPS sequences - the training sequences sampled from these four families do not represent the sequence diversity in the family well - we constructed a larger training set by pulling in all the 436 sequences in the four families as additional positive training sequences. We found that several missed proteases were promoted to the top of the PF-SVM prediction lists. However, this change also introduced more false positives, and the overall ranking deteriorated.

(2) Why might PF-SVM be better for remote homology detection?

There are two reasons why PF-SVM may outperform PSI-Blast. Firstly, PF-SVM is not misled by widely shared structural motifs. For example, we found that a disproportionate number of the false positive PSI-Blast predictions fell into the S9 and S33 protease families. This is largely due to the presence of an alpha/beta fold in their peptidase unit. This alpha/beta fold structure is commonly shared with a large number of hydrolytic enzymes including the S9 and S33 proteases and other non-protease hydrolases with broad substrate specificity. These enzymes are believed to derive from a common ancestor with the basic arrangement of the catalytic residues. The false positive hits from PSI-Blast searches included a number of lipases that have that typical alpha/beta fold. By contrast, these proteins did not appear at the top of the rank list in PF-SVM ranking, since even if there is a match of alpha/beta folds in S9 or S33 in the positive training sequences, they are also present in the negative training sequences such as lipases, and thus, features describing these domains are assigned relatively low importance in protease classification.

Secondly, PF-SVM does not suffer from the so-called “profile-drift” problem: the incorporation of the additional weakly matched sequences dilutes the signal in the original sequence. In applying PSI-Blast, we used both single iteration search and five iteration searches to generate predictions. Most of the verified predictions were not highly ranked due to a large number of false positives that were introduced by the iterative PSI-Blast search. Thus, we carefully analyzed only the predictions produced by the single iteration PSI-Blast. This is probably a specific case of the profile drifting problem in PSI-Blast. Instead of relying on estimating the statistical significance of a particular alignment, profile kernels measure the overall similarity between two sequence profiles, and thus are more robust in preserving the original sequence signal while evolutionary information is introduced in the profile for effective remote homology detection.

(3) The degradome distributions in malaria parasites

The degradome complements of two human malaria parasites (*P. falciparum* and *P. vivax*) and three rodent parasites (*P. yoelii yoelii*, *P. berghei*, and *P. chabaudi*) have been revealed by SVM-based remote homology detection combining conventional PSI-Blast homology search. The proteolytic repertoire of *Plasmodium* consists of about 115–137 predicted proteins of 5 catalytic classes (aspartic, cysteine, metallo, serine and threonine). They can be further classified into 37 families according to the MEROPS protease nomenclature, which is based on intrinsic evolutionary and structural relationships (Rawlings et al. 2008) (Table 2 and Table 3). The detailed predicted characteristics of the proteases are summarized in Supplementary Table 1–Supplementary Table 5 (URL: http://compbio.cs.umn.edu/Protease_Class/). The fractions of proteases relative to predicted proteome complexity vary from 0.9% to 2.3% in five *Plasmodium* species: the human parasites appear to have relatively more abundant proteases than their rodent kin. The overall protease fraction in *Plasmodium* is similar to that in the 363 organisms with completed genomes that have been sequenced and annotated (2.9%) (Southan 2001; Puente et al. 2005; Rawlings et al. 2008).

(4) The core degradome

Our results indicate that malaria parasites possess a core degradome structure consisting of twenty-nine families of proteases. This degradome may be common to all *Apicomplexan* parasites. The proteases in this set have been found to play diverse roles in metabolism, cell cycle regulation, invasion and infection (Table 2). These families fall into four of the most important catalytic classes of proteases, and we discuss them below.

Cysteine proteases—Cysteine proteases comprise about 30% of the degradome; the two most prominent families from this class are the papain family (C1) and the ubiquitin carboxyl-

terminal hydrolase 2 family (UCH2, C19) (Table 2). The papain family (C1) includes well-characterized members of the falcipains and Serine-Repeat Antigens (SERAs). The functions of falcipains range from hemoglobin digestion, erythrocyte rupture to erythrocyte invasion as indicated by protease inhibition assay (Rosenthal 2002;Shenai, Semenov, and Rosenthal 2002), biochemical characterization (Shenai et al. 2000;Sijwali et al. 2001), RNA interference (Malhotra et al. 2002;Mohmmed et al. 2003) and gene disruption knockout experiments (Sijwali and Rosenthal 2004;Sijwali et al. 2006)(See (Rosenthal 2004) for a review). SERAs are potential vaccine targets since their gene products are immunogenic, and at least one member of the SERA family, SERA-5 (PFB0340c) in *P. falciparum*, may have proteolytic activity (Hodder et al. 2003;McCoubrie et al. 2007). Recently, a *P. berghei* SERA (PB000649.01.0) was suggested to be a protease that functions at sporozoite egress from oocyst (Aly and Matuschewski 2005;Arisue et al. 2007). The UCH2 (C19) family is another highly expanded gene family. This feature has likely arisen from the large-scale gene duplication events, as evidenced by the preservation of multiple copies of threonine proteases (T1 family) in multiple proteasome α and β subunits, and the ubiquitin C-terminal hydrolase family (C12). Such a massive retention of duplicates reflects the crucial role of the ATP-dependent ubiquitin-proteasome system, which has been implicated in cell-cycle control and stress responses in parasite life cycle (Gantt et al. 1998). Another cysteine protease family that can be of critical importance for parasite cell cycle is the metacaspase family (C14). We found that multiple copies (2–4) of metacaspases are present in *Plasmodium*, and they have the histidine and cysteine residues that are predicted to form the typical catalytic dyad (Wu et al. 2003). These paralogs may play complementary functions in parasite development and apoptosis in *P. falciparum* and *P. berghei* (Le Chat, Sinden, and Dessens 2007;Meslin et al. 2007).

Metallo and serine proteases—Although metallo and serine proteases are also abundant in *Plasmodium*, very little is known about their biological functions. Eleven metalloproteases are conserved in *Plasmodium*. For example, falcilysin, which belongs to the pitrilysin family (M16), is thought to be involved in hemoglobin degradation in the food vacuole (Eggleson, Duffin, and Goldberg 1999; Goldberg 2005). Recently its potential role in the degradation of apicoplast targeting peptides has been explored (Ponpuak et al. 2007). Our analysis shows that at least one copy of a falcilysin ortholog is present in each of the five *Plasmodium* genomes; two copies are found in the two rodent parasites *P. berghei* and *P. chabaudi*, and at least five copies of the M16 paralogs are present. As with the metalloproteases, only one of the seven families of serine proteases that seem to be conserved in *Plasmodium*, the subtilisin family (S8), has been extensively studied as a potential new drug target due to its apparent role in parasite invasion and egress (Blackman et al. 1998; Barale et al. 1999; Hackett et al. 1999; Wu et al. 2003; Withers-Martinez, Jean, and Blackman 2004; Yeoh et al. 2007). We confirmed the existence of multiple paralogs of subtilisins in the *Plasmodium* genomes. Moreover, the S8 family has experienced an expansion to four copies in *P. vivax* and five copies in *P. berghei*.

Aspartic proteases—Two families of aspartic proteases are conserved in *Plasmodium*. Plasmepsin, the pepsin family (A1) in *P. falciparum*, has long thought to play important roles in hemoglobin digestion (Coombs et al. 2001; Goldberg 2005). We identified a large family of plasmepsins in the other *Plasmodium* species which supports the speculation that it is an ancient family that has undergone domain shuffling, possibly rounds of gene duplications, gene loss, and gene gain by lateral gene transfers (Jean et al. 2001). We identified a new family of presenilin in the aspartic clan (A22). It may be involved in regulated intermembrane proteolysis.

Threonine protease—One single proteasome family (T1) forms the threonine protease clan in *Plasmodium* and plays a central function in degrading damaged or unused proteins by proteolysis. Although the detailed pathways and the entities of the substrates remain unclear,

the core complex structure of protease subunits (seven α- and seven β- subunits) and regulatory subunits have been revealed by our previous comparative genomic analysis (Wu et al. 2003). Independent microarray expression assays have shown apparent co-expressed patterns of the predicted threonine proteases (Bozdech et al. 2003a; Le Roch et al. 2003; Wang and Wu 2004). A schematic map can be found at Dr. Hagai Ginsburg's Malaria Parasite Metabolic Pathway, (http://sites.huji.ac.il/malaria/maps/proteaUbiqpath.html). In addition, we identified two new threonine proteases in *P. falciparum*: a proteasome catalytic subunit 3 homolog (PF10_0111) and an ATP-dependent heat shock protease hslV (PFL1465c). Both proteins possess a characteristic domain for threonine protease (pfam PF00227) with high statistical support (E=5.1e-64 and E=1.6e-13, respectively). Their potential importance will be discussed in the next section.

(5) Potentially important under-characterized proteases

To date, the studies of malaria proteases as potential drug or vaccine targets have been mainly focused on a small number of proteases. Several newly discovered proteases could be worth functional characterization.

Threonine protease - proteasome catalytic subunit (PF10_0111): protein-protein interactions?—It is particularly interesting that PF10_0111 showed 15 possible protein-protein interactions in yeast two-hybrid assays (Suthram, Sittler, and Ideker 2005). Given the substantial evolutionary distance between the two species, their different life styles and the relatively high rate of false-positive predictions in such assays, caution must be used when using yeast to predict protein networks in *P. falciparum*. Nonetheless, there is a high likelihood that PF10_0111 is an active component in protein networks. The nature of the protein interaction network(s) awaits further experimentation since these 15 interacting proteins seems to span a variety of functional categories, including (1) a ubiquitin transferase that could be a component of the ubiquitin-proteasome conjugated proteolysis, (2) a translation elongation factor, (3) a ribosome protein L15, (4) a ribosomal protein L4/L1, (5) a CCAAT-box DNA binding protein, (6) a nucleosome assembly protein, (7) a merozoite surface protein, (8) an erythrocyte membrane protein, (9) and seven hypothetical proteins.

Threonine protease hslV PFL1465c: prokaryotic origin—The proteasome inhibitor lactacystin has been shown to block the cell growth and cell division in malaria parasites, suggesting the proteasome can be targeted for drug development (Gantt et al. 1998). Which components in proteasome should be targeted? Malaria parasites, which are a group of primordial eukaryotes, seem to have a mosaic proteasome structure: a catalytic core 20S complex that is typically found in eukaryotes and a structurally complex HslV that is typically found in eubacteria are simultaneously present. The core complex is less attractive from a drug development perspective since it is conserved in the eukaryote domain. For example, a number of α and β subunits of threonine proteases in *Plasmodium* show considerable homology to the human proteases, suggesting their inhibitors could have potential side effects. By contrast, inhibitors for the prokaryotic version of the proteasome are more feasible. We confirmed that a putative heat shock protein PFL1465c is a homolog of the hslV theronine protease. It has several desirable features: (1) it is expressed at the erythrocytic stage, especially at the schizont stage, as suggested by multiple microarray experiments (Bozdech et al. 2003a; Bozdech et al. 2003b; Le Roch et al. 2003) and RT-PCR (Ramasamy et al. 2007); (2) it is likely catalytically active. The recombinant protein showed threonine, chymotrypsin and peptidyl glutamyl peptide hydrolase activity and the active sites are conserved between *P. falciparum* and the template *E. coli* protein, as shown by homology modeling (Ramasamy et al. 2007); (3) it may be a soluble protein as shown by localization assays; (4) it is distantly related to the host, as shown by phylogenetic analysis (Figure 2); (5) it is feasible to develop inhibitors specific to PFL1465c. In fact, a small-molecule inhibitor Nip-Leu-Leu-LeuVS-Me has been developed

for general HslV proteases. It shows irreversible inhibition due to covalent modification of the catalytic threonine (Powers et al. 2002). It is possible that the inhibitors for malaria HslV could have none or low side effects as there is no human homolog.

Regulated intramembrane proteolysis (RIP)—The discovery of RIP overturned the traditional paradigm of cell signaling where receptors transmit signals across membrane via binding specific molecules or ions (Brown et al. 2000). In the RIP pathways, proteases are the central players that cleave receptors and then release the fragments, which become messengers for the downstream signaling process. We identified two families of proteases in *Plasmodium* that may conduct RIP using different structure motifs and mechanisms.

(a) Rhomboid proteases (S54) –potential roles in invasion?: Rhomboid is a serine protease that is involved in regulated intramembrane proteolysis. It is ubiquitously present in archaea, bacteria and eukaryotes (Urban, Schlieper, and Freeman 2002). It has been shown to be important for animal development by activating epidermal growth factor receptor (EGFR) signaling in *Drosophila melanogaster* (Urban, Lee, and Freeman 2001) and for mitochondrial morphology and remodeling in yeast and human (Herlan et al. 2003; McQuibban, Saurya, and Freeman 2003). The function of rhomboid protease in *Apicomplexa*, the phylum to which malaria parasites belong, was first revealed in *Toxoplasma gondii*: four rhomboids were shown to cleave surface MIC adhesions, which are essential for invasion (Brossier et al. 2005; Dowse et al. 2005); Dowse and Soldati (2005) proposed a uniform nomenclature for *Apicomplexan* rhomboids, which we adopt here. These authors detected 8 rhomboid-like proteins in *P. falciparum* and seven of these had homologs in *P. berghei*. More recently, reports showed that two of these malarial rhomboid proteases, PF11_0150 (PfROM1) and PFE0340c (PfROM4), could cleave multiple adhesions during invasion (Baker, Wijetilaka, and Urban 2006), and that PFE0340c (PfROM4) specifically mediated shedding of the erythrocyte-binding antigen (EBA-175) (O'Donnell et al. 2006).

Our analysis found that homologs of the rhomboids detected by Dowse and Soldati (2005) are also found in the three additional species we examined. Based on our phylogenetic analysis, there are from 5 to 8 homologs of rhomboid proteases present in the *Plasmodium* species. They can be divided into at least five clusters based on their sequence similarity, depending on the bootstrap values used to establish the groups: ROM1/2, ROM3, ROM4/5, and ROM6/7/9 appeared to be conserved in the *Apicomplexa* parasites, while ROM8/10 seemed to be *Plasmodium*-specific (Figure 3a). Note that the homologs we uncovered in *P. vivax*, *P. yoelii yoelii*, and *P. chabaudi* were not uniformly distributed among the five clusters; there are two rhomboids from *P. vivax* in ROM8/10 and no *P. chabaudi* homolog in ROM4/5. We also uncovered a second *P. berghei* homolog in ROM6/7/9. It remains unknown why the rhomboid family has been greatly expanded in *Plasmodium*. One possible evolutionary driver for such a lineage specific expansion is to meet the needs of parasite or parasite-host signaling: different rhomboids might modulate the proteolysis of substrates such as adhesions and dynamins with diverse structures.

All the predicted *Plasmodium* rhomboids have a typical rhomboid domain (PF01694). As clearly shown in the alignment (Figure 3b), seven of the eight rhomboids in *P. falciparum* possess a conserved dyad: a serine (S) and a histidine (H) in two separate transmembrane domains. This dyad is a characteristic of the active sites required for rhomboid catalytic function as revealed by the crystal structure of the GlpG protein, a rhomboid protease from *E. coli* (Wang, Zhang, and Ha 2006). The S-H dyad is missing in PFF0900c (PfROM10), which appears to be quite divergent from the other rhomboids (Figure 3a).

(b) Signal peptide peptidase (SPP, presenilin family A22): The second family of the proteases that may govern RIP in malaria parasites is the SPP or presenilin. The four human

homologs of this family have been under extensive investigation because their mutation is strongly associated with the early onset of Alzheimer's disease. SPP has also been implicated in a variety of developmental and physiological functions. We found only single copies in four *Plasmodium* species; the exception was *P. berghei* where two paralogous copies are found. The *P. chabaudi* SPP homolog is a 68-residue partial fragment. It is remarkable that the plasmodial SPPs have two invariant catalytic motifs that are believed to be active sites for this protease family: a Tyr-Asp (YD) motif in a transmembrane domain and a Gly-Leu-Gly-Asp (GLGD) motif in a downstream transmembrane domain (Figure 4). Recently, Nyborg et al. (Nyborg et al. 2006) showed that the *P. falciparum* SPP (PF14_0543), when cloned into a mammalian vector, was capable of cleaving a SPP substrate. Microarray experiments have shown that PF14_0543 is expressed during the erythrocyte stage; the mass-spectrometry proteomics assay also pinpointed its expression at the merozoite stage, which is critical for invasion. If the plasmodial SPPs are bona fide proteases, it would be intriguing to test whether the well-known adhesins are the potential substrates of SPP. Moreover, because a line of inhibitors and compound libraries targeting animal SPPs have already been established, it should be relatively straightforward to design inhibitors of the plasmodial SPP, making it a good potential antimalarial target.

(6) Unclassified proteases

We identified four protease homologs that do not fall into any typical protease clan classification: (1) U48 (prenyl protease 2 family). Very little is known about this protease family, the majority of which are hypothetical proteins in diverse species from all the domains. The membrane-bound, prenyl protease is a new member of the *Plasmodium* degradome, which may be involved in secretion and protein modification. (2) A new signal peptidase. We previously predicted the two signal peptidases in *P. falciparum*, both belonging to the S26 family, which resemble the bacterial signal peptidase I and the eukaryotic mitochondrial 21KD signal peptidase (Wu et al. 2003). The new putative protease resembles the signal peptidase complex SPC22 unit in yeast and mammals. Apparently, the signal peptide processing machinery in *Plasmodium* is a mosaic of prokaryotic and eukaryotic types. The plasmodial SPC22 may have an important function, as the yeast SPC22 is essential for processing newly synthesized secreted proteins. (3) The PPPDE protease. This novel protease family has a circularly permuted papain-like fold and may function in the deubiquitination pathway and cell cycle control (Iyer, Koonin, and Aravind 2004). (4) A putative zinc protease that has a weak prosite motif.

(7) Comparison of the degradome in parasitic protozoa *Plasmodium* and the free-living ciliate *Tetrahymena thermophila*

—We compared the *Plasmodium* degradomes with the degradome in the ciliate *T. thermophila* (Eisen et al. 2006), the fully sequenced free-living organism most closely related to the malaria parasites. Twenty-one protease families are present in both genomes. For example, the members in the ATP-dependent ubiquitin-proteasome system (proteases C12, C19, and T1) are well conserved. There are more abundant proteases in *T. thermophila*, including 19 protease families that seem to be unique to *T. thermophila*. Surprisingly, leishmanolysin (M8), which was originally identified in the kinetoplastid parasite *Leishmania major* (Gruszynski et al. 2003; LaCount et al. 2003), is not present in any *Plasmodium* species despite their close evolutionary relatedness. However, a huge number (48) of leishmanolysins are found in the free-living *T. thermophila*, including 15 members in a tandem array. It remains unclear why leishmanolysin are expanded in nonkinetoplastid eukaryotes. Similarly, the carboxypeptidase A (M14) family is expanded to 28 members in *T. thermophila*, while only one copy is present in *Plasmodium*; The carboxypeptidase Y (S10) family includes 25 members, while none is found in *Plasmodium*.

Seven protease families are unique to *Plasmodium*: The metacaspase family (C14), a prototype caspase that has been implicated in apoptosis-like signal transduction (Madeo et al. 2002); the rhomboid family (S54) that can be essential for regulated intramembrane proteolysis during invasion and parasite development; the otubain-1 family (C65) and the Poh1 peptidase family (M67) that includes the isopeptidases that release ubiquitin from polyubiquitin for recycling; the thimet oligopeptidase family (M3) that regulates the intracellular degradation of oligopeptides such as cleaved signal peptides, and degraded protein products; the S2P protease family (M50), which has been shown in mammals to be involved in transcriptional regulation by proteolysis of transcription regulators; and the ClpP endopeptidase family (S14) which is a component of the ClpXP and ClpAP complexes responsible for the degradation of nascent polypeptides whose synthesis is interrupted.

Conclusion

We explored an approach combining PSI-Blast search and supervised SVM learning using profile kernels (PF-SVM) for improving the prediction of malaria degradomes. The PF-SVM was proved to be able to identify new proteases that were not detectable by PSI-Blast. Furthermore, when we restricted the number of false positives to be small, the PF-SVM also achieves higher sensitivity and accuracy than PSI-Blast. Our approach captured a global picture of the degradome of the five malaria parasite genomes, and is readily extensible to the study of organisms with remote homology to known model systems. The addition of the degradomes from four other species of *Plasmodium* to the existing one for *P. falciparum* revealed the core degradome for this important group of parasite. Our study also extended the list of proteases in all the species examined, unveiling proteases that are known to play key roles in other organisms in regulation, protein processing and housekeeping.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Abbreviations

EGFR, epidermal growth factor receptor; ER, endoplasmic reticulum; PF-SVM, support vector machine using profile kernels; ORF, open reading frame; RIP, regulated intramembrane proteolysis; SBDD, structured based drug design; SERA, serine-repeat antigen; SPP, signal peptide peptidase; SVM, support vector machine.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. We thank PlasmoDB for providing an all-in-one portal for malaria genomic data. The project described is supported by grants 1SC1GM081068, 8SC1AI080579, and R21AI067543 from the National Institute of General Medical Sciences and National Institute of Allergy and Infectious Diseases to Y. Wang. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences, National Institute of Allergy and Infectious Diseases or the National Institutes of Health. YW is also supported by NIH grant G12RR013646, and San Antonio Area Foundation Biomedical Research Funds. RK is supported by Grant-in-Aid of Research, Artistry and Scholarship at University of Minnesota, and the Biomedical Informatics and Computational Biology Seed Grant for UM-Mayo-IBM Collaboration. JG is supported by PSC-CUNY 37 Research Award and Summer Research Award for faculty at College of Staten Island / CUNY.

References

Aly AS, Matuschewski K. A malarial cysteine protease is necessary for Plasmodium sporozoite egress from oocysts. J Exp Med 2005;202:225–230. [PubMed: 16027235]

Arisue N, Hirai M, Arai M, Matsuoka H, Horii T. Phylogeny and evolution of the SERA multigene family in the genus Plasmodium. J Mol Evol 2007;65:82–91. [PubMed: 17609844]

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–29. [PubMed: 10802651]

Baker RP, Wijetilaka R, Urban S. Two Plasmodium rhomboid proteases preferentially cleave different adhesins implicated in all invasive stages of malaria. PLoS Pathog 2006;2:e113. [PubMed: 17040128]

Barale JC, Blisnick T, Fujioka H, Alzari PM, Aikawa M, Braun-Breton C, Langsley G. Plasmodium falciparum subtilisin-like protease 2, a merozoite candidate for the merozoite surface protein 1–42 maturase. Proc Natl Acad Sci U S A 1999;96:6445–6450. [PubMed: 10339607]

Blackman MJ, Fujioka H, Stafford WH, Sajid M, Clough B, Fleck SL, Aikawa M, Grainger M, Hackett F. A subtilisin-like protein in secretory organelles of Plasmodium falciparum merozoites. J Biol Chem 1998;273:23398–23409. [PubMed: 9722575]

Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum. PLoS Biol 2003a;1:E5. [PubMed: 12929205]

Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL. Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biol 2003b;4:R9. [PubMed: 12620119]

Brossier F, Jewett TJ, Sibley LD, Urban S. A spatially localized rhomboid protease cleaves cell surface adhesins essential for invasion by Toxoplasma. Proc Natl Acad Sci U S A 2005;102:4146–4151. [PubMed: 15753289]

Brown MS, Ye J, Rawson RB, Goldstein JL. Regulated intramembrane proteolysis: a control mechanism conserved from bacteria to humans. Cell 2000;100:391–398. [PubMed: 10693756]

Carlton J. The Plasmodium vivax genome sequencing project. Trends Parasitol 2003;19:227–231. [PubMed: 12763429]

Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL, Peterson JD, Pop M, Kosack DS, Shumway MF, Bidwell SL, Shallom SJ, van Aken SE, Riedmuller SB, Feldblyum TV, Cho JK, Quackenbush J, Sedegah M, Shoaibi A, Cummings LM, Florens L, Yates JR, Raine JD, Sinden RE, Harris MA, Cunningham DA, Preiser PR, Bergman LW, Vaidya AB, Van Lin LH, Janse CJ, Waters AP, Smith HO, White OR, Salzberg SL, Venter JC, Fraser CM, Hoffman SL, Gardner MJ, Carucci DJ. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. Nature 2002;419:512–519. [PubMed: 12368865]

Carroll CD, Patel H, Johnson TO, Guo T, Orlowski M, He ZM, Cavallaro CL, Guo J, Oksman A, Gluzman IY, Connelly J, Chelsky D, Goldberg DE, Dolle RE. Identification of potent inhibitors of Plasmodium falciparum plasmepsin II from an encoded statine combinatorial library. Bioorg Med Chem Lett 1998;8:2315–2320. [PubMed: 9873534]

Coombs GH, Goldberg DE, Klemba M, Berry C, Kay J, Mottram JC. Aspartic proteases of Plasmodium falciparum and other parasitic protozoa as drug targets. Trends in Parasitology 2001;17:532–537. [PubMed: 11872398]

Cristianini, N.; Shawe-Taylor, J. Cambridge, UK: Cambridge University Press; 2000. An Introduction to Support Vector Machines.

Dowse TJ, Pascall JC, Brown KD, Soldati D. Apicomplexan rhomboids have a potential role in microneme protein cleavage during host cell invasion. Int J Parasitol 2005;35:747–756. [PubMed: 15913633]

Eggleson KK, Duffin KL, Goldberg DE. Identification and characterization of falcilysin, a metallopeptidase involved in hemoglobin catabolism within the malaria parasite Plasmodium falciparum. Journal of Biological Chemistry 1999;274:32411–32417. [PubMed: 10542284]

Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM, Tallon LJ, Delcher AL, Salzberg SL, Silva JC, Haas BJ, Majoros WH, Farzad M, Carlton JM, Smith RK Jr, Garg J, Pearlman RE, Karrer KM, Sun L, Manning G, Elde NC, Turkewitz AP, Asai DJ, Wilkes DE, Wang Y, Cai H, Collins K, Stewart BA, Lee SR, Wilamowska K, Weinberg Z, Ruzzo WL, Wloga D, Gaertig J, Frankel J, Tsao CC, Gorovsky MA, Keeling PJ, Waller RF, Patron NJ,

Cherry JM, Stover NA, Krieger CJ, del Toro C, Ryder HF, Williamson SC, Barbeau RA, Hamilton EP, Orias E. Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. PLoS Biol 2006;4:e286. [PubMed: 16933976]

Ersmark K, Samuelsson B, Hallberg A. Plasmepsins as potential targets for new antimalarial therapy. Med Res Rev 2006;26:626–666. [PubMed: 16838300]

Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 1981;17:368–376. [PubMed: 7288891]

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. The Pfam protein families database. Nucleic Acids Res 2008;36:D281–D288. [PubMed: 18039703]

Florens L, Liu X, Wang YF, Yang SG, Schwartz O, Peglar M, Carucci DJ, Yates JR, Wu YM. Proteomics approach reveals novel proteins on the surface of malaria-infected erythrocytes. Molecular and Biochemical Parasitology 2004;135:1–11. [PubMed: 15287581]

Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, Witney AA, Wolters D, Wu YM, Gardner MJ, Holder AA, Sinden RE, Yates JR, Carucci DJ. A proteomic view of the Plasmodium falciparum life cycle. Nature 2002;419:520–526. [PubMed: 12368866]

Gantt SM, Myung JM, Briones MRS, Li WD, Corey EJ, Omura S, Nussenzweig V, Sinnis P. Proteasome inhibitors block development of Plasmodium spp. Antimicrobial Agents and Chemotherapy 1998;42:2731–2738. [PubMed: 9756786]

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DMA, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 2002;419:498–511. [PubMed: 12368864]

Goldberg DE. Hemoglobin degradation. Curr Top Microbiol Immunol 2005;295:275–291. [PubMed: 16265895]

Gruszynski AE, DeMaster A, Hooper NM, Bangs JD. Surface coat remodeling during differentiation of Trypanosoma brucei. J Biol Chem 2003;278:24665–24672. [PubMed: 12716904]

Hackett F, Sajid M, Withers-Martinez C, Grainger M, Blackman MJ. PfSUB-2: a second subtilisin-like protein in Plasmodium falciparum merozoites. Mol Biochem Parasitol 1999;103:183–195. [PubMed: 10551362]

Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, Berriman M, Florens L, Janssen CS, Pain A, Christophides GK, James K, Rutherford K, Harris B, Harris D, Churcher C, Quail MA, Ormond D, Doggett J, Trueman HE, Mendoza J, Bidwell SL, Rajandream MA, Carucci DJ, Yates JR 3rd, Kafatos FC, Janse CJ, Barrell B, Turner CM, Waters AP, Sinden RE. A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. Science 2005;307:82–86. [PubMed: 15637271]

Haque TS, Skillman AG, Lee CE, Habashita H, Gluzman IY, Ewing TJ, Goldberg DE, Kuntz ID, Ellman JA. Potent, low-molecular-weight non-peptide inhibitors of malarial aspartyl protease plasmepsin II. J Med Chem 1999;42:1428–1440. [PubMed: 10212129]

Herlan M, Vogel F, Bornhovd C, Neupert W, Reichert AS. Processing of Mgm1 by the rhomboid-type protease Pcp1 is required for maintenance of mitochondrial morphology and of mitochondrial DNA. J Biol Chem 2003;278:27781–27788. [PubMed: 12707284]

Hodder AN, Drew DR, Epa VC, Delorenzi M, Bourgon R, Miller SK, Moritz RL, Frecklington DF, Simpson RJ, Speed TP, Pike RN, Crabb BS. Enzymic, phylogenetic, and structural characterization of the unusual papain-like protease domain of Plasmodium falciparum SERA5. J Biol Chem 2003;278:48169–48177. [PubMed: 13679369]

Iyer LM, Koonin EV, Aravind L. Novel predicted peptidases with a potential role in the ubiquitin signaling pathway. Cell Cycle 2004;3:1440–1450. [PubMed: 15483401]

Jaakkola T, Diekhans M, Haussler D. A discriminative framework for detecting remote protein homologies. J Comput Biol 2000;7:95–114. [PubMed: 10890390]

Jean L, Long M, Young J, Pery P, Tomley F. Aspartyl proteinase genes from apicomplexan parasites: evidence for evolution of the gene structure. Trends Parasitol 2001;17:491–498. [PubMed: 11587964]

Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856. [PubMed: 9927713]

Kasam V, Zimmermann M, Maass A, Schwichtenberg H, Wolf A, Jacq N, Breton V, Hofmann-Apitius M. Design of new plasmepsin inhibitors: a virtual high throughput screening approach on the EGEE grid. J Chem Inf Model 2007;47:1818–1828. [PubMed: 17727268]

Knop M, Finger A, Braun T, Hellmuth K, Wolf DH. Der1, a novel protein specifically required for endoplasmic reticulum degradation in yeast. Embo J 1996;15:753–763. [PubMed: 8631297]

Kuang R, Ie E, Wang K, Wang K, Siddiqi M, Freund Y, Leslie C. Profile-based string kernels for remote homology detection and motif extraction. J Bioinform Comput Biol 2005;3:527–550. [PubMed: 16108083]

LaCount DJ, Gruszynski AE, Grandgenett PM, Bangs JD, Donelson JE. Expression and function of the Trypanosoma brucei major surface protease (GP63) genes. J Biol Chem 2003;278:24658–24664. [PubMed: 12707278]

Lasonder E, Ishihama Y, Andersen JS, Vermunt AMW, Pain A, Sauerwein RW, Eling WMC, Hall N, Waters AP, Stunnenberg HG, Mann M. Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. Nature 2002;419:537–542. [PubMed: 12368870]

Le Chat L, Sinden RE, Dessens JT. The role of metacaspase 1 in Plasmodium berghei development and apoptosis. Mol Biochem Parasitol 2007;153:41–47. [PubMed: 17335919]

Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder AA, Carucci DJ, Yates JR 3rd, Winzeler EA. Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. Genome Res 2004;14:2308–2318. [PubMed: 15520293]

Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder AA, Batalov S, Carucci DJ, Winzeler EA. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 2003;301:1503–1508. [PubMed: 12893887]

Leslie CS, Eskin E, Cohen A, Weston J, Noble WS. Mismatch string kernels for discriminative protein classification. Bioinformatics 2004;20:467–476. [PubMed: 14990442]

Li R, Chen X, Gong B, Selzer PM, Li Z, Davidson E, Kurzban G, Miller RE, Nuzum EO, McKerrow JH, Fletterick RJ, Gillmor SA, Craik CS, Kuntz ID, Cohen FE, Kenyon GL. Structure-based design of parasitic protease inhibitors. Bioorg Med Chem 1996;4:1421–1427. [PubMed: 8894100]

Liao L, Noble WS. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. J Comput Biol 2003;10:857–868. [PubMed: 14980014]

Madeo F, Herker E, Maldener C, Wissing S, Lachelt S, Herian M, Fehr M, Lauber K, Sigrist SJ, Wesselborg S, Frohlich KU. A caspase-related protease regulates apoptosis in yeast. Molecular Cell 2002;9:911–917. [PubMed: 11983181]

Malhotra P, Dasaradhi PV, Kumar A, Mohmmed A, Agrawal N, Bhatnagar RK, Chauhan VS. Double-stranded RNA-mediated gene silencing of cysteine proteases (falcipain-1 and −2) of Plasmodium falciparum. Mol Microbiol 2002;45:1245–1254. [PubMed: 12207693]

McCoubrie JE, Miller SK, Sargeant T, Good RT, Hodder AN, Speed TP, de Koning-Ward TF, Crabb BS. Evidence for a common role for the serine-type Plasmodium falciparum serine repeat antigen proteases: implications for vaccine and drug design. Infect Immun 2007;75:5565–5574. [PubMed: 17893128]

McQuibban GA, Saurya S, Freeman M. Mitochondrial membrane remodelling regulated by a conserved rhomboid protease. Nature 2003;423:537–541. [PubMed: 12774122]

Meslin B, Barnadas C, Boni V, Latour C, Monbrison FDe, Kaiser K, Picot S. Features of apoptosis in Plasmodium falciparum erythrocytic stage through a putative role of PfMCA1 metacaspase-like protein. J Infect Dis 2007;195:1852–1859. [PubMed: 17492602]

Mohmmed A, Dasaradhi PV, Bhatnagar RK, Chauhan VS, Malhotra P. In vivo gene silencing in Plasmodium berghei--a mouse malaria model. Biochem Biophys Res Commun 2003;309:506–511. [PubMed: 12963018]

Murata CE, Goldberg DE. Plasmodium falciparum falcilysin: a metalloprotease with dual specificity. J Biol Chem 2003a;278:38022–38028. [PubMed: 12876284]

Murata CE, Goldberg DE. Plasmodium falciparum falcilysin: an unprocessed food vacuole enzyme. Mol Biochem Parasitol 2003b;129:123–126. [PubMed: 12798513]

Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540. [PubMed: 7723011]

Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302:205–217. [PubMed: 10964570]

Nyborg AC, Ladd TB, Jansen K, Kukar T, Golde TE. Intramembrane proteolytic cleavage by human signal peptide peptidase like 3 and malaria signal peptide peptidase. Faseb J 2006;20:1671–1679. [PubMed: 16873890]

O'Donnell RA, Hackett F, Howell SA, Treeck M, Struck N, Krnajski Z, Withers-Martinez C, Gilberger TW, Blackman MJ. Intramembrane proteolysis mediates shedding of a key adhesin during erythrocyte invasion by the malaria parasite. J Cell Biol 2006;174:1023–1033. [PubMed: 17000879]

Pandey KC, Singh N, Arastu-Kapur S, Bogyo M, Rosenthal PJ. Falstatin, a cysteine protease inhibitor of Plasmodium falciparum, facilitates erythrocyte invasion. PLoS Pathog 2006;2:e117. [PubMed: 17083274]

Ponpuak M, Klemba M, Park M, Gluzman IY, Lamppa GK, Goldberg DE. A role for falcilysin in transit peptide degradation in the Plasmodium falciparum apicoplast. Mol Microbiol 2007;63:314–334. [PubMed: 17074076]

Powers JC, Asgian JL, Ekici OD, James KE. Irreversible inhibitors of serine, cysteine, and threonine proteases. Chem Rev 2002;102:4639–4750. [PubMed: 12475205]

Puente XS, Gutierrez-Fernandez A, Ordonez GR, Hillier LW, Lopez-Otin C. Comparative genomic analysis of human and chimpanzee proteases. Genomics 2005;86:638–647. [PubMed: 16162398]

Ramasamy G, Gupta D, Mohmmed A, Chauhan VS. Characterization and localization of Plasmodium falciparum homolog of prokaryotic ClpQ/HslV protease. Mol Biochem Parasitol 2007;152:139–148. [PubMed: 17270290]

Rangwala H, Karypis G. Profile-based direct kernels for remote homology detection and fold recognition. Bioinformatics 2005;21:4239–4247. [PubMed: 16188929]

Rawlings ND, Morton FR, Kok CY, Kong J, Barrett AJ. MEROPS: the peptidase database. Nucleic Acids Res 2008;36:D320–D325. [PubMed: 17991683]

Rosenthal PJ. Hydrolysis of erythrocyte proteins by proteases of malaria parasites. Current Opinion in Hematology 2002;9:140–145. [PubMed: 11844998]

Rosenthal PJ. Cysteine proteases of malaria parasites. Int J Parasitol 2004;34:1489–1499. [PubMed: 15582526]

Rosenthal PJ, Sijwali PS, Singh A, Shenai BR. Cysteine proteases of malaria parasites: targets for chemotherapy. Curr Pharm Des 2002;8:1659–1672. [PubMed: 12132997]

Scheidt KA, Roush WR, McKerrow JH, Selzer PM, Hansell E, Rosenthal PJ. Structure-based design, synthesis and evaluation of conformationally constrained cysteine protease inhibitors. Bioorg Med Chem 1998;6:2477–2494. [PubMed: 9925304]

Sharma A. Malarial protease inhibitors: potential new chemotherapeutic agents. Curr Opin Investig Drugs 2007;8:642–652.

Shenai BR, Semenov AV, Rosenthal PJ. Stage-specific antimalarial activity of cysteine protease inhibitors. Biol Chem 2002;383:843–847. [PubMed: 12108550]

Shenai BR, Sijwali PS, Singh A, Rosenthal PJ. Characterization of native and recombinant falcipain-2, a principal trophozoite cysteine protease and essential hemoglobinase of Plasmodium falciparum. Journal of Biological Chemistry 2000;275:29000–29010. [PubMed: 10887194]

Sijwali PS, Koo J, Singh N, Rosenthal PJ. Gene disruptions demonstrate independent roles for the four falcipain cysteine proteases of Plasmodium falciparum. Mol Biochem Parasitol 2006;150:96–106. [PubMed: 16890302]

Sijwali PS, Rosenthal PJ. Gene disruption confirms a critical role for the cysteine protease falcipain-2 in hemoglobin hydrolysis by Plasmodium falciparum. Proc Natl Acad Sci U S A 2004;101:4384–4389. [PubMed: 15070727]

Sijwali PS, Shenai BR, Gut J, Singh A, Rosenthal PJ. Expression and characterization of the Plasmodium falciparum haemoglobinase falcipain-3. Biochem J 2001;360:481–489. [PubMed: 11716777]

Southan C. A genomic perspective on human proteases. FEBS Lett 2001;498:214–218. [PubMed: 11412860]

Suthram S, Sittler T, Ideker T. The Plasmodium protein network diverges from those of other eukaryotes. Nature 2005;438:108–112. [PubMed: 16267557]

Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 2007;24:1596–1599. [PubMed: 17488738]

Urban S, Lee JR, Freeman M. Drosophila rhomboid-1 defines a family of putative intramembrane serine proteases. Cell 2001;107:173–182. [PubMed: 11672525]

Urban S, Schlieper D, Freeman M. Conservation of intramembrane proteolytic activity and substrate specificity in prokaryotic and eukaryotic rhomboids. Curr Biol 2002;12:1507–1512. [PubMed: 12225666]

Vapnik, VN. Adaptive and learning systems for signal processing, communications, and control. New York, NY: John Wiley & Sons; 1998. Statistical Learning Theory.

Wang Y, Wu Y. Computer assisted searches for drug targets with emphasis on malarial proteases and their inhibitors. Curr Drug Targets Infect Disord 2004;4:25–40. [PubMed: 15032632]

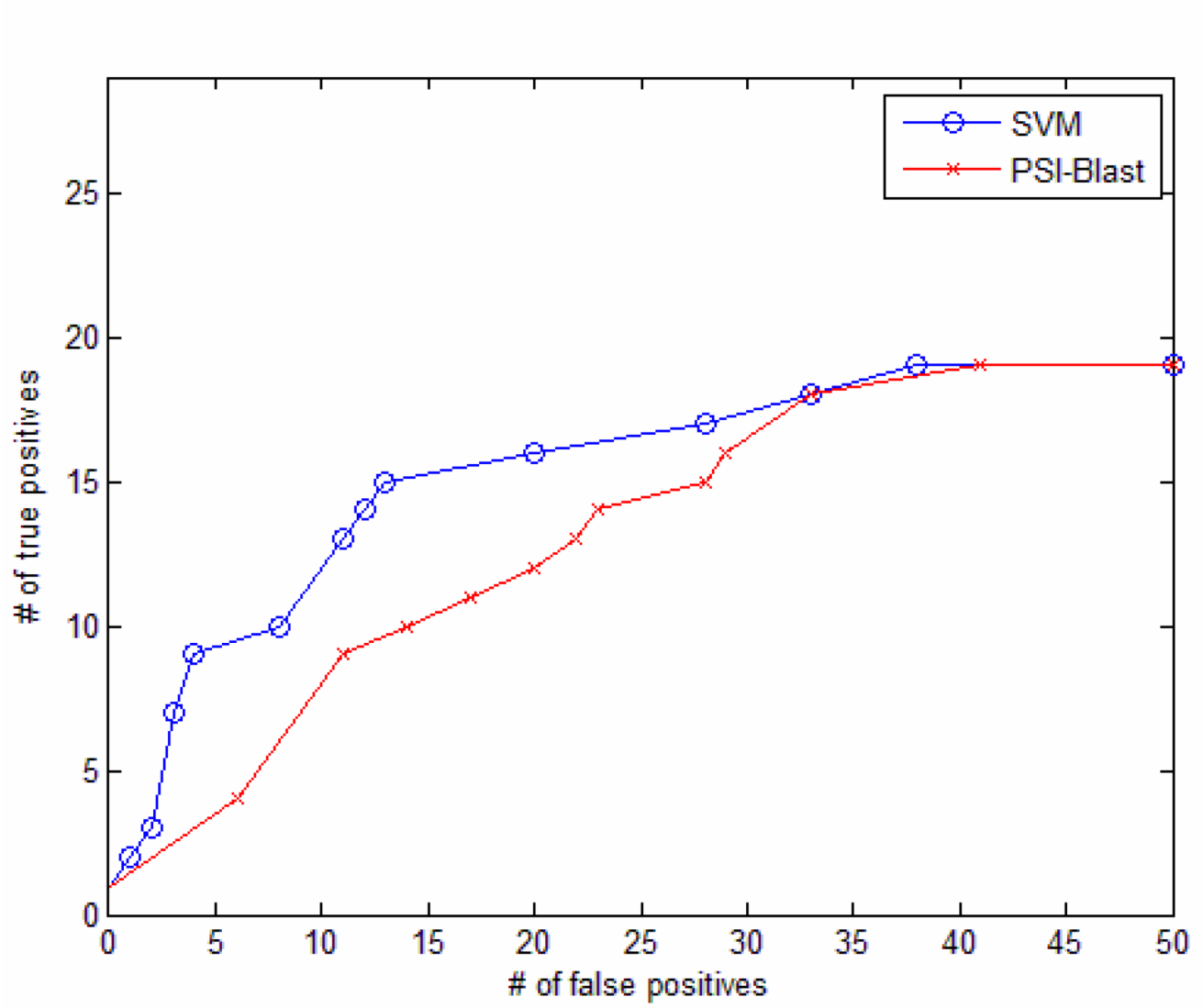
Wang Y, Zhang Y, Ha Y. Crystal structure of a rhomboid family intramembrane protease. Nature 2006;444:179–180. [PubMed: 17051161]

Withers-Martinez C, Jean L, Blackman MJ. Subtilisin-like proteases of the malaria parasite. Mol Microbiol 2004;53:55–63. [PubMed: 15225303]

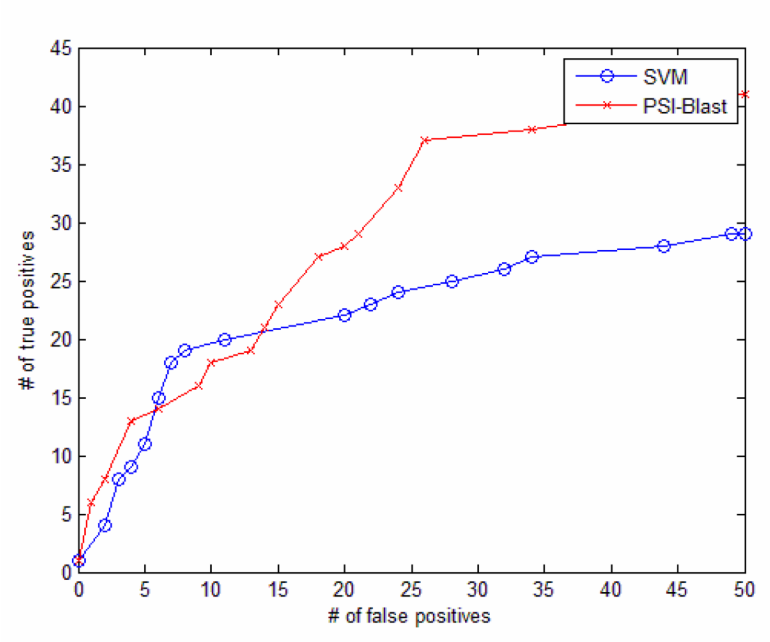
Wu YM, Wang XY, Liu X, Wang YF. Data-mining approaches reveal hidden families of proteases in the genome of malaria parasite. Genome Research 2003;13:601–616. [PubMed: 12671001]

Yeoh S, O'Donnell RA, Koussis K, Dluzewski AR, Ansell KH, Osborne SA, Hackett F, Withers-Martinez C, Mitchell GH, Bannister LH, Bryans JS, Kettleborough CA, Blackman MJ. Subcellular discharge of a serine protease mediates release of invasive malaria parasites from host erythrocytes. Cell 2007;131:1072–1083. [PubMed: 18083098]

(a) Prediction performance on *P. falciparum* genome



(b) Prediction performance on *P. vivax* genome



(c) Prediction performance in *P. yoelii yoelii* genome

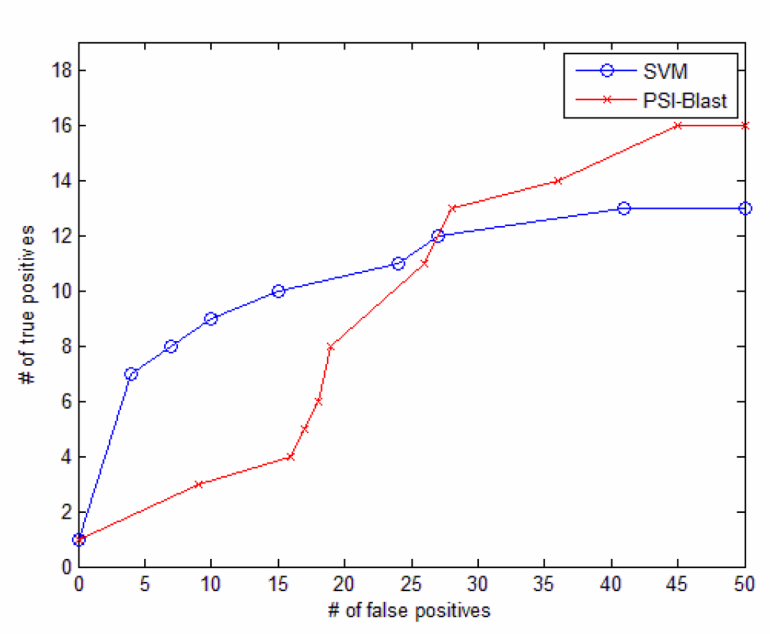


Figure 1. Performance comparison of SVM and PSI-Blast

(a) Prediction performance on *P. falciparum* genome. (b) Prediction performance on *P. vivax* genome. (c) Prediction performance on *P.yoelii yoelii* genome.

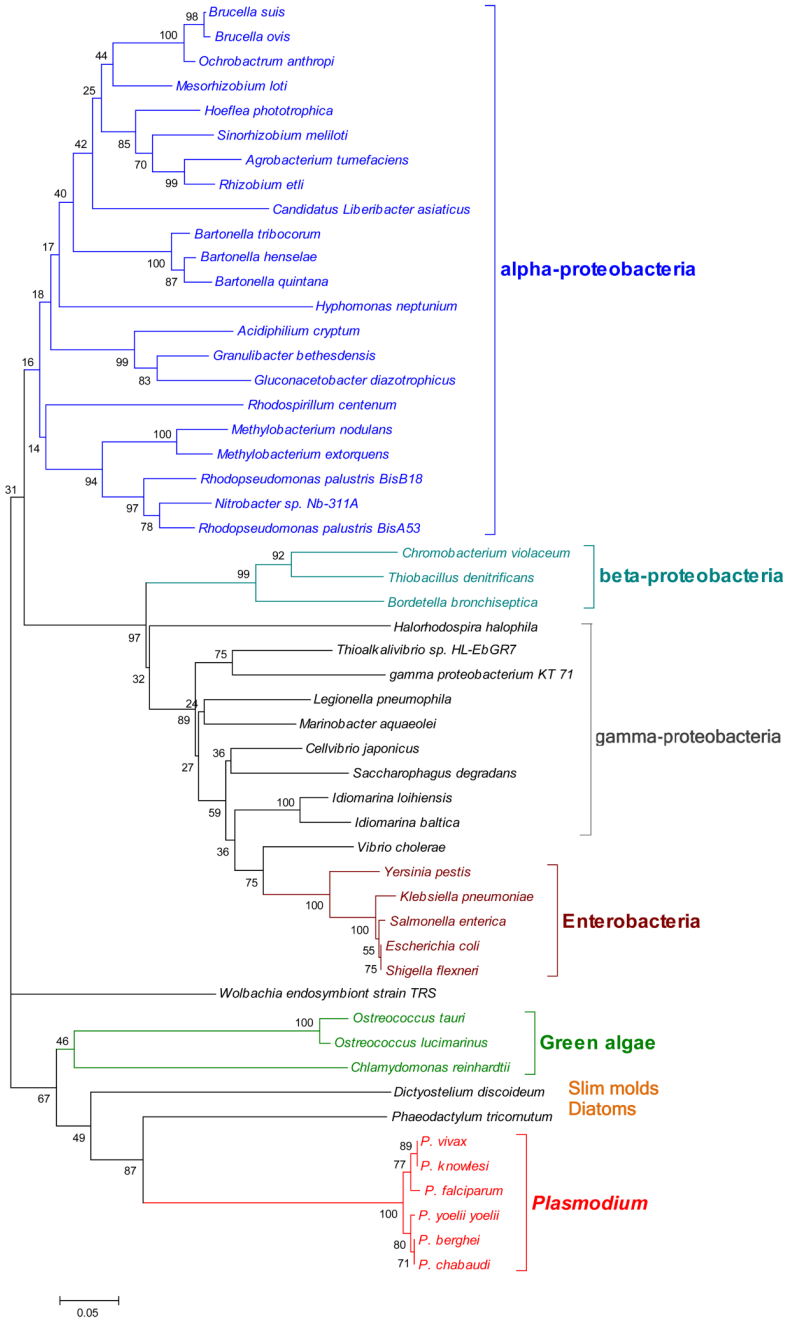
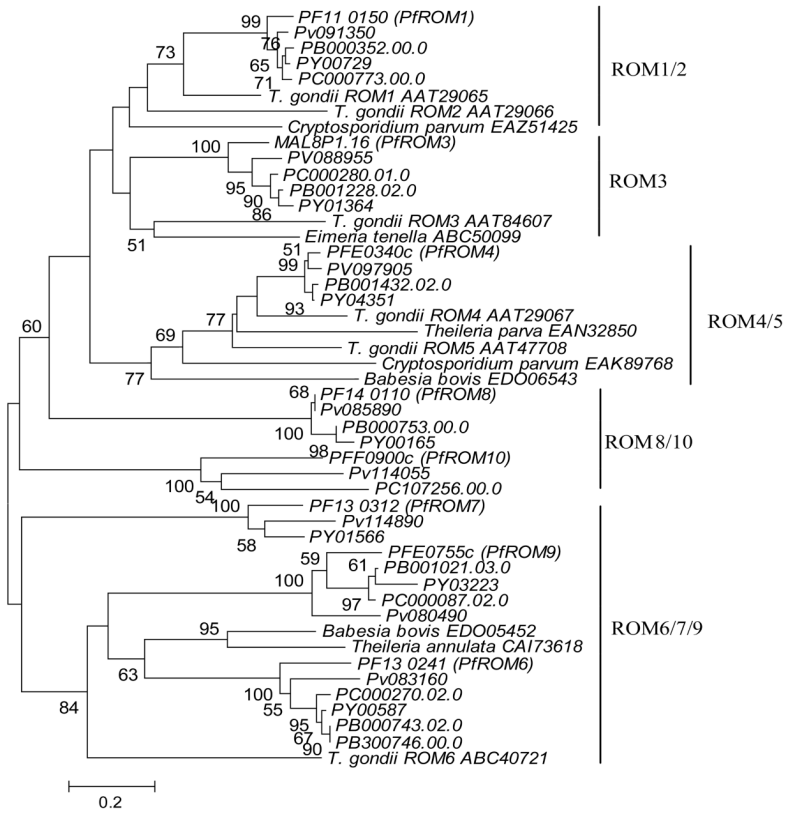


Figure 2. The phylogenetic tree of the HslV threonine proteases, inferred by the neighbor-joining method based on the amino acid sequences with Poisson corrected distance

The option of complete deletion of gaps was used for tree construction. 1000 bootstrap replicates were used to infer the reliability of branching points. The scale bar indicates the number of amino acid substitutions per site. The accession numbers are: NP_699055 (*Brucella suis* 1330), YP_001259885 (*Brucella ovis* ATCC 25840), YP_001369390 (*Ochrobactrum anthropi* ATCC 49188), NP_105747 (*Mesorhizobium loti* MAFF303099), ZP_02164574 (*Hoeflea phototrophica* DFL-43), NP_384162 (*Sinorhizobium meliloti* 1021), NP_353083 (*Agrobacterium tumefaciens* str. C58), YP_001976232 (*Rhizobium etli* CIAT 652), ZP_03287799 (*Candidatus Liberibacter asiaticus* str. psy62), YP_001608702 (*Bartonella*

tribocorum CIP 105476), YP_033060 (*Bartonella henselae* str. Houston-1), YP_031904 (*Bartonella quintana* str. Toulouse), YP_760731 (*Hyphomonas neptunium* ATCC 15444), YP_001234539 (*Acidiphilium cryptum* JF-5), YP_746057 (*Granulibacter bethesdensis* CGDNIH1), YP_001603516 (*Gluconacetobacter diazotrophicus* PAl 5), YP_002299467 (*Rhodospirillum centenum* SW), ZP_02118779 (*Methylobacterium nodulans* ORS 2060), YP_001638129 (*Methylobacterium extorquens* PA1), YP_530197 (*Rhodopseudomonas palustris* BisB18), YP_779310 (*Rhodopseudomonas palustris* BisA53), ZP_01045070 (*Nitrobacter* sp. Nb-311A), NP_900071 (*Chromobacterium violaceum* ATCC 12472), YP_316299 (*Thiobacillus denitrificans* ATCC 25259), NP_886730 (*Bordetella bronchiseptica* RB50), YP_001002780 (*Halorhodospira halophila* SL1), ZP_03278234 (*Thioalkalivibrio* sp. HL-EbGR7), ZP_01102923 (*gamma proteobacterium* KT 71), YP_094676 (*Legionella pneumophila* subsp. *pneumophila* str. Philadelphia 1), YP_958102 (*Marinobacter aquaeolei* VT8), YP_001980929 (*Cellvibrio japonicus* Ueda107), YP_528169 (*Saccharophagus degradans* 2–40), YP_156839 (*Idiomarina loihiensis* L2TR), ZP_01042581 (*Idiomarina baltica* OS145), NP_232303 (*Vibrio cholerae* O1 biovar eltor str. N16961), NP_667636 (*Yersinia pestis* KIM), YP_001337873 (*Klebsiella pneumoniae* subsp. pneumoniae MGH 78578), NP_457960 (*Salmonella enterica* subsp. enterica serovar Typhi str. CT18), ZP_02780625 (*Escherichia coli* O157:H7 str. EC4401), NP_709736 (*Shigella flexneri* 2a str. 301), YP_198552 (*Wolbachia* endosymbiont strain TRS of *Brugia malayi*), CAL54733 (*Ostreococcus tauri*), XP_001418801 (*Ostreococcus lucimarinus* CCE9901), XP_001692687 (*Chlamydomonas reinhardtii*), XP_645845 (*Dictyostelium discoideum AX4*), ACI65383 (*Phaeodactylum tricornutum* CCAP 1055/1), PFL1465c (*Plasmodium falciparum* 3D7), Pv124160 (*P. vivax* SaI-1), PY03772 (*P. yoelii yoelii* str. 17XNL), PB000649.02.0 (*P. berghei* strain ANKA), PC000270.03.0 (*P. chabaudi chabaudi*), CAQ42254 (*P. knowlesi* strain H).

(a)



(b)

```
MAL8P1.16  137 TTI----------------------------CNPCQ----LAVGASTSGFGLICCSI----FEIFLAWKNL-TKKAKNYYI
PF11_0150   81 SSS----------------------------ITYCP----LKVGASTSGMGLVGIVT----SELLLLWHILRHSERVVPNI
PF13_0241  386 YIL----------------------------LNKKN----QLLGASGSIMGLVYILSTLKPNEVFLSLFPLPYLKLTALQL
PF13_0312  229 SYYNHFKKKNEEIYLNDIIDQNYINKKIFNFKNNK----LICGSSSAIYALYG----MHITHVLFFYF----KNHY---V
PF14_0110  613 SSP----------------------------LTYCY----STTESSSSSSGTIGIFF----SELLMMTNFNVDKISISVHL
PFE0340c   513 SAV----------------------------CDPCG----VTLGSSGSLYGLIGALF----AYYLEYWKTI---PRPCCVL
PFF0900c   156 TCA----------------------------TSKCL----DVRNYLKYNENIL------------VHYYELIDKLS-IFGN
PFE0755c   392 QIL----------------------------YHKNNYNNIYWFGASGSISSLLSTYTFMYPSQNLYLYGVLALPLA-LFSS
consensus  721 s i                              c     livgastsi glig       seiil w  l  r    y i

MAL8P1.16  181 LNIFLFLL-----FFMFVSFSPSYDLFGHIGGLLCGAFLCCHY----------------------------------NK
PF11_0150  126 IFFSLISF-----FYYFTFNGSNLDHVGHLGGLISGISLGILY----------------------------------NE
PF13_0241  435 CHLSILTNIFFTLFFRKKKVNHFNIAWSAHMFGLLGGALYNIYQ----------------------------------R
PF13_0312  294 INTSFLYNFFYSFISS--LLLENVSHFNHIVGFLCGFVFSFLI----------------------------------IA
PF14_0110  658 FCFF----LLLFLK-FSLNTVSLNIYSHFFGFLGG--------------------------------------------
PFE0340c   555 IFMF----LVVMFGIIVGM-FGYTDNYAHIGGCLGGVLFGFSTITTVSAADKCTLGERMLVSAPFSWFLSNETKELIIAK
PFF0900c   192 FIFS----FLSLY-LMISLFSYNGNVLGNVGGILAGVSYPYI----------------------------------FKS
PFE0755c   444 LYF-----LNELYCV-LSNRQDNTGHIAHLTGMGLGLIYYI----------------------------------
consensus  801 l fs    i  lf    tv   nv  fghigGllgGily f
```

Figure 3. (a) The phylogenetic tree of the rhomboid protease homologs in Apicomplexa, inferred by the neighbor-joining method based on the amino acid sequences with Poisson corrected distance

The option of complete deletion of gaps was used for tree construction. 1000 bootstrap replicates were used to infer the reliability of branching points. The scale bar indicates the number of amino acid substitutions per site. The rhomboids in *Toxoplasma gondii* were used as the reference for the naming system ROM1-10 (Dowse and Soldati 2005). (b) The alignment of the rhomboid domain region of the eight ***P. falciparum*** homologs. The putative catalytic dyad Serine (S) and Histidine (H) are highlighted in red.

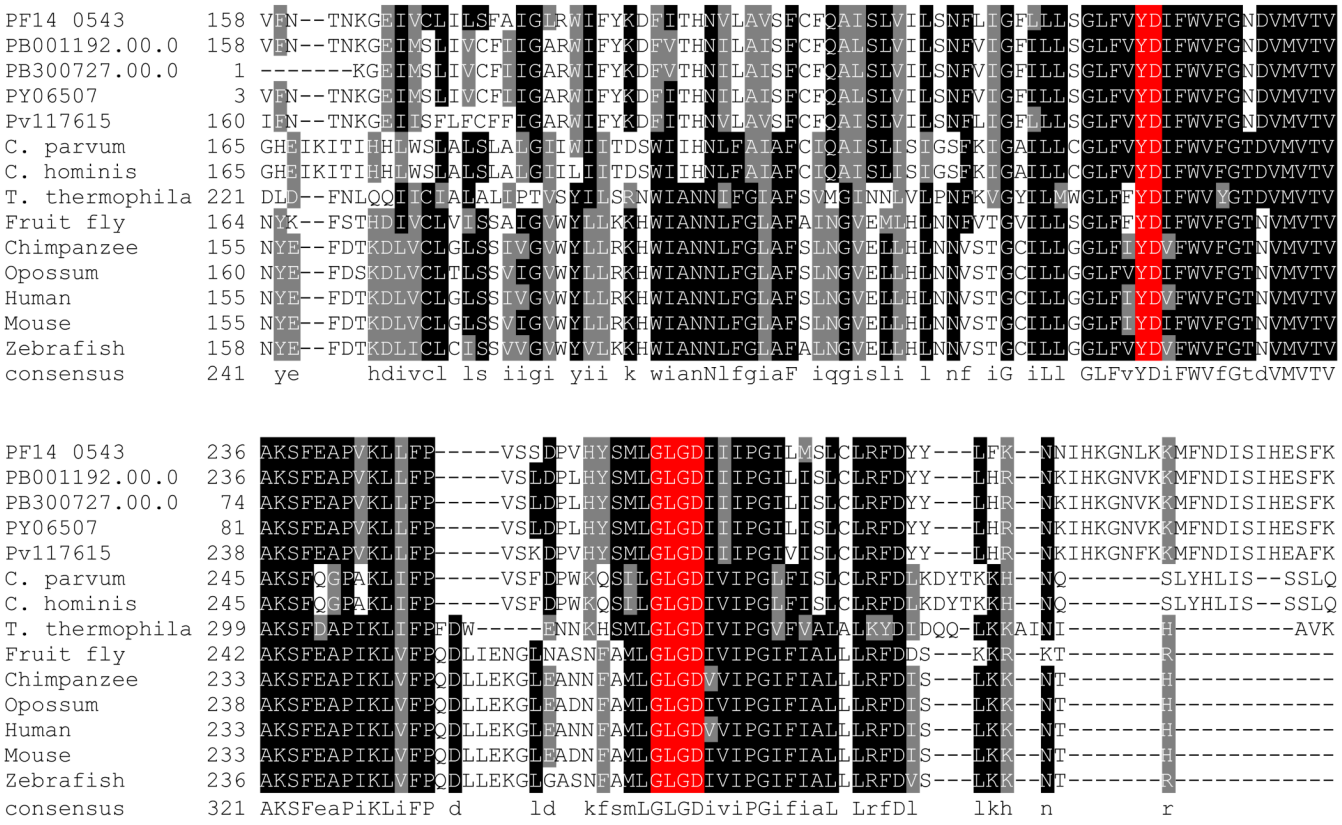


Figure 4. The alignment of the active site region of the signal peptide peptidases in representative species

The putative catalytic motifs Tye-Asp (YD) and Gly-Leu-Gly-Asp (GLGD) are highlighted in red.

Table 1

Newly identified putative proteases by SVM or PSI-Blast but not by both

	P. falciparum ORF (PSI-Blast e-value)	*P. vivax* ORF (PSI-Blast e-value)	*P. yoelii yoelii* ORF (PSI-Blast e-value)
PF-SVM	PFI0940c (0.017) PF13_0260 (1.3) PFI0215c (3.3) PF10_0317 (6)	Pv099375 (2.7)	PY04820 (2.4) PY03756 (5)
PSI-Blast	PF14_0363 (3E-15) PF14_0171 (2E-12) PFI0660c (7E-12) MAL8P1.113 (6E-10) PF14_0160 (0.000001) PF14_0692 (0.000009) PFI1135c (0.00003)	Pv085125 (2.00E-72) Pv081585 (6E-26) Pv085585 (2.00E-17) Pv080490 (6E-13) Pv123300 (8.00E-09) Pv111155 (3.00E-08) Pv084700 (4.00E-32) Pv093655 (1.00E-12) Pv085640 (0.000001)	PY05983 (2.00E-07) PY03056 (4.00E-16) PY00663 (2.00E-10) PY04718 (3.00E-09)

Table 2

Protease complements in *Plasmodium* species and other model organisms

Organism	Catalytic Class						Total	Percentage of the Proteome[a]
	Aspartic	Cysteine	Metallo	Serine	Threonine	Unclassified		
Plasmodium falciparum	12 (9.8%)[b]	39 (31.7%)	28 (22.8%)	25 (20.3%)	15 (12.2%)	4 (3.2%)	123	2.3
Plasmodium vivax	9 (7.7%)	41 (35.0%)	26 (22.2%)	24 (20.5%)	15 (12.8%)	2 (1.7%)	117	2.2
Plasmodium yoelii	12 (10.4%)	32 (27.8%)	29 (25.2%)	25 (21.7%)	15 (13.0%)	2 (1.7%)	115	1.5
Plasmodium bergheri	12 (9.8%)	39 (31.7%)	29 (23.6%)	24 (19.5%)	16 (13.0%)	3 (2.4%)	123	1.0
Plasmodium chabaudi	11 (8.0%)	48 (35.0%)	39 (28.5%)	18 (13.1%)	18 (13.1%)	3 (2.2%)	137	0.9
Tetrahymena thermophila[c]	43 (9.0%)	211 (44.0%)	139 (28.9%)	73 (15.2%)	14 (2.9%)	0 (0%)	480	1.7
Paramecium tetraurelia[c]	48 (8.3%)	225 (38.9%)	168 (29.1%)	95 (16.4%)	42 (7.3%)	0 (0%)	578	1.5
Neurospora crassa	13 (6.1%)	36 (16.7%)	71 (33.0%)	75 (34.9%)	20 (9.3%)	0 (0%)	215	2.2
Saccharomyces cerevisiae	14 (9.5%)	39 (26.4%)	51 (34.5%)	26 (17.6%)	17 (11.5%)	1 (0.6%)	148	2.4
Caenorhabditis elegans	27 (6.0%)	114 (25.3%)	180 (39.9%)	105 (23.3%)	24 (5.3%)	1 (0.2%)	451	2.2
Drosophila melanogaster	46 (6.6%)	80 (11.4%)	191 (27.2%)	351 (50.0%)	33 (4.7%)	1 (0.1%)	702	5.1
Homo sapiens	312 (31.6%)	167 (16.9%)	223 (22.6%)	247 (25.0%)	37 (3.8%)	1 (0.1%)	987	4.1
Arabidopsis thaliana	205 (24.8%)	147 (17.8%)	113 (13.7%)	317 (38.3%)	40 (4.9%)	5 (0.6%)	827	2.7
Escherichia coli	16 (5.8%)	35 (12.7%)	84 (30.5%)	120 (43.6%)	4 (1.5%)	16 (5.8%)	275	6.1

[a] The percentage of the whole genome that encodes putative proteases.

[b] Percentage of individual catalytic class in the protease complement is included in parentheses.

[c] The distributions of *T. thermophila and Paramecium tetraurelia* are based on Eisen et al. (2006) and unpublished data (Wang et al.) The distributions of the other model organisms are based on the results published in Merops database Release 7.60.

Table 3

Protease homologs in five *Plasmodium* genomes.

Catalytic Type	Protease Family	*P. falciparum*	*P. vivax*	*P. yoelii yoelii*	*P. berghei*	*P. chabaudi*
Aspartic	A1 (pepsin family)	PF14_0076 PF14_0077 PF14_0078 PF14_0075 PF13_0133 PFC0495w PF10_0329 PF14_0625 PF14_0281 PF08_0108 PFL1660c	Pv116695 Pv086040 Pv085030 Pv088125 Pv111035 Pv117180 Pv119690 Pv100670	PY06899 PY02085 PY02004 PY00469 PY06692 PY01268 PY01714 PY01716 PY03145 PY00470 PY06201	PB000298.03.0 PB000864.03.0 PB001294.02.0 PB301549.00.0 PB001202.00.0 PB001233.02.0 PB001439.02.0 PB301005.00.0 PB301520.00.0 PB000828.02.0	PC000654.04.0 PC000527.00.0 PC000734.01.0 PC000289.01.0 PC000244.02.0 PC300601.00.0 PC000243.00.0 PC000233.02.0 PC000173.02.0 PC301713.00.0
	A22 (presenilin family)	PF14_0543	Pv117615	PY06507	PB001192.00.0 PB300727.00.0	PC300096.00.0
Cysteine	C1 (papain family)	PF14_0553 PF11_0165 PF11_0162 PF11_0161 PFL2290w PFD0230c PF11_0174 PFI0135c PFB0360c PFB0325c PFB0330c PFB0335c PFB0340c PFB0345c PFB0350c PFB0355c	Pv091405 Pv091415 Pv117565 Pv003800 Pv003850 Pv003845 Pv003840 Pv003835 Pv003830 Pv091410 Pv001005 Pv003790 Pv003795 Pv003805 Pv003810 Pv003820 Pv003825 Pv091465 Pv101280	PY00783 PY00109 PY05365 PY02063 PY02062 PY00293 PY00292 PY00291 PY01568 PY02150 PY01068	PB000590.03.0 PB000980.00.0 PB000649.01.0 PB000676.00.0 PB107093.00.0 PB000108.03.0 PB000107.03.0 PB000352.01.0 PB301518.00.0 PB000233.03.0 PB000856.03.0 PB300133.00.0 PB000209.02.0 PB001079.01.0 PB301229.00.0 PB000888.02.0	PC302158.00.0 PC000458.02.0 PC000740.02.0 PC000637.04.0 PC001159.02.0 PC000780.04.0 PC000054.00.0 PC302520.00.0 PC000451.00.0 PC301446.00.0 PC301037.00.0 PC000068.00.0 PC000151.05.0 PC000253.01.0 PC302514.00.0 PC300826.00.0 PC000840.01.0 PC301973.00.0 PC000792.00.0 PC000617.03.0 PC000818.02.0
	C2 (calpain family)	MAL13P1.310	Pv115075	PY00976	PB001043.00.0	PC000118.05.0
	C12 (ubiquitin C-terminal hydrolase family)	PF14_0576 PF11_0177	Pv091485 Pv117435	PY01755 PY04400	PB000162.01.0 PB001566.02.0	PC000213.02.0 PC001172.02.0 PC000135.04.0
	C13 (legumain family)	PF11_0298	Pv092055	PY03470	N/A	N/A

Catalytic Type	Protease Family	*P. falciparum*	*P. vivax*	*P. yoelii yoelii*	*P. berghei*	*P. chabaudi*
	C14 (caspase family)	PF13_0289 PF14_0363 PF14_0160	Pv118575 Pv114725 Pv085640	PY00663 PY04718	PB001074.02.0 PB000485.03.2 PB000564.02.0 PB301215.00.0	PC000661.00.0 PC000188.05.0 PC000375.02.0
	C19 (ubiquitin-specific protease family)	PFA0220w PFA0220w PFD0165w PFD0680c PFE1355c PFE0835w MAL7P1.147 PFI0225w PF13_0096 PF14_0145 PFD0655c	Pv001075 Pv079880 Pv080415 Pv081540 Pv081630 Pv089655 Pv098675 Pv122670 Pv085715	PY04608 PY03410 PY02443 PY01440 PY00546 PY03738 PY05772 PY03802 PY01242	PB001276.02.0 PB001352.02.0 PB000920.01.0 PB000462.01.0 PB000975.01.0 PB300888.00.0 PB001543.02.0 PB000788.02.0 PB301209.00.0 PB001171.00.0 PB000494.03.0 PB000468.01.0	PC000383.03.0 PC000091.04.0 PC000330.04.0 PC001064.02.0 PC000088.01.0 PC001431.02.0 PC000952.01.0 PC000386.05.0 PC300262.00.0 PC000240.01.0 PC301801.00.0 PC301887.00.0 PC000558.00.0 PC000736.03.0 PC300411.00.0
	C48 (Ulp1endopeptidase family)	PFL1635w MAL8P1.157	Pv100650 Pv093655	PY03464 PY02388	PB000399.03.0 PB000598.02.0	PC000686.02.0 PC302567.00.0 PC301919.00.0 PC000942.01.0
	C50 (separase family)	MAL8P1.113	Pv123300	PY06348	PB000196.03.0	PC000026.00.0
	C54 (Aut2 peptidase family)	PF14_0171	Pv085585	PY03056	N/A	N/A
	C56 (PfpI endopeptidase family)	MAL6P1.153	N/A	PY04638	N/A	N/A
	C65 (otubain-1 family)	PFI1135c	Pv111155	PY05983	PB000440.01.0	N/A
Metallo	M1 (aminopeptidase N family)	MAL13P1.56 PF14_0692	Pv122425	PY01557	PB000843.02.0	PC001408.02.0 PC302364.00.0
	M3 (thimet oligopeptidase)	PF10_0058 MAL13P1.184	Pv082780 Pv094475	PY07695 PY06253 PY01285 PY03756	PB000279.03.0 PB301030.00.0	PC000493.04.0 PC001365.02.0

Catalytic Type	Protease Family	*P. falciparum*	*P. vivax*	*P. yoelii yoelii*	*P. berghei*	*P. chabaudi*
	M14 (carboxypeptidase A family)	PFA0170c	Pv081585	PY03811	PB001118.02.0	PC000329.04.0
	M16 (pitrilysin family)	PFE1155c PFI1625c PF11_0189 PF11_0226 PF13_0322 PF14_0382	Pv087035 Pv115000 Pv080095 Pv091550 Pv118475	PY01832 PY07032 PY00244 PY04232	PB001664.02.0 PB000738.02.0 PB300569.00.0 PB000990.01.0 PB000541.01.0	PC000562.01.0 PC300280.00.0 PC000079.01.0 PC000457.00.0 PC302182.00.0
	M17 (leucyl aminopeptidase family)	PF14_0439	Pv118180	PY07818 PY01898	PB000863.03.0	PC000418.00.0 PC000352.05.0 PC000869.01.0 PC300334.00.0 PC301661.00.0 PC000394.00.0
	M18 (aminopeptidase I)	PFI1570c	Pv087090	PY03205	PB000725.01.0 PB000622.00.0	PC000238.00.0
	M22 (O-sialoglycoprotein peptidase)	PF10_0299 PFD0440w	Pv111195 Pv000800	PY00526 PY00451	PB000774.03.0 PB300640.00.0 PB001212.00.0	PC000645.02.0 PC000543.03.0
	M24 (methionyl aminopeptidase 1)	PFE1360c MAL8P1.140 PF10_0150 PF14_0327 PF14_0517	Pv079875 Pv084805 Pv093540 Pv094985 Pv117760 Pv085125	PY04617 PY01653 PY02559 PY00802 PY05380 PY00855	PB000590.02.0 PB000961.00.0 PB001353.02.0 PB000628.00.0 PB000873.02.0 PB301269.00.0	PC000881.01.0 PC000583.01.0 PC000971.01.0 PC302079.00.0 PC000111.05.0 PC302144.00.0 PC301953.00.0 PC302381.00.0 PC000923.02.0 PC302399.00.0 PC300703.00.0
	M41 (FtsH endopeptidase family)	PF11_0203 PFL1925w PF14_0616	Pv100935 Pv117215 Pv091615	PY04402 PY05070 PY05838	PB000535.03.0 PB000162.00.0 PB001012.01.0 PB301353.00.0	PC000026.03.0 PC000187.04.0 PC000425.02.0 PC302098.00.0 PC300501.00.0
	M50 (S2P protease family)	PF13_0028 PF10_0317	Pv122115	PY00562	PB000986.02.0	PC301334.00.0
	M67 (Poh1 peptidase)	MAL13P1.343 PFI0895c PFI0630w	Pv115365 Pv099335 Pv099080	PY03078 PY05051 PY03442 PY02659	PB000245.01.0 PB000359.01.0 PB001445.02.0	PC001205.02.0 PC000291.02.0 PC301634.00.0

Catalytic Type	Protease Family	*P. falciparum*	*P. vivax*	*P. yoelii yoelii*	*P. berghei*	*P. chabaudi*
Serine	S1 (chymotrypsin family)	MAL8P1.126 PF14_0067	Pv088155 Pv123160	PY01797	N/A	N/A
	S8 (subtilisin family)	PFE0370c PF11_0381 PFE0355c	Pv092460 Pv097920 Pv097935 Pv097925	PY04329 PY01222	PB001288.02.0 PB300857.00.0 PB000701.02.0 PB000680.03.0 PB300552.00.0	PC001276.02.0 PC000778.00.0 PC000265.03.0
	S9 (prolyl oligopeptidase family)	PFC0950c	Pv095160	PY06542 PY02956 PY02448 PY02677	PB000494.00.0	N/A
	S14 (ClpP endopeptidase family)	PFC0310c PF14_0348 PF08_0063 PF14_0063 PF11_0175	Pv119490 Pv084700	PY00557 PY06630	PB001115.03.0 PB000565.01.0	PC001282.02.0 PC000189.00.0
	S16 (lon protease family)	PF14_0147	Pv086100 Pv089580 Pv091470 Pv085705	PY06406 PY06430 PY00565 PY05364 PY04458	PB000872.00.0 PB001190.00.0 PB300335.00.0 PB301282.00.0 PB301538.00.0	PC000511.02.0 PC000279.05.0 PC000118.01.0 PC000468.04.0 PC300948.00.0
	S26 (signal peptidase I family)	PF13_0118 MAL13P1.167	Pv122830 Pv082500	PY07139 PY00480	PB001244.00.0 PB001226.00.0	PC000159.03.0 PC000349.04.0
	S33 (prolyl aminopeptidase)	PFC0065C PF08_0022 PF14_0015	Pv089050	PY05572 PY04076	PB300951.00.0	PC000901.02.0
	S54 (Rhomboid family)	PFE0340c MAL8P1.16 PFE0755C PF11_0150 PF14_0110 PF13_0241 PFF0900c PF13_0312	Pv097905 Pv091350 Pv088955 Pv083160 Pv085890 Pv080490 Pv114055 Pv114890	PY04351 PY00729 PY01364 PY00165 PY03223 PY00587 PY01566	PB001432.02.0 PB000352.00.0 PB301437.00.0 PB001228.02.0 PB000743.02.0 PB300746.00.0 PB001021.03.0 PB000753.00.0	PC000773.00.0 PC000280.01.0 PC000270.02.0 PC000087.02.0 PC107256.00.0

Catalytic Type	Protease Family	*P. falciparum*	*P. vivax*	*P. yoelii yoelii*	*P. berghei*	*P. chabaudi*
Threonine	T1 (proteasome family)	PF14_0716 PFF0420c PFC0745c PF13_0282 PF07_0112 MAL8P1.128 MAL13P1.270 PFE0915c MAL8P1.142 PFA0400c PF14_0676 PFI1545c PF13_0156 PF10_0111 PFL1465c	Pv080330 Pv081375 Pv081675 Pv082355 Pv087115 Pv088170 Pv093555 Pv094790 Pv095380 Pv113585 Pv114680 Pv114685 Pv116925 Pv118620 Pv124160	PY04957 PY06767 PY03772 PY03034 PY02352 PY02094 PY00152 PY06176 PY06665 PY02351 PY02685 PY03212 PY04190 PY00267 PY00806	PB000867.01.0 PB001404.02.0 PB000672.03.0 PB000702.03.0 PB000705.00.0 PB001210.00.0 PB000393.02.0 PB001079.03.0 PB300285.00.0 PB000374.02.0 PB000621.00.0 PB000874.02.0 PB001452.02.0 PB000332.00.0 PB000776.01.0 PB000649.02.0	PC000988.01.0 PC301501.00.0 PC000739.02.0 PC302536.00.0 PC000536.01.0 PC000488.02.0 PC000358.02.0 PC300354.00.0 PC000491.02.0 PC000421.01.0 PC301073.00.0 PC000401.02.0 PC000270.04.0 PC000219.01.0 PC301900.00.0 PC000230.01.0 PC000737.02.0 PC000270.03.0
Unknown	U48 (prenyl protease 2 family)	PFI0660c	N/A	N/A	PB000372.01.0	PC000507.04.0
	Zinc protease	PF13_0260	N/A	N/A	N/A	N/A
	Signal peptidase	PFI0215c	Pv098665	PY04820	PB000789.02.0	PC000121.03.0
	PPPDE peptidase family	PFI0940c	Pv099375	PY05337	PB000018.00.0	PC000008.05.0