Molecular Medicine 7(10): 698–710, 2001

The Impact of Whole Genome Sequence Data on Drug Discovery—A Malaria Case Study

Marcin P. Joachimiak,[1,3] Calvin Chang,[2] Philip J. Rosenthal,[2] and Fred E. Cohen[*1,3]

[1]Graduate Group in Biophysics

[2]Department of Medicine, San Francisco General Hospital

[3]Department of Cellular and Molecular Pharmacology, University of California at San Francisco, San Francisco, CA, USA

Contributed by F. E. Cohen. Accepted September 5, 2001

Abstract

Background: Identification and validation of a drug discovery target is a prominent step in drug development. In the post-genomic era it is possible to reevaluate the association of a gene with a specific biological function to see if a homologous gene can subsume this role. This concept has special relevance to drug discovery in human infectious diseases, like malaria. A trophozoite cysteine protease (falcipain-1) from the papain family, thought to be responsible for the degradation of erythrocyte hemoglobin, has been considered a promising target for drug discovery efforts owing to the antimalarial activity of peptide based covalent cysteine protease inhibitors. This led to the development of non-peptidic non-covalent inhibitors of falcipain-1 and their characterization as antimalarials. It is now clear from sequencing efforts that the malaria genome contains more than one cysteine protease and that falcipain-1 is not the most important contributor to hemoglobin degradation. Rather, falcipain-2 and falcipain-3 appear to account for the majority of cysteine hemoglobinase activity in the plasmodium trophozoite.

Materials and Methods: We have modeled the falcipain-2 cysteine protease from one of the major human malaria species, *Plasmodium falciparum* and compared it to our original work on falcipain-1. As with falcipain-1, computational screening of the falcipain-2 active site was conducted using DOCK. Using structural superpositions within the protease family and evolutionary analysis of substrate specificity sites, we focused on the commonalities and the protein specific features to direct our drug discovery effort.

Results: Since 1993, the size of the Available Chemicals Directory had increased from 55313 to 195419 unique chemical structures. For falcipain-2, eight inhibitors were identified with IC_{50}'s against the enzyme between 1 and 7 μM. Application of three of these inhibitors to infected erythrocytes cured malaria in culture, but parasite death did not correlate with food vacuole abnormalities associated with the activity of mechanistic inhibitors of cysteine proteases like the epoxide E64.

Conclusions: Using plasmodial falcipain proteases, we show how a protein family perspective can influence target discovery and inhibitor design. We suspect that parallel drug discovery programs where a family of targets is considered, rather than serial programs built on a single therapeutic focus, will become the dominant industrial paradigm. Economies of scale in assay development and in compound synthesis are expected owing to the functional and structural features of individual family members. One of the remaining challenges in post-genomic drug discovery is that inhibitors of one target are likely to show some activity against other family members. This lack of specificity may lead to difficulties in functional assignments and target validation as well as a complex side effect profile.

Introduction

Cysteine proteases play a number of degradative and regulatory roles in a wide range of organisms. One measure of the success of this enzymatic motif is the degree of cysteine protease speciation. For example, the malaria genome is predicted to contain at least 5 cysteine proteases. This protein family is defined by a unique fold, which has speciated functionally many times producing subfamilies with unique substrate specificities. This proliferation of proteases creates the likelihood that more than one enzyme could subsume the same function *in vivo* and complicates the task of identifying the best targets for drug discovery. In the pre-genomic era, drug discovery targets were identified via a reductionist approach where genes were sought that carried out a physiological role. Further proof of principle was obtained using chemical inhibitors of the gene products. If the chemical inhibitor used had a broad specificity, the conclusions reached could be subject to question. In the post-genomic era, the genetic "deck of cards" is known and process of elimination logic can play a more prominent role in the identification of targets.

Given the success of angiotensin-converting enzyme (ACE) inhibitors in the treatment of hypertension (1) and HIV protease inhibitors in AIDS (2),

Address correspondence and reprint requests to: Fred E. Cohen, Department of Cellular and Molecular Pharmacology, UCSF, Box 0450, San Francisco, CA, 94143-0450. Fax: 415-476-6515; e-mail: cohen@cmpharm.ucsf.edu.

proteases have become popular drug discovery targets. However, several protease targets, such as the renin aspartyl protease for hypertension and matrix metalloproteases for cancer and arthritis, have not led to marketable products. These difficulties originated not from problems in the sequencing, cloning or annotation efforts but rather because of the redundant and homeostatic nature of biological systems, including the presence of genes performing back-up functions. The proteolytic cascade of the Renin Angiotensin Aldosterone (RAA) system mediates cleavage of angiotensinogen to angiotensin I by the aspartyl protease renin and subsequent cleavage of angiotensin I to the effector peptide angiotensin II by ACE. By the mid 1990's renin inhibitors were widely known to have negligible effects on hypertension (3), while to date dozens of ACE inhibitors have been proven to be effective human therapeutics for hypertension in spite of their side effects profile. Renin, the upstream enzyme in this pathway has a single unique substrate. While this molecular specificity would be expected to yield a better target for drug discovery efforts, compensatory homeostatic mechanisms undermine this thesis.

Like renin in humans, the plasmodial cysteine proteases that degrade hemoglobin exist as a family of homologs in the *P. falciparum* genome. As hemoglobin is the major nutritional source for the parasite in the erythrocytic stage, and proteases have been the target of successful drug discovery efforts, inhibitors of hemoglobin degradation have been sought as a new class of antimalarials. In 1987, Rosenthal et al. identified three *P. falciparum* proteases by gel electrophoresis. Two of these had an active site cysteine (4). A papain-like cysteine protease thought to be necessary for hemoglobin degradation in the trophozoite stage of the malaria human life cycle, and now known as falcipain-1, was cloned and sequenced (5). In 1993, a model of falcipain-1 based on its sequence homology to papain and actinidin was used in a structure-based drug discovery effort to identify a symmetric acylhydrazide inhibitor with antimalarial properties at a 6 μM concentration (6). However, optimization of the lead compound was complicated by difficulties in reconciling the activity of the lead analogs with the model protease structure. In the past year, *P. falciparum* genomic sequencing efforts led to the identification of a number of homologs of falcipain-1 and it now seems likely that the falcipain-2 and falcipain-3 gene products are the major plasmodial cysteine hemoglobinases (7).

During the erythrocytic phase of the life cycle, malaria parasites rely on hemoglobin degradation as the predominant source of amino acids. Interruption of hemoglobin degradation with mechanistic inhibitors of cysteine protease leads to accumulation of undigested hemoglobin, swelling of the food vacuole and parasite death (8). The precise order of events in the hemoglobin degradation pathway still remains to be clarified. In 1994, two aspartyl proteases, plasmepsins I (9) and II (10), were isolated from the *P. falciparum* food vacuole and shown to perform the first cleavage of hemoglobin. Recently plasmepsin II has been shown to cleave other erythrocyte proteins (11). Falcilysin, a plasmodial metallopeptidase, was reported to act against partially degraded hemoglobin fragments (12). However, our understanding of the pathway of hemoglobin hydrolysis remains limited, as falcipain-2 and falcipain-3 also readily hydrolyze native hemoglobin, while multiple plasmodial aspartic protease genes are predicted from the genome.

The antimalarial properties of peptide fluoromethyl-ketones and vinyl-sulphones as cysteine protease inhibitors (13,14) have encouraged their evaluation in animal models of infection. Unfortunately, activity against murine malaria required high doses and the toxicity of peptide fluoromethylketones in experimental animals has stalled their development (14). These results amplify our need to understand which proteases are most essential to hemoglobin degradation. Using modeling, drug design and inhibitor studies for the falcipain hemoglobinases, we illustrate how a gene family approach to drug targets can enhance the understanding of biological phenotype and its inhibition, and hence expedite the drug development process.

Results

Comparative Analysis of the Falcipain-2 and Falcipain-1 Model Structures

The original drug discovery effort directed at falcipain-1 led to structure-activity relationships that could not be reconciled with the protease model structure. Analogs of the acyl-hydrazide lead compound designed to take advantage of specific interactions in the protein's binding sites were synthesized. Unfortunately, these customized analogs did not lead to improvements in inhibitor affinity (15,16). In retrospect, we attribute this to the fact that the falcipain used in these assays was purified from parasite extract that is now known to be predominantly falcipain-2. Thus, the design was directed against falcipain-1 but the compounds were tested against predominantly falcipain-2. Inhibition studies are now conducted with recombinant falcipain-2.

Falcipain-2 and falcipain-1 share 37% sequence identity in the mature protease domain. The original model of falcipain-1 was based on papain and actinidin crystal structures (6), and both of the templates were 33% identical to falcipain-1 in sequence. The current model of falcipain-2 is based on a crystal structure of human cathepsin K, which is 35% identical to falcipain-2 in sequence over the mature protease domain. Using standard homology modeling procedures, we constructed a model of the active form of the falcipain-2 hemoglobinase.

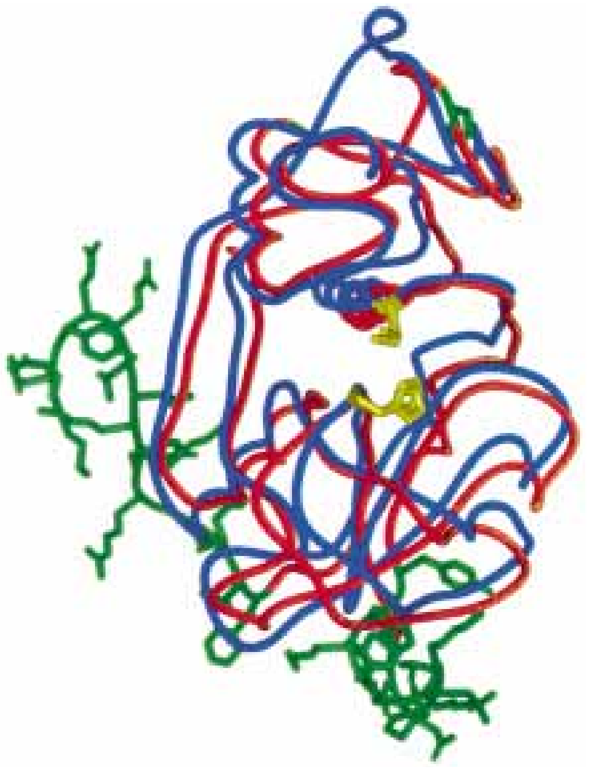


Fig. 1. Superposition of falcipain-1 and falcipain-2 model structures. A structural alignment of the falcipain-2 (red) and falcipain-1 (blue) model structures was performed with MINAREA (17). Structural positions present in falcipain-2 and absent in falcipain-1 are colored green. The catalytic dyad is shown in yellow for reference. The figure was generated with CHIMERA (18).

A comparison of the falcipain-1 and falcipain-2 models was carried out to determine features that could explain functional differences of these proteases and to direct our drug design efforts. The majority of the sequence changes, and all of the three insertion-deletion events, are on the face of the protein opposite the active site (Fig. 1). Nevertheless, there are a number of changes that significantly alter the features of the specificity sites, predominantly on the non-prime side of the peptide binding cleft that recognizes the side-chains on residues N-terminal to the cleavage site.

Protease specificity is frequently studied in the context of subsites that flank the catalytic residues and provide the enzyme with specific preferences for peptide or protein substrates. Following the nomenclature of Berger and Schechter (19), these sites are referred to as S_4, S_3, S_2, S_1, S_1', S_2', S_3' (Fig. 2), and they correspond to substrates with the sequence P_4, P_3, P_2, P_1, P_1', P_2', P_3', where the P_1-P_1' peptide bond is cleaved. The papain cysteine protease family has well-defined sites from S_3 to S_1', with some individual proteases having more extended specificity. The S_2 and S_1 sites contribute the strongest preference to substrate binding in the case of falcipain-2 (7) and many other papain family proteases (21).

In all, there are 11 amino acid differences in the S_2, S_3 and S_4 sites of falcipain-2 relative to falcipain-1 (Fig. 3). The S_2 site is most variable with a total of six differences ranging from conservative to functionally significant ones. The combination of conservative changes retains the overall hydrophobic character of this site; however, there is a net gain of two non-hydrogen atoms in side-chains on the part of falcipain-1, decreasing the free volume available for binding in this site. Two pairs of these sequence differences appear to be compensating substitutions: S46A and A175S conserve the serine, while N86F and S149N conserve the asparagine. At the other end of the spectrum, the sequence difference I85P is predicted to have a pronounced effect on the local backbone geometry. This is evident from a superposition of the two models. Together, these sequence differences in the S_2 site are predicted to have significant impact on the binding and kinetics of the substrate-protease interaction.

The S_3 site in these proteases is composed of less than half as many residues as the S_2 pocket (four versus nine). In this context, the three sequence differences in S_3 site change an even greater percentage of the binding site's surface. It should be noted that in the current specificity site designation, the S_3 and S_4 sites share one residue that differs between falcipain-2 and falcipain-1 (N86F). Overall there is less hydrophilic functionality lining the site in falcipain-2 (Y78F and N86F), though falcipain-1 has an additional basic functionality (L84H). These S_3 site sequence differences are predicted to shrink the substrate binding volume and give rise to a preference for hydrophobic residues for falcipain-1.

Falcipain-2 is predicted to have an additional strand at the edge of the beta sheet structure forming part of the S_4 binding site (Fig. 1). Although this region is more likely to play a part in extended specificity, it appears to be the largest global structural difference between falcipain-2 and falcipain-1. This has direct implications for substrate binding. Four out of six residues in the S_4 site differ between falcipain-2 and falcipain-1. Three of these changes result in gain of hydrophobic functionality in the falcipain-1 S_4 site (Fig. 3). Interestingly, in spite of the sequence differences in the S_4 site, the molecular dimensions of the S_4 pocket differ only by one non-hydrogen atom. It appears that these two proteases do not share substrate specificity at the P4 position, although amino acids with similar volumes may be preferred.

Genome and Protein Family Based Drug Discovery: Falcipain-2 and Human Homologs in the Papain Superfamily

A reality of the post-genomic era is access to a seemingly endless array of genome sequences. It is now possible to annotate proteins and analyze the homologies and variations between the pathogen proteins and the human homologs. For oncologic

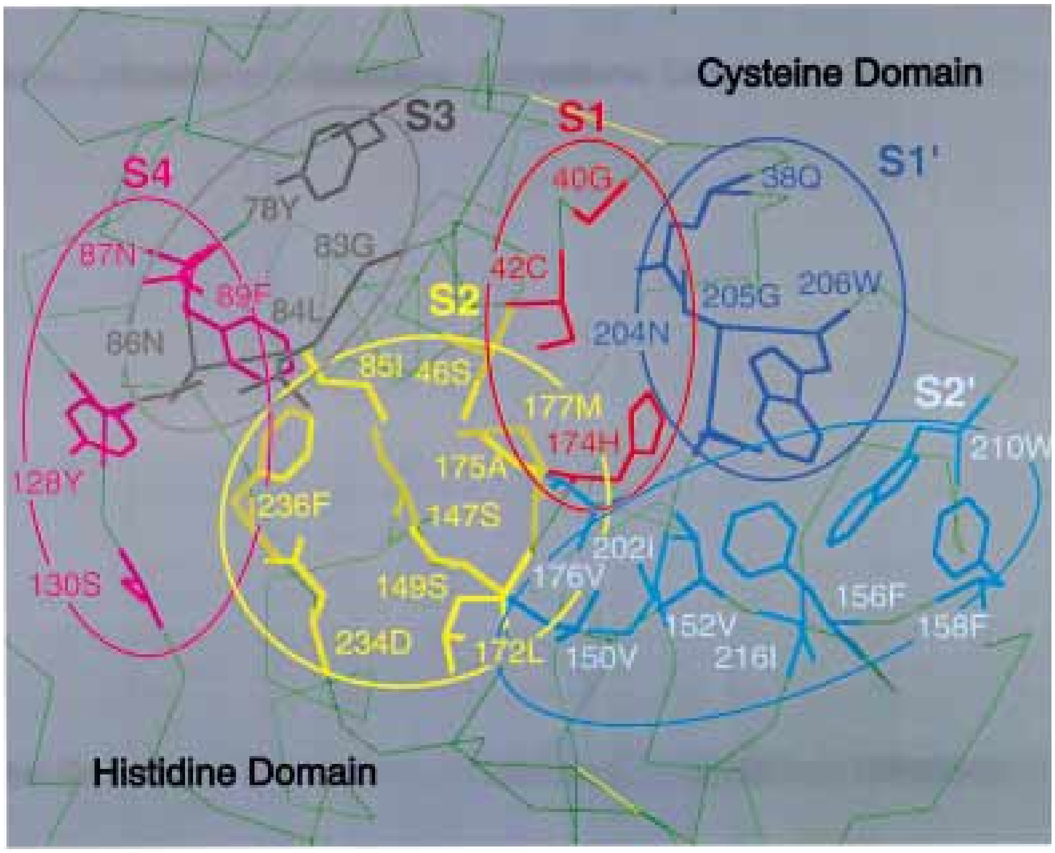
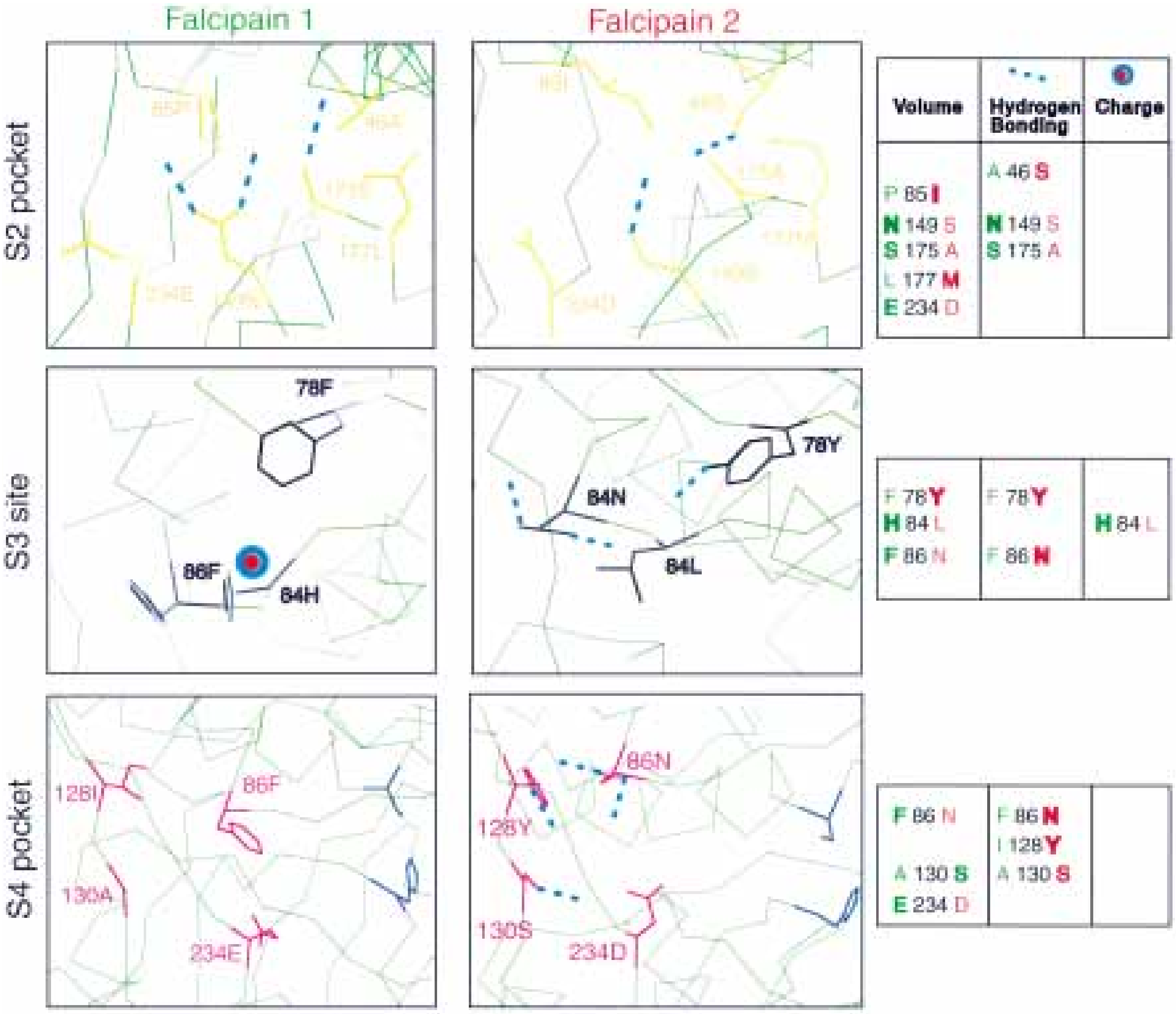


Fig. 2. The falcipain-2 model specificity sites. The residues that line the substrate specificity sites are displayed on the falcipain-2 model. The natural peptide substrate orientation prefers the non-prime side for the *N*-terminus, and the prime side for the *C*-terminus. Residues at the boundaries of a site may contribute to neighboring sites. The figure was generated with WEBMOL (20) and a sequence to structure alignment and family analysis JAVA^{TM} package (M.P. Joachimiak, *in preparation*).

disease, the pathogenic protein(s) may be mutated, up regulated or in the case of tumor supressors, down regulated. While it is possible to imagine small molecule inhibitors of mutated or up regulated proteins (e.g., Gleevec for BCR-ABL (22)), down regulated systems are less likely to be amenable to small molecule approaches. In all of these cases, the impact of a small molecule inhibitor on the related protein targets must be considered. In the case of Gleevec, c-kit and gastrointestinal stromal tumors, an unexpected benefit is found (23,24). However, it is more likely that untoward side effects will result. As is common in the case of infectious disease drug targets, drug resistance of oncoprotein targets can occur by amino acid substitution of residues involved in the drug interaction (25).

Subsite specificity analysis of the falcipain-2 model suggested a number of favorable features for drug design. The dominant feature of the falcipain-2 S_2 site is a deep hydrophobic binding pocket. As judged by peptide substrate binding data, falcipain-2 (7) and the modeling template cathepsin K (26) share a marked P_2 preference for leucine. The S_3 site is quite small and largely solvent accessible, in keeping with the trend of the papain superfamily. Even the extended specificity S_4 site of falcipain-2 has a potential binding pocket. However, the extended non-prime specificity sites are problematic for drug design because they appear poorly defined structurally and substrate analog binding data shows no preferences in this region (27). To direct computational small molecule selection calculations and to further understand structure-specificity relationships, we proceeded to analyze the commonalities and distinctions between falcipain-2 and the other plasmodial and human papain-like cysteine proteases.

A multiple sequence alignment based on the available sequences and structures of human cysteine proteases was created and used to assign falcipain-2 residues to substrate specificity sites. More than half of the residues on the prime side of the specificity sites are conserved. The few differences have little impact upon site volume, hydrogen bonding or charge. The S_1/S_1' catalytic site including the catalytic dyad, a glycine residue and a number of backbone atoms, is absolutely conserved within this family. This leaves fewer than half of the specificity sites as possible unique structural sites for differential drug design.



Volume	Hydrogen Bonding	Charge
P 85 I N 149 S S 175 A L 177 M E 234 D	A 46 S N 149 S S 175 A	
F 78 Y H 84 L F 86 N	F 78 Y F 86 N	H 84 L
F 86 N A 130 S E 234 D	F 86 N I 128 Y A 130 S	

Fig. 3. Falcipain-1 versus falcipain-2 S_2, S_3 and S_4 specificity site analysis. To highlight differences that affect the S_2, S_3 and S_4 specificity sites, the sites were analyzed with respect to sequence differences between the two plasmodial proteases (see Methods). Residues that gained hydrogen bonding functionality relative to the other sequence are marked with blue dashes, residues that gained charge functionality are marked by a red ball within a blue ball. The table of residue changes highlights the sequence that gained functionality with a boldfaced font. The figure was generated with WEBMOL (20) and a sequence to structure alignment and family analysis JAVA^{TM} package (M.P. Joachimiak, *in preparation*).

Excluding glycines and main chain atoms, the S_2, S_3 and S_4 sites are variable across the papain cysteine protease family. These sites have diverged during evolution to optimize different functional substrate specificities. Certain sequences exhibit compensating changes, but for nearly all the specificity site sequence positions there exist variations in volume, hydrogen bonding potential and even charge. Due to the contribution of these specificity sites to substrate binding, such patterns of sequence variation represent the unique functional signature of the papain family. Within the functional variations of a protein family resides an important aspect of protease differential specificity-how changes in sequence affect the binding site volume. If we assume that the backbone positions remain relatively fixed, then mutations to a smaller residue will result in a larger available volume for binding and vice versa. Such unique differences can be exploited with distinct substituents attached to a common small molecule scaffold. In contrast, conserved signatures like the S_1/S_1' catalytic site are problematic for targeted drug design because of their ubiquitous presence within the protein family.

Using the available sequence data we performed a variation of the Evolutionary Trace method (28), with a JAVA^{TM} implementation of the ET analysis (data not shown). Combined with a specificity site annotation by analogy to characterized homologs, the analysis identified amino acids that were unique in falcipain-2 relative to the known human homologs (Fig. 4). A surface-exposed cluster of residues was identified at the boundaries of the S_2, S_3 and S_4 specificity sites. The S_2 site, the main determinant of

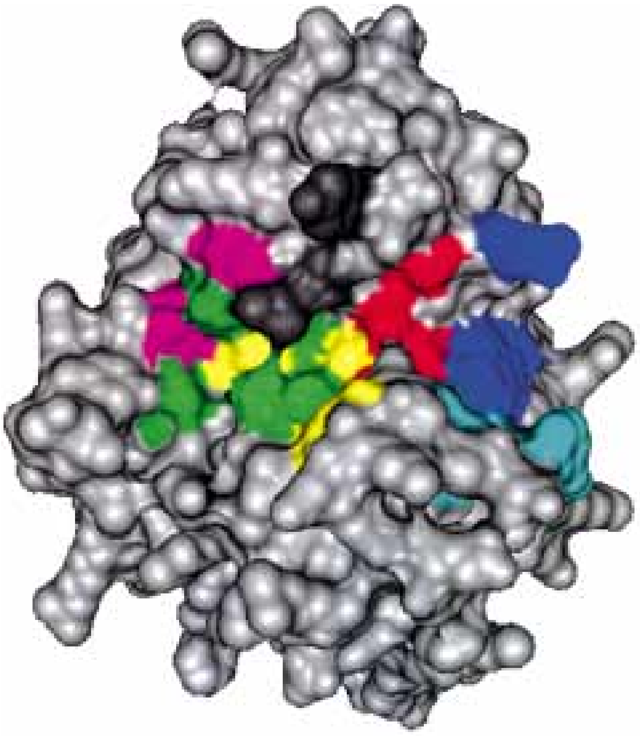


Fig. 4. Unique site analysis of falcipain-2 in the context of human homologs. Cathepsins B, C, H, K, L, L2, O, S, Z and stefin B were the human homologs used in this evolutionary analysis. In green are residues unique in falcipain-2 relative to the human sequences. Other colors correspond to the defined specificity sites as represented in Figure 2. See Methods for definition of the specificity sites and the unique site analysis. This figure was generated with a sequence to structure alignment JAVA^{TM} application (M.P. Joachimiak, *in preparation*), CHIMERA (18) and MSMS (29).

substrate binding, contained the majority of these unique site positions (three out of six).

The unique site identified in the falcipain-2 specificity sites is solvent accessible, spans two well defined binding pockets and exhibits marked sequence differences relative to human papain family protease homologs. Together this evidence suggested that the identified cluster of residues was a promising candidate drug target site for an antimalarial with minimized specificity towards human homologs of the target. The results of this analysis were directly applied to both the in silico and visual screening steps.

Drug Discovery Results Against Falcipain-2 Compared to Falcipain-1

The original falcipain modeling and drug design effort (6) led to three inhibitors with an IC_{50} less than 100 μM. The best compound was a naphthylhydrazide, which inhibited the plasmodial protease extract with an IC_{50} of 6 μM in an *in vitro* enzyme assay, and had activity against the parasite in culture at a similar concentration as judged by inhibition of hypoxanthine uptake. Overall 31 compounds were tested in the original DOCK screen (6), resulting in a 3% hit rate at the $<$10 μM cutoff.

The current modeling and drug design effort has had a considerably higher hit rate in terms of active compounds found with the aid of a computational screen. Of the 44 compounds tested, eight had an IC_{50} below 10 μM in an *in vitro* enzyme assay with values ranging from 1 to 7 μM (Table 1). Three of the eight best compounds against falcipain-2 (2,4,7), were also effective in killing parasites with an IC_{50} of about 20 μM. For these compounds there was complete inhibition of parasite multiplication at 50 μM.

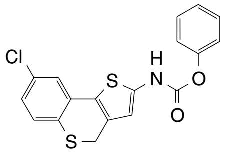
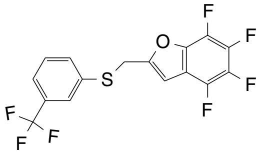
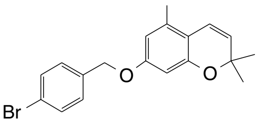
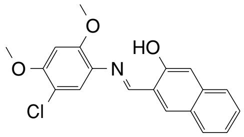
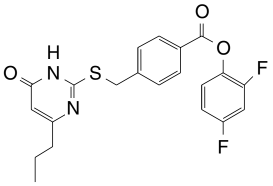
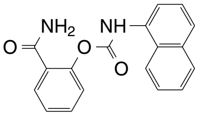
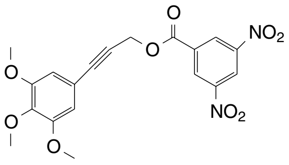
Discussion

There are several advantages to pursuing a family of proteins as drug discovery targets, including: 1) compounds related to inhibitors of one family member are likely to be active against other members; 2) the structure of one member provides substantial insights into the structure and function of other members; 3) assay development can proceed in parallel; and 4) experience developed for one target is frequently relevant to the homologous targets. Confounding these advantages, inhibitor specificity can be a significant challenge and the relatedness of the targets means that the inhibitors, especially those that are mechanistic in nature, are likely to have several distinct activities.

In the case of cysteine proteases, not only are there thousands of cysteine protease sequences in the sequence databases, but this ubiquitous sequence family also has tens of well-determined crystal structures complexed with inhibitors. This is a favorable situation for modeling, predicting substrate specificity and forming inhibitor structure-activity relationships.

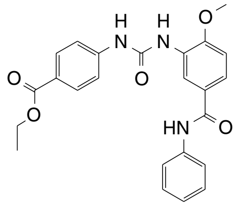
Based on our evolutionary analysis, the S_2, S_3 and S_4 specificity sites of falcipain-2 exhibit unique functional differences that can be targeted with drug design. In contrast, conserved signatures traditionally targeted with mechanistic inhibitors, like the S_1/S_1' catalytic site, logically lead to the selection of compounds with specificity towards human homologs of the drug target. In addition, the non-prime sites encompassing the unique falcipain-2 site should lead to decreased drug side effects. We predict however, that for instances involving a target whose interactions are strictly selected for and where the binding partners (small or macromolecules) also exhibit chemical conservation, the present analysis will not guarantee a unique cluster of residues. Nevertheless, for many protein families, functional speciation can be observed in the coevolution of binding interfaces and how side-chain variation at the interfaces occurs in a correlated manner (30). All instances of speciation of function result in differences that can be exploited to direct drug design efforts.

Table 1. Inhibition of falcipain-2 and cultured malaria parasites by compounds identified with a computational screen

	Enzyme IC$_{50}$ * ** (μM)	Cell Culture IC$_{50}$ ** (μM)
1. (8-chloro-4*H*-1,5-dithia-cyclopenta[a]naphthalen-2-yl)-carbamic acid phenyl ester  	1.1 +/− 0.5	94
2. 4,5,6,7-tetrafluoro-2-(3-trifluoromethyl-phenylsulfanylmethyl)-benzofuran  	1.4 +/− 0.6	20
3. 7-(4-bromobenzyloxy)-2,2,5-trimethyl-2*H*-chromene  	2.5 +/− 1.3	no inhibition
4. 1-(5-chloro-2,4-dimethoxyphenyliminomethyl)-2-naphthol  	3.5 +/− 0.6	25
5. 4-(6-oxo-4-propyl-1,6-dihydroprymidin-2-ylsulfanylmethyl)-benzoic acid 2,4-difluoro-phenyl ester  	4.1 +/− 1.3	ND
6. Naphthalen-1-yl-carbamic acid 2-carbamoyl-phenyl ester  	4.7 +/− 2.8	103
7. 3,5-dinitro-benzoic acid 3-(3,4,5-trimethoxy-phenyl)-prop-2-ynyl-ester  	6.4 +/− 1.1	21

(Continued)

Table 1. (Continued)

	Enzyme IC_{50} * ** (μM)	Cell Culture IC_{50} ** (μM)
8. 4-[3-(2-methoxy-5-phenylcarbamoyl-phenyl)-ureido]-benzoic acid ethyl ester 	6.9 +/− 2.6	ND

*Leupeptin, the enzyme assay positive control, had an IC_{50} of 50 nM.

**See Methods for details.

Comparison of the structural sites targeted with computational screening shows important differences between falcipain-1 and falcipain-2. For falcipain-1 the most active compound was predicted to bind to the $S_2/S_1/S_1'$ sites (Fig. 5). Given that S_1 and S_1' are the conserved signature of the papain family, this inhibitor has the potential to cross-react with human cysteine proteases. The set of inhibitors generated by virtually screening the falcipain-2 model produced eight diverse compounds, all selected to bind the unique functional signature of the S_2 and extended non-prime sites.

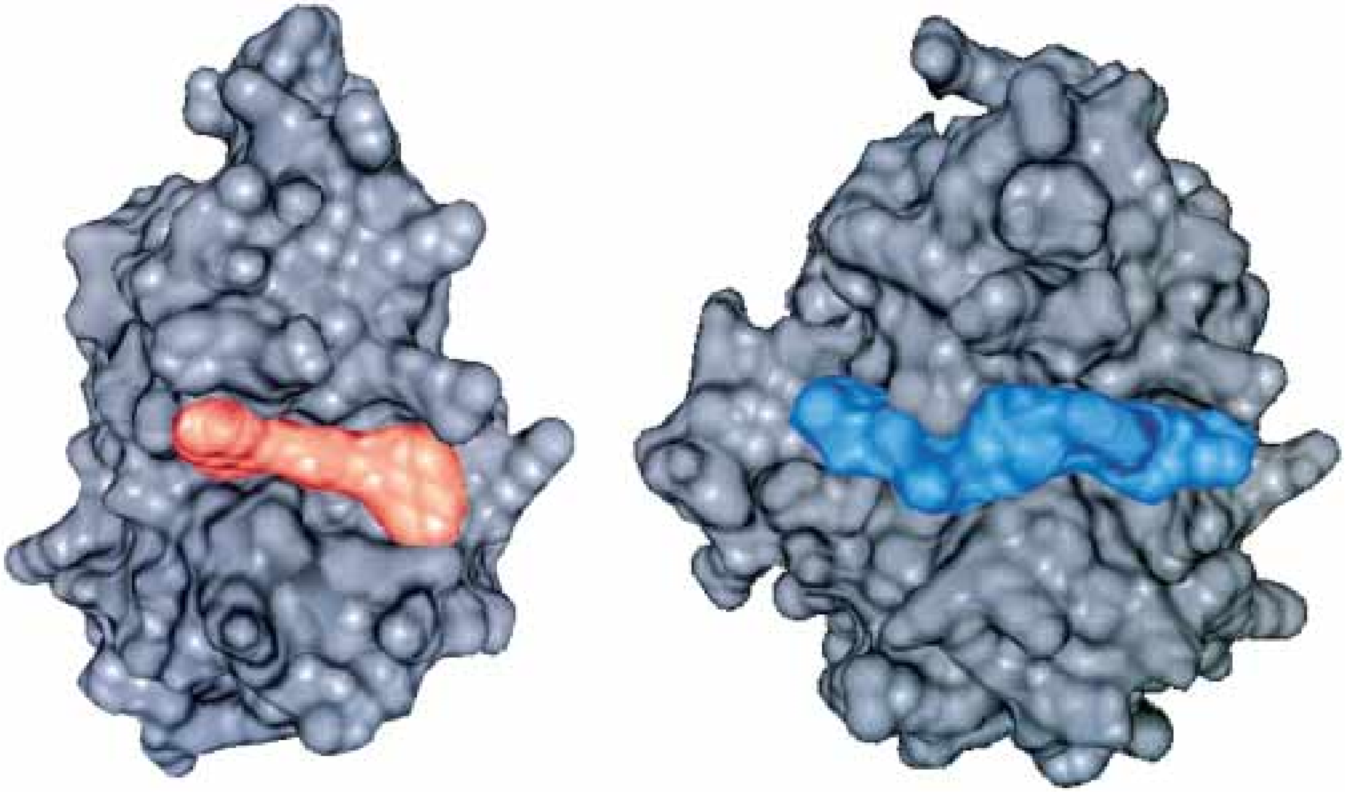


Fig. 5. Inhibitor binding modes of falcipain-1 compared to falcipain-2. Models show the predicted binding modes of the best inhibitors for falcipain-1 (orange, left) and falcipain-2 (blue, right). The falcipain-1 model structure is shown with the predicted binding mode of its best symmetric acyl-hydrazide inhibitor from the work of Ring et al (6). The falcipain-2 model is shown with the predicted binding modes for the top eight active compounds. The figure was generated with MSMS (29) and the MSV software (31).

The common feature of many of these compounds, including the falcipain-1 inhibitor, is a pseudopeptide backbone of length 2–4 atoms adopting a planar conformation owing to the double bonds and the two flanking functionalized ring systems. We continue to believe that the length of the linker and chemical substituents on the functional groups are what determine the specificity and uniqueness of the target-inhibitor interaction (15,16).

It is difficult to extrapolate the present cell culture results to the antimalarial action of other known falcipain inhibitors because earlier cellular assays varied in method. In addition, the actual concentration of compound reaching the parasite food vacuole in the cell culture experiments is likely to be affected by permeability through the multiple cell membranes involved. All of the compounds tested in this study had molecular weights less than 350 Da, and all were soluble in water at appreciable concentrations. Nevertheless, it cannot be ruled out that the cell culture IC_{50} measurements in our experiments do not reflect the effective concentration of compound at the target site, presumably the parasite food vacuole.

There are some factors convoluting the results of both the *in vitro* enzyme and the *in vivo* culture inhibition assays. An important difference in the drug discovery effort against falcipain-1 compared to falcipain-2 is the size of the chemical database used in the computational screen. The Fine Chemicals Directory used in 1993 contained 55313 compounds, whereas the version of the Available Chemical Directory used in the current drug discovery effort consisted of 195419 commercially available compounds. This 4-fold increase in database size results in significantly greater chemical diversity available for computational screening and drug discovery. The percentage sequence identity to the modeling target, a standard measure of model accuracy, is predicted to have had a negligible effect in this case. The percentage sequence identity between falcipain-1 and papain and actinidin was 33%, while that of falcipain-2 to cathepsin K was 35%. Substantially more sequence similarity was observed in the region around the active site that should be most important to drug design efforts.

The most powerful convolution in the falcipain-1 drug discovery effort was the assay of enzyme inhibition performed using a parasite extract now known to be primarily composed of falcipain-2. Knowledge about the expected and associated phenotypes has considerably increased in the past few years—assessing the *in vivo* inhibition phenotype began with a general metabolism assay, continued with food vacuole swelling as exhibited by broad-spectrum inhibitors, and now with multiple homologous targets has returned to more general assays of parasite health and development. Significantly, technological improvements in modeling, structure analysis and docking, have become combined with the accumulation of sequences, crystal structures, and available small molecules. An important remaining rate limiting step in post-genomic drug discovery is knowledge about the target. Such knowledge includes the cellular and disease contexts, other gene products modulating the targets' function, as well as the functional family it belongs to within a specific genome and beyond.

Presently the function of falcipain-1 remains unknown and the current analysis may serve as a lead in the search for its endogenous targets. The sites responsible for substrate specificity in falcipain-1 and falcipain-2 are notably different given the high degree of sequence similarity. It is postulated that these two cysteine proteases have different endogenous target sequences and therefore different *in vivo* functions.

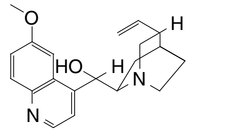
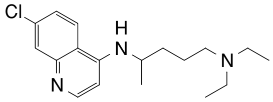
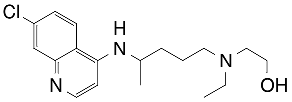
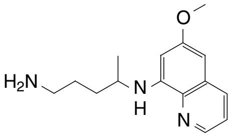
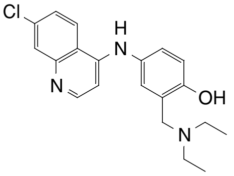
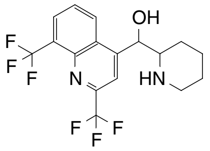
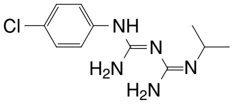
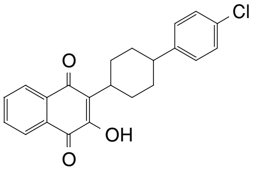
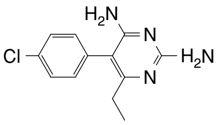
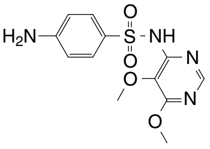
Our drug discovery effort against falcipain-2 has resulted in eight compounds with activities <7 μM, three of which kill parasites in a cell culture assay. Presumably, their potency can be increased through a combined medicinal and computational chemistry effort directed at the unique site of falcipan-2. Most importantly, nearly all antimalarials in current use have pronounced side effects and/or have encountered plasmodial drug resistance. The molecules identified in the falcipain-2 drug discovery effort represent diverse chemical scaffolds and functionalities relative to the known antimalarial drugs (Table 2). It follows, that the structures we have identified as falcipain-2 inhibitors provide new avenues for antimalarial drug development with potential for minimized toxicity and drug resistance.

Conclusions

Based on modeling and drug design we have explored the effects of the association between a gene product and its phenotype in the context of a gene family. Discovery of the new sequences and their experimental confirmation as targets immediately led to new models of a new target. Higher expectations for success were set for drug design efforts, specifically structure-activity correlations and improved specificity for the parasite enzyme, both in culture and in animal tests. Based on the current drug design effort and other examples of protease inhibition, a primary cause of side effects is accumulation of undesirable substrates. In molecular terms such side effects are a signal for the presence of genes potentially unrelated to the target phenotype (e.g., ACE inhibition leads to accumulation of bradykinin and substance P (33)). In the case of malaria infection in humans, the side effects of broad papain family inhibitors would most likely mean accumulation of many substrates, the majority of which are host proteins.

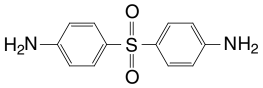
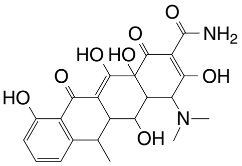
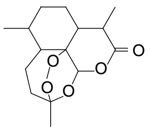
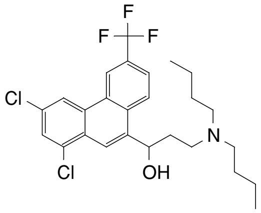
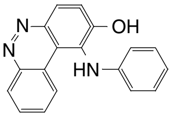
Biological phenotypes are fundamentally complex due to the presence of back-up functionalities, possibly distributed across cell types, tissues and time, as well as clearance and other protective mechanisms. These pitfalls of drug development illustrate the

Table 2. Chemical structures and properties of prescribed antimalarial drugs

Name	Resistance [32]	Toxicity [32]	Compound
quinine	Yes	delayed 'glaucoma'	
chloroquine	Yes	heart liver	
hydroxychloroquine	Yes	retina macula	
primaquine	Yes	anemia	
amodiaquine	like chloroquine	agranulocytosis with long term treatment	
mefloquine	Yes	cardiotoxicity, vivid dreams, psychosis	
proguanil	Yes DHFR	ulcers, alopecia	
atovaquone (with proguanil)	Yes selective cytochrome b	minor	
pyrimethamine (with sulfadoxine or sulfonylbisbenzenamine)	Yes DHFR	severe skin disease	
sulfadoxine (with pyrimethamine)	Yes	severe skin disease	

(Continued)

Table 2. (Continued)

Name	Resistance [32]	Toxicity [32]	Compound
sulfonylbisbenzenamine (with pyrimethamine)		anemia, allergy, fever	
doxycycline (with quinine)		sun sensitivity	
Artemisinin		CNS toxicity	
halofantrine		cardiotoxicity	
Pyronaridine		teratogen	

detailed knowledge required for pursuing targets even with established phenotypes. A gene family perspective can lead to unique structural sites relative to the 'host' or 'pathogen' families. Knowledge of gene families can aid in the identification of the inhibitory spectrum of a molecule and provide the insight and unanticipated auxillary functions of the original target or back-up functions subsumed by family members.

Materials and Methods

Homology Methods and Structural Modeling

A structural model of falcipain-2 was constructed by homology to other members of the papain cysteine protease subfamily. In order to identify a structural template, the protein structure database (PDB) was searched for falcipain-2 homologs using the PSI-BLAST algorithm (34). The closest homologs were the cathepsin K zymogen (e-value of 3E-43 over 321 residues), the cathepsin L zymogen (2E-41 over 326 residues), the caricain zymogen (4E-41 over 343 residues), and the ginger rhizome cysteine protease (1E-38 over 220 residues). Typically, the best template for modeling corresponds to the sequence with the longest significant alignment and the highest score in the mature protease region. At a per-residue level, human cathepsin K was found to be 39% identical in the mature region of the protease (35% identity over all aligned residues).

A model of falcipain-2 based on the cathepsin K zymogen structure (PDB code: 1BY8, resolution 2.6 Å) was built using MODELLER (35). This software derives distance and angle constraints based on conserved sequence features in the alignment and structural features of the template. Given a correct alignment, sequences that share 40% identity are expected to align within 1 Å RMS over 90% of their residues, approximately the accuracy expected in the present model. Falcipain-2 has an additional predicted disulfide bond relative to cathepsin K, and this disulfide bridge was added with the modeling software SYBYL (TRIPOS corp.). The model structure was refined at the side-chain level using the backbone-dependent side-chain rotamer library algorithm SCWRL (36). For pairs of sequences with gaps inserted to yield an alignment with identical residues in 30–40% of the positions and using a template structure determined using 2 Å resolution x-ray data, SCWRL predicts the χ_1 side-chain angles with an accuracy of 65%. All identical aligned residues were fixed in their template conformation. Of the eight residues with unlikely conformations identified by SCWRL, all were distant from the active site and occurred in regions of insertions relative to the structural template.

Annotation of Specificity Subsites and Unique Site Analysis

The substrate specificity sites in the falcipain-2 model structure were identified by analogy to the extensive family of papain-like cysteine proteases. Following the nomenclature of Berger and Schechter (19) and sequence alignments to known papain family crystal structures, the falcipain-2 S_4 to S_2' substrate side-chain binding sites on either side of the scissile amide bond were identified. A more extensive multiple sequence alignment was built using CLUSTALW (37) and edited to include the alignments derived by structure alone.

As we seek inhibitors that are unlikely to be active against human proteases from the papain family, we performed a variation of the Evolutionary Trace method (22) on falcipain-2 and its human homologs. The variation consists of restricting the sequence data to a subset of the full sequences, corresponding to the aligned specificity sites. The definition of residue conservation and subfamily comparison was modified, by considering only amino acids that were unique in human sequences relative to the falcipain-2 target. Finally, residue similarity filters (BLOSUM62 (38)) and coloring by groups were applied, to analyze the unique positions in terms of volume, hydrogen bonding and charge properties. All of the above functions and the resulting mapping of phylogenetic and sequence data onto the falcipain-2 model structure (Fig. 4). were performed with a JAVA^{TM} application (data not shown).

Docking

DOCK 4.0 (39) was used to screen the falcipain-2 sites against the Available Chemicals Directory release 97.2 containing 195419 unique compounds (MDL Inc). The screening procedure took about six weeks of CPU time on a 4 processor MIPS R12000 SGI server. 5000 compounds were saved from the energy-scoring scheme, and 5000 from the shape-scoring scheme. Visual selection of these hits was performed in duplicate, to arrive at a set of 160 compounds in an unbiased fashion. This selection step relied on knowledge of the specificity sites unique in falcipain-2 relative to known human sequences, as well as standard drug-like properties of small molecules including: hydrophobicity, molecular weight, and absence of chemical functionalities with tendencies to form covalent adducts with amino acid side-chains.

Falcipain-2 Enzyme Assays

A set of 44 compounds manually selected from the DOCK computational screen of the falcipain-2 active site was tested in a fluorescence-based assay against recombinant falcipain-2. Recombinant falcipain-2 was prepared (7) and the falcipain-2 fluorescence-based assay performed as previously described (13). All compounds were dissolved in DMSO to make a 10 mM stock solution. Each compound was incubated with the enzyme in 0.1 M sodium acetate (pH 5.5) and 10 mM dithiothreitol (DTT) for 30 minutes at room temperature before addition of the substrate benzyloxycarbonyl-Phe-Arg-7-amino-4-methyl-coumarin (Z-Phe-Arg-AMC). The fluorescence caused by the cleavage of the substrate was monitored continuously over 30 minutes with a Fluoroskan II spectrofluorometer (Labsystems). The rate of hydrolysis of Z-Phe-Arg-AMC in the presence of the compounds was compared with the rates of hydrolysis in the negative (equivalent volume of dimethyl sulfoxide (DMSO)) and positive (100 μM leupeptin) controls. Using the PRISM 3.0 software (Graphpad Software Inc.), the 50% inhibitory concentration of each compound (IC_{50}) was determined from plots of falcipain-2 activity inhibition over a series of compound concentrations. Initial fluorescent assay screens were carried out for 44 DOCK compounds and those with IC_{50}'s below 10 μM were selected for further testing.

Cell Culture Assays

The six best compounds, 1–4, 6 and 7 (Table 1) were selected for characterization in a cell-based assay. Final concentrations of compounds in the parasite cultures were 100 μM, 50 μM, 25 μM and 10 μM and the final concentration of the DMSO control was 1%. W2 strain *P. falciparum* parasites were cultured with human erythrocytes at 2% hematocrit in RPMI-1640 medium supplemented with 10% heat inactivated human serum (13). Parasite synchrony was maintained by serial treatments with 5% sorbitol (40). In order to assess the effects of inhibitors on parasite development, *P. falciparum* parasites were incubated for 48 hours with different concentrations of compound added from 100× stocks in DMSO (13). The experiment was started at the synchronized young ring stage and continued until the control cultures contained nearly all new ring stage parasites (48 hours). Giemsa-stained smears were made at 24 and 48 hours. At 24 hours parasite morphology was evaluated and at 48 hours the number of new ring forms per 1000 erythrocytes were counted and compared to control cultures incubated with DMSO. IC_{50}'s for compounds 1–4, 6 and 7 were calculated using the PRISM 3.0 software.

Acknowledgments

We are grateful to Xiaohui Du, Florence Horn, Barney May and Elaine Meng for their comments and advice. This work was supported by grants from the Burroughs-Wellcome Fund and the National Institute of Health.

References

1. Mark KS, Davis TP. (2000) Stroke: development, prevention and treatment with peptidase inhibitors. *Peptides* **21**: 1965–1973.
2. Tebas P, Powderly WG. (2000) Nelfinavir mesylate. *Expert. Opin. Pharmacother.* **1**: 1429–1440.
3. Fisher ND, Hollenberg NK. (2001) Is there a future for renin inhibitors? *Expert. Opin. Investig. Drugs* **10**: 417–426.
4. Rosenthal PJ, Kim K, McKerrow JH, Leech JH. (1987) Identification of three stage-specific proteinases of *Plasmodium falciparum*. *J. Exp. Med.* **166**: 816–821.
5. Rosenthal PJ, Nelson RG. (1992) Isolation and characterization of a cysteine protease gene of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **51**: 143–152.
6. Ring CS, Sun E, McKerrow JH, et al. (1993) Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc. Natl. Acad. Sci. USA* **90**: 3583–3587.
7. Shenai BR, Sijwali PS, Singh A, Rosenthal PJ. (2000) Characterization of native and recombinant falcipain-2, a principal trophozoite cysteine protease and essential hemoglobinase of *Plasmodium falciparum*. *J. Biol. Chem.* **275**: 29000–29010.
8. Rosenthal PJ, McKerrow JH, Aikawa M, Nagasawa H, Leech JH. (1988) A malarial cysteine proteinase is necessary for hemoglobin degradation by *Plasmodium falciparum*. *J. Clin. Invest.* **82**: 1560–1566.
9. Francis SE, Gluzman IY, Oksman A, et al. (1994) Molecular characterization and inhibition of a *Plasmodium falciparum* aspartic hemoglobinase. *EMBO J.* **13**: 306–317.
10. Hill J, Tyas L, Phylip LH, Kay J, Dunn BM, Berry C. (1994) High level expression and characterisation of Plasmepsin II, an aspartic proteinase from *Plasmodium falciparum*. *FEBS Lett.* **352**: 155–158.
11. Le Bonniec S, Deregnaucourt C, Redeker V, et al. (1999) Plasmepsin II, an acidic hemoglobinase from the *Plasmodium falciparum* food vacuole, is active at neutral pH on the host erythrocyte membrane skeleton. *J. Biol. Chem.* **274**: 14218–14223.
12. Eggleson KK, Duffin KL, Goldberg DE. (1999) Identification and characterization of falcilysin, a metallopeptidase involved in hemoglobin catabolism within the malaria parasite *Plasmodium falciparum*. *J. Biol. Chem.* **274**: 32411–32417.
13. Rosenthal PJ, Wollish WS, Palmer JT. (1991) Antimalarial effects of peptide inhibitors of a *Plasmodium falciparum* cysteine protease. *J. Clin. Invest.* **88**: 1467–1472.
14. Rosenthal P, Olson JE, Lee K, Palmer JT. (1996) Antimalarial effects of vinyl sulfone cysteine protease inhibitors. *Antimicrob. Agents. Chemother.* **50**: 1600–1603.
15. Li Z, Chen X, Davidson E, et al. (1994) Anti-malarial drug development using models of enzyme structure. *Chem. Biol.* **1**: 31–37.
16. Li R, Chen X, Gong B, et al. (1996) Structure-based design of parasitic protease inhibitors. *Bioorg. Med. Chem.* **4**: 1421–1427.
17. Falicov A, Cohen FE. (1996) A surface of minimum area metric for the structural comparison of proteins. *J. Mol. Biol.* **258**: 871–892.
18. Huang CC, Couch GS, Pettersen EF, Ferrin TE. (1996) Chimera: An Extensible Molecular Modeling Application Constructed Using Standard Components. *Pac. Symp. Biocomput.* **1**: 724.
19. Schechter I, Berger A. (1968) On the active site of proteases. 3. Mapping the active site of papain; specific peptide inhibitors of papain. *Biochem. Biophys. Res. Commun.* **32**: 898–902.
20. Walther, D. (1997) WebMol—a Java based PDB viewer. *Trends Biochem. Sci.*, **22**: 274–275.
21. McGrath ME. (1999) The lysosomal cysteine proteases. *Annu. Rev. Biomol. Struct.* **28**: 181–204.
22. Druker BJ, Talpaz M, Resta DJ, et al. (2001) Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* **344**: 1031–1037.
23. Tuveson DA, Willis NA, Jacks T, et al. (2001) STI571 inactivation of the gastrointestinal stromal tumor c-KIT oncoprotein: biological and clinical implications. *Oncogene* **20**: 5054–5058.
24. Strickland L, Letson GD, Muro-Cacho CA. (2001) Gastrointestinal stromal tumors. *Cancer Control* **8**: 252–261.
25. Gorre ME, Mohammed M, Ellwood K, et al. (2001) Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science* **293**: 876–880.
26. Bossard MJ, Tomaszek TA, Thompson SK, et al. (1996) Proteolytic activity of human osteoclast cathepsin K. Expression, purification, activation, and substrate identification. *J. Biol. Chem.* **271**: 12517–12524.
27. Turk D, Guncar G, Podobnik M, Turk B. (1998) Revised definition of the substrate binding sites of papain-like cysteine proteases. *Biol. Chem.* **379**: 137–147.
28. Lichtarge O, Bourne HR, Cohen FE. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**: 342–358.
29. Sanner MF, Olson AJ, Spehner JC. (1995) Fast and robust computation of molecular surfaces. *Proc. 11th ACM Symp. Comp. Geom.* **C6–C7**.
30. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. (2000) Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299**: 283–293.
31. Sanner MF, Spehner JC, and Olson AJ. (1996) Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**: 305–320.
32. Katzung, BG ed. (2001) *Basic and Clinical Pharmacology*. 8th edition, Lange Medical Books/McGraw-Hill, New York, pp. 882–893.
33. Emanueli C, Grady EF, Madeddu P, et al. (1998) Acute ACE inhibition causes plasma extravasation in mice that is mediated by bradykinin and substance P. *Hypertension* **31**: 1299–1304.
34. Altschul SF, Madden TL, Schaffer AA, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402.
35. Sanchez R, Sali A. (1997) Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins* **Suppl 1**: 50–58.
36. Bower MJ, Cohen FE, Dunbrack RL. (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J. Mol. Biol.* **267**: 1268–1282.
37. Thompson JD, Higgins DG, Gibson TJ. (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.
38. Henikoff S, Henikoff JG. (1993) Performance evaluation of amino acid substitution matrices. *Proteins* **17**: 49–61.
39. Ewing TJA, Makino S, Skillan AG, Kuntz ID. (2001) DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided. Mol. Des.* **5**: 411–428.
40. Lambros C, Verderberg JP. (1979) Synchronization of *Plasmodium falciparum* erythrocytic stages in culture. *J. Parasitol.* **65**: 418–420.