

# Modeling and structural analysis of evolutionarily diverse S8 family serine proteases

Aparna Laskar<sup>1</sup>, Euan James Rodger<sup>2</sup>, Aniruddha Chatterjee<sup>2, 3\*</sup>, Chhabinath Mandal<sup>1</sup>

<sup>1</sup>Indian Institute of Chemical Biology (CSIR Unit, Government of India), Kolkata-700032, West Bengal, India; <sup>2</sup>Department of Pathology, Dunedin School of Medicine, University of Otago, Dunedin-9054, New Zealand; <sup>3</sup>National Research Centre for Growth and Development, New Zealand. Aniruddha Chatterjee - Email: aniruddha.chatterjee@otago.ac.nz; Phone: +64-210701558; \*Corresponding author

Received October 14, 2011; Accepted October 18, 2011; Published October 31, 2011

## Abstract:

Serine proteases are an abundant class of enzymes that are involved in a wide range of physiological processes and are classified into clans sharing structural homology. The active site of the subtilisin-like clan contains a catalytic triad in the order Asp, His, Ser (S8 family) or a catalytic tetrad in the order Glu, Asp and Ser (S53 family). The core structure and active site geometry of these proteases is of interest for many applications. The aim of this study was to investigate the structural properties of different S8 family serine proteases from a diverse range of taxa using molecular modeling techniques. In conjunction with 12 experimentally determined three-dimensional structures of S8 family members, our predicted structures from an archaeon, protozoan and a plant were used for analysis of the catalytic core. Amino acid sequences were obtained from the MEROPS database and submitted to the LOOPP server for threading based structure prediction. The predicted structures were refined and validated using PROCHECK, SCRWL and MODELYN. Investigation of secondary structures and electrostatic surface potential was performed using MOLMOL. Encompassing a wide range of taxa, our structural analysis provides an evolutionary perspective on S8 family serine proteases. Focusing on the common core containing the catalytic site of the enzyme, the analysis presented here is beneficial for future molecular modeling strategies and structure-based rational drug design.

**Keywords:** serine protease; SB clan; S8 family; homology; threading; modeling.

## Background:

Serine proteases are among the most abundant enzymes in nature and are involved in a wide range of biological processes, including digestion, blood clotting, embryo development, signal transduction and the immune response [1]. This diverse class of enzymes is characterized by the presence of three critical amino acids in the catalytic site - serine (Ser), histidine (His) and aspartate (Asp), which is referred to as the catalytic triad. The presence of this catalytic triad in at least three structurally unrelated contexts indicates that it has evolved independently at least three times [2]. The MEROPS database (<http://merops.sanger.ac.uk/>) is an exceptional resource for information on proteases that employs a useful classification system [3]. Under this system, families of proteases are classified according to statistically significant similarities in the

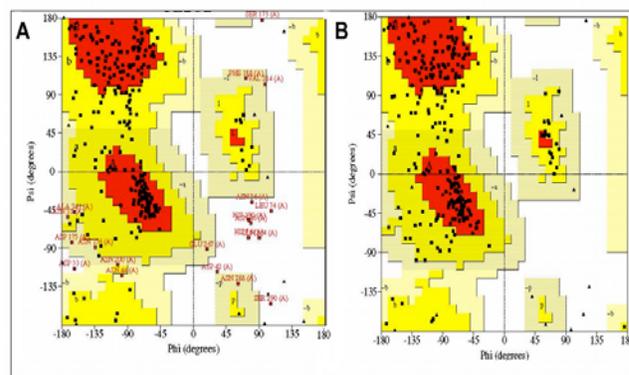
amino acid sequence. These protease families are further grouped into clans that have dissimilar amino acid sequences, but typically have structural homology and/or the same linear order of catalytic triad residues. The subtilisin-like (SB) clan of serine proteases is comprised of two distinct families, S8 (subtilases) and S53 (sedolisins). The catalytic residues are in the order Asp, His, Ser in the S8 family, but the S53 family contains a catalytic tetrad in the order Glu, Asp and Ser. Only the Ser in both families and the His in family S8 and the Glu in family S53 are in equivalent positions. The Ser in both families is found within a Gly-Thr-Ser-Xaa-Xaa-Xbb-Pro motif (where Xaa is an aliphatic amino acid and Xbb is a small amino acid) [3]. This study focuses solely on the S8 family, which contains the bacterial serine endopeptidase subtilisin and homologs such as proteinase K, kexin and furin. In addition to the distinct

catalytic triad, the protease structure typically comprises three layers of seven-stranded  $\beta$  sheets between two layers of  $\alpha$  helices. As is the case with all classical serine proteases, the catalytic mechanism involves nucleophilic attack by the Ser hydroxyl group on the carbonyl atom of the substrate, which is catalyzed by the His imidazole group as a general base (supported by a hydrogen bond to the Asp). The resulting tetrahedral intermediate is stabilized by an oxyanion hole. The His imidazole group transfers the proton to the amine leaving group and the tetrahedral intermediate breaks down to an acylenzyme intermediate. In a second addition-elimination reaction, the acylenzyme is attacked by a water molecule to form a second tetrahedral intermediate. With the protonation of the Ser by the His imidazole group, this intermediate breaks down and the C-terminus of the substrate is released. Most S8 family members are non-specific endopeptidases that preferentially cleave after hydrophobic residues, with some exceptions such as kexin and furin that cleave after dibasic amino acids [2]. S8 family serine proteases have a wide variety of biological functions, but they are particularly involved in nutrition and protein processing. Most notably, studies have shown that they are implicated in various diseases and are therefore a potential target for pharmacological intervention. For example, mutations in human proprotein convertase subtilisin-like kexin type 9 (*PCSK9*) have been associated with hypercholesterolemia [4] and other potential targets include virulence factors such as the *Streptococcus pyogenes* C5a peptidase [5]. The S8 family serine proteases have also been very popular candidates for protein engineering and directed evolution, which have yielded commercially successful results [6]. A structural analysis of these proteases, particularly the active site geometry, is of interest for many applications such as the above. Indeed, much of the information we have on this abundant class of enzymes has been deduced from experimental three dimensional (3D) structure analysis [7], but for many family members there is no structural data available. In such cases, *in silico* molecular modeling is a powerful tool to predict the 3D structure of a protein based on the amino acid sequence. The LOOPP homology modelling server (<http://loopp.org/>) is used to computationally build atomically detailed models using a set of candidates as structural templates and then it ranks the best models [8]. In the present study, we investigate the structural properties of different S8 family serine proteases using contemporary molecular modeling techniques. Here we present predicted structures from the extremophilic archaeon *Pyrococcus furiosus*, the malarial protozoan parasite *Plasmodium falciparum*, and Asian rice (*Oryza sativa*). By encompassing a wide range of taxa, our structural analysis provides an evolutionary perspective on this protease family. The potential for structure-based rational drug design and protein engineering, and future molecular modeling strategies is also discussed.

## Methodology:

Experimental structure data of SB serine proteases (S8 family) for 9 bacteria, 2 fungi, and 1 animal (Table S1, see supplementary material) were obtained from the Protein Data Bank (PDB, <http://www.rcsb.org/pdb>). Our in-house modeling software MODELYN [9] was used to analyze structural parameters, including the distance between Ca atoms of the catalytic triad. In addition to the experimental structures, amino acid sequences of S8 family SB serine proteases (Table

S1, see supplementary material) for 1 archaeon (*Pyrococcus furiosus*), 1 protozoan (*Plasmodium falciparum*), and 1 plant (*Oryza sativa*) were obtained from the MEROPS protease database (<http://merops.sanger.ac.uk>) in FASTA format [3]. Upon unsuccessful homology-based structure prediction using SWISS-MODEL (due to less than 30% sequence similarity with known experimental structures), these sequences were submitted to the LOOPP server [8] for threading based structure prediction. This analysis reported a ranked list of 10 possible structure predictions (Tables S2, S3 and S4, see Supplementary material) for each of the protease sequences, including match scores, sequence identity (%) and the extent of sequence coverage (%). Predicted structures were superposed with respect to a selected set of Ca atoms and a suitable starting scaffold was determined. Root mean square deviation (RMSD) values helped to identify the common segments, corresponding to the structurally conserved regions. For further refinement of the model, PROCHECK was used to check the distribution of  $\phi$ - $\psi$  dihedral angles and eliminate Ramachandran outliers. After side chain regeneration using SCRWL [10], the general structural parameters of the refined model, such as deviations of bond lengths, bond angles from standard values, overall atom clashscores and rotamer outliers were validated against experimental structure data using MOLPROBITY [11] and MODELYN. The ribbon structure and electrostatic potential surface of the structures were determined by MOLMOL [12]. To determine sequence conservation between species, CLUSTALW [13] was used for multiple sequence alignment and PEPSTATS [14] was used to analyze amino acid composition.



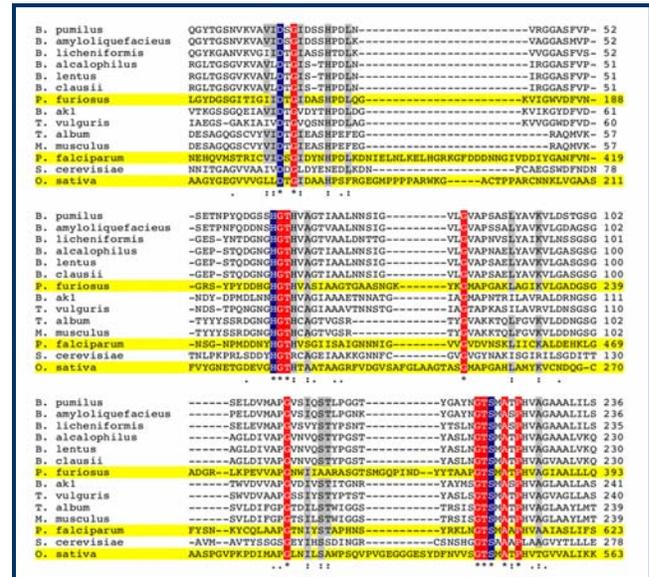
**Figure 1:** Representative Ramachandran plot of  $\phi$ - $\psi$  dihedral angles of a modeled SB serine protease structure before and after backbone refinement. PROCHECK was used to check the distribution of  $\phi$ - $\psi$  dihedral angles and eliminate Ramachandran outliers in each modeled protease structure (A, before; B, after refinement). Residues whose  $\phi$ - $\psi$  pairs fell outside the most favourable (red) and additional allowed (yellow) zones are annotated in red. Representative Ramachandran plots from the *P. falciparum* modeling are shown.

## Results and Discussion:

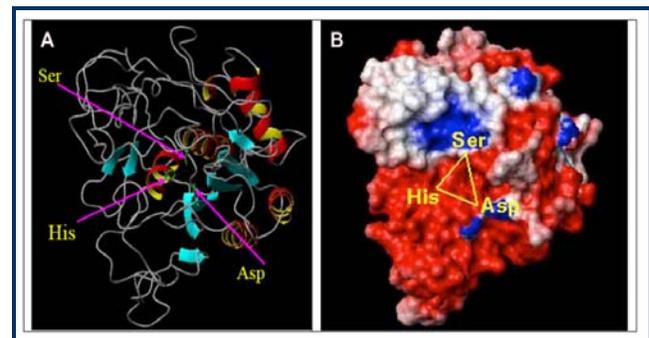
For the *P. furiosus*, *P. falciparum* and *O. sativa* proteases (Tables S2, S3, and S4, see supplementary material), the threading based server LOOPP predicted 10 structures for each sequence from 10 different PDB experimental structures (for *O. sativa*, only the first 5 probable structures were considered). The best matched structures for each showed high confidence scores

ranging from 4.7 to 9.0 and sequence identity ranging from 20 to 30%, with best length coverage between 61 and 93%. For *P. furiosus*, the matched structures were superposed with respect to a selected set of C $\alpha$  atoms (36%), with the structure 1EA7 having the best score of 4.73 (RMSD values were between 0.332 and 0.746 Å, which helped to identify common segments corresponding to structurally conserved regions). From these superposed structures, the variable loop regions were identified on the starting scaffold derived from 1EA7. For *P. falciparum*, structures were superposed with respect to selected C $\alpha$  atoms (41%) with the structure 1MEE having the highest score of 3.07 (RMSD values between 0.298 and 0.642 Å). For *O. sativa* were superposed with respect to selected C $\alpha$  atoms (34%), with the structure 1XF1 having the best score of 6.3 (RMSD values were between 0.431 and 0.612 Å). After this regularization, the overall backbone conformations of the predicted structures were measured to identify and eliminate Ramachandran outliers (Figure 1 and Table S5, see supplementary material). The general structural parameters of experimental and predicted structures were comparable (Table S6, see supplementary material), which validated the threading based modeling. Superposition of selected SB proteases on the representative X-ray structure (1MEE) of the *Bacillus pumilus* protease found that 22 to 44% of the C $\alpha$  atoms superposed with a RMSD below 1Å (Table S7, see supplementary material). The superposed structures showed highly conserved core structures with large variation in loops outside the core (data not shown). The mean Asp, His, Ser C $\alpha$  atom distances in the experimentally determined structures were  $7.4 \pm 0.06$ ,  $8.7 \pm 0.04$  and  $10.0 \pm 0.03$  Å respectively. The small standard deviations indicated that the structural environment around the catalytic triad was highly conserved. The mean values of the predicted structures were  $7.3 \pm 0.09$ ,  $9.0 \pm 0.30$  and  $10.3 \pm 0.19$  Å respectively, which was in good agreement with the values above. Multiple sequence alignment (Figure 2) confirmed sequence conservation of the catalytic triad residues at Asp32, His64, and Ser221. In addition to the previously described Gly-Thr-Ser-Xaa-Xaa-Xbb-Pro motif (where Xaa is an aliphatic amino acid and Xbb is a small amino acid) at Ser221 [3], our analysis indicated a Xaa-Xaa-Asp-Xcc-Gly-Xaa motif (where Xcc is a small polar amino acid) at the Asp32 and a His-Gly-Thr-Xdd-Xcc motif (where Xdd is either His or Arg) at His64. Other highly conserved residues contribute to important structural features, such as the Asn155 (not shown) involved in forming the oxyanion hole. The protease model from *P. furiosus* had 6  $\alpha$ -helical segments and 7 short  $\beta$ -sheets. The surface electrostatic potentials around the catalytic site were mostly negative with a patch of positive potential above the catalytic Ser residue (Figure 3). In comparison with other species, the *P. furiosus* protease had a higher proportion of aromatic residues (12.5%) and less smaller amino acids (59.3%). These distinctive features may be associated with increased stabilization and hyperthermophilic adaptation [15, 16], which could be utilized for protein engineering strategies. The protease model from *P. falciparum* had 9  $\alpha$ -helical segments and 6  $\beta$ -sheets. The pattern of surface electrostatic potential was very different from others analyzed, with the surface containing mostly neutral regions around the catalytic site (Figure 4). This malarial protease potentially mediates one or more of the serine protease activities associated with parasite erythrocyte invasion [17] and therefore the unique neutral regions around the catalytic site could be utilized in rational anti-malarial drug design. The

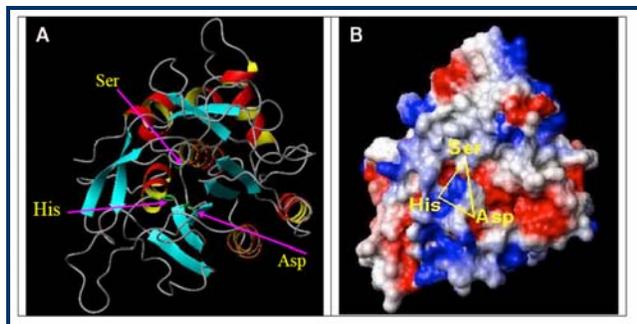
*Oryza sativa* SB protease model had 6  $\alpha$ -helical segments and 7  $\beta$ -sheets. The electrostatic potentials around the catalytic triad residues were mostly negative with a positive region near the His of the catalytic triad (Figure 5). This protease is expressed in the seed and seedling shoots of the rice plant [18] and could potentially have a role in signal transduction or a role in fungal resistance like its tobacco ortholog [19]. This protease is a potential candidate for protein engineering strategies to improve rice crops.



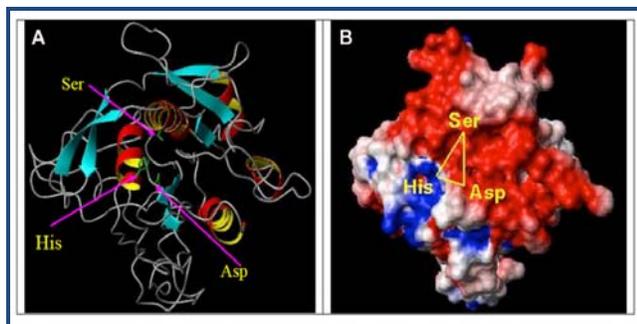
**Figure 2:** Multiple amino acid sequence alignment of SB serine proteases. CLUSTALW was used to align amino acid sequences of SB serine proteases with experimentally determined and predicted 3D structures (highlighted in yellow). Only the regions showing the conserved catalytic residues Asp (D), His (H), and Ser (S) are shown. Amino acid residues with 100% conservation (\*) between aligned sequences are either highlighted in blue (catalytic residues) or red (other). Other residues showing high (:) conservation (highlighted in gray) or medium (.) conservation are also indicated.



**Figure 3:** Modeled SB protease structure from *Pyrococcus furiosus* (PMDB ID: PM0075943). **A)** Ribbon model showing beta-sheets are shown in (light blue) and with arrow directed to C-terminus, alpha-helices (red and yellow), turn/loops (gray), and catalytic triad residue side chains (green sticks). **B)** Surface electrostatic potential model showing negative (red), positive (blue), and neutral (white) charges.



**Figure 4:** Modeled SB protease structure from *Plasmodium falciparum* (PMDB ID: PM0075941). **A)** Ribbon model showing beta-sheets are shown in (light blue) and with arrow directed to C-terminus, alpha-helices (red and yellow), turn/loops (gray), and catalytic triad residue side chains (green sticks). **B)** Surface electrostatic potential model showing negative (red), positive (blue), and neutral (white) charges.



**Figure 5:** Modeled SB protease structure from *Oryza sativa* (PMDB ID: PM0075942). **A)** Ribbon model showing beta-sheets are shown in (light blue) and with arrow directed to C-terminus, alpha-helices (red and yellow), turn/loops (gray), and catalytic triad residue side chains (green sticks). **B)** Surface electrostatic potential model showing negative (red), positive (blue), and neutral (white) charges.

The following predicted structures are available in the Protein Model Database (PMDB) (<http://mi.caspur.it/PMDB/>):

1. SB serine protease from *Pyrococcus furiosus* (PMDB ID: PM0075943)
2. SB serine protease from *Plasmodium falciparum* (PMDB ID: PM0075941)
3. SB serine protease from *Oryza sativa* (PMDB ID: PM0075942)

### Conclusion:

The structural geometry of the catalytic core was highly conserved across the diverse range of taxa analyzed and this was reflected in the sequence conservation motifs flanking the catalytic triad residues. Evolutionary divergence was exhibited by large variation in secondary structure features outside the core, differences in overall amino acid distribution, and unique surface electrostatic potential patterns between species. These features are probably associated with environmental

adaptation, subcellular localisation, and functional diversity of the different protease orthologs. Indeed, the higher proportion of aromatic residues in the extremophilic archaeon *P. furiosus* protease potentially provides stabilization [16] and the negatively charged residues around the catalytic site could also confer stabilization [15]. The mostly neutral surface electrostatic potential pattern around the catalytic site of the *P. falciparum* protease was very different from others analyzed. Significantly, as this protease is potentially associated with malarial erythrocyte invasion [17], the unique neutral regions around the catalytic site could be utilized in rational anti-malarial drug design [20].

### Acknowledgement:

Authors A.L. and C.M. are grateful for the funding and infrastructural support provided by the Indian Institute of Chemical Biology, Kolkata, West Bengal, India. Authors E.J.R. and A.C. acknowledge the support given by the Department of Pathology, University of Otago, Dunedin, the Health Research Council (E.J.R.), and the National Research Centre for Growth and Development (A.C.), New Zealand.

### References:

- [1] Hedstrom L *Curr Protoc Protein Sci.* 2002 **21**: 10 [PMID: 18429229]
- [2] Page MJ & Di Cera E, *Cell Mol Life Sci.* 2008 **65**: 1220 [PMID: 18259688]
- [3] Rawlings ND *et al. Nucleic Acids Res.* 2010 **38**: D227 [PMID: 19892822]
- [4] Abifadel M *et al. Nat Genet.* 2003 **34**: 154 [PMID: 12730697]
- [5] Cheng Q *et al. Infect Immun.* 2002 **70**: 2408 [PMID: 11953377]
- [6] Bryan PN. *Biochim Biophys Acta.* 2000 **1543**: 203 [PMID: 11150607]
- [7] Wright CS *et al. Nature.* 1969 **221**: 235 [PMID: 5763076]
- [8] Meller J & Elber R, *Proteins.* 2001 **45**: 241 [PMID: 11599028]
- [9] Mandal C, MODELYN: A molecular modelling program, version PC-1.0. Indian copyright No. 9/98. 1998.
- [10] Canutescu AA *et al. Protein Sci.* 2003 **12**: 2001 [PMID: 12930999]
- [11] Davis IW *et al. Nucleic Acids Res.* 2004 **32**: W615 [PMID: 15215462]
- [12] Koradi R *et al. J Mol Graph.* 1996 **14**: 29 [PMID: 8744573]
- [13] Thompson JD *et al. Nucleic Acids Res.* 1994 **22**: 4673 [PMID: 7984417]
- [14] Rice P *et al. Trends Genet.* 2000 **16**: 276 [PMID: 10827456]
- [15] Dym O *et al. Science.* 1995 **267**: 1344 [PMID: 17812611]
- [16] Siezen RJ *et al. Protein Eng.* 1991 **4**: 719 [PMID: 1798697]
- [17] Hackett F *et al. Mol Biochem Parasitol.* 1999 **103**: 183 [PMID: 10551362]
- [18] Yamagata H *et al. Biosci Biotechnol Biochem.* 2000 **64**: 1947 [PMID: 11055401]
- [19] Kinal H *et al. Plant Cell.* 1995 **7**: 677 [PMID: 7647561]
- [20] Turk B *Nat Rev Drug Discov.* 2006 **5**: 785 [PMID: 16955069]

Edited by P Kanguane

Citation: Laskar *et al.* Bioinformation 7(5): 239-245 (2011)

**License statement:** This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

## Supplementary material:

**Table S1:** Experimental structures and predicted structures of SB serine proteases (S8 family) across different taxa

Species	Structure	MEROPS ID
<b>Bacteria</b>		
<i>Bacillus pumilus</i>	PBD: 1MEE	MER001755
<i>Bacillus licheniformis</i>	PBD: 1AF4	MER000309
	PBD: 1BH6	MER000310
<i>Bacillus sp. ak1</i>	PBD: 1DBI	MER000332
<i>Bacillus alcalophilus</i>	PBD: 1AH2	MER000315
<i>Bacillus lentus</i>	PBD: 1GCI	MER001906
<i>Bacillus clausii</i>	PBD: 1MPT	MER001520
<i>Bacillus amyloliquefaciens</i>	PBD: 1SBT	MER000311
<i>Thermoactinomyces vulguris</i>	PBD: 1THM	MER000333
<b>Archaea</b>		
<i>Pyrococcus furiosus</i>	PMDB: PM0075943	MER017399
<b>Protozoa</b>		
<i>Plasmodium falciparum</i>	PMDB: PM0075941	MER004690
<b>Plantae</b>		
<i>Oryza sativa</i>	PMDB: PM0075942	MER054575
<b>Fungi</b>		
<i>Saccharomyces cerevisiae</i>	PBD: 1OT5	MER000364
<i>Tritirachium album</i>	PBD: 1IC6	MER000335
<b>Animalia</b>		
<i>Mus musculus</i>	PBD: 1P8J	MER000375

**Table S2:** LOOPP server results for secondary structure matches of *Pyrococcus furiosus* SB serine protease target sequence with known PDB structures

<i>Pyrococcus furiosus</i> (MER017399)							
PDB ID	Secondary Structure			Score	Sequence Identity (%)	Length (%)	
	Helical structure (%)	Extended (%)	Loops/Other (%)				
Target	15.83	16.02	68.15	-	-	-	
1EA7	22.98	18.45	58.58	4.729	26.14	55.02	
1S01	29.32	17.67	53.01	4.481	29.49	48.65	
1MEE	29.32	17.67	53.01	4.452	28.93	48.65	
1BH6	25.28	18.49	56.23	4.422	28.65	48.46	
1AU9	29.32	17.67	53.01	4.290	28.65	48.65	
1SBI	28.57	17.67	53.76	4.275	29.97	48.84	
1GCI	30.60	18.66	50.75	3.874	25.56	47.49	
1THM	27.41	18.52	54.07	3.696	28.21	48.46	
1IC6	25.48	22.39	52.12	3.043	23.01	47.30	
1GT9	23.12	20.42	56.46	1.932	15.30	61.58	

**Table S3:** LOOPP server results for secondary structure matches of *Plasmodium falciparum* SB serine protease target sequence with known PDB structures

<i>Plasmodium falciparum</i> (MER004690)						
PDB Name	Secondary Structure			Score	Sequence Identity (%)	Length (%)
	Helical structure (%)	Extended (%)	Loops /Other (%)			
Target	24.08	18.73	57.19	-	-	-
1MEE	29.93	17.15	52.92	8.961	91.30	27.95
1SBI	29.56	17.15	53.38	8.874	91.64	27.85
1SCJ	30.29	17.15	52.55	8.757	91.30	26.94
1BH6	29.67	17.95	52.38	8.700	90.97	25.93
1THM	27.34	17.95	54.68	8.626	88.96	29.93
1GCI	30.60	18.66	50.75	7.880	89.30	28.28
1R0R	0.00	0.00	100.00	7.610	90.97	26.94
1DBI	26.30	18.15	55.56	7.232	86.29	28.23
1EA7	26.00	19.00	55.00	6.715	96.99	22.07
1IC6	24.28	21.74	53.99	5.968	91.30	24.57

**Table S4:** LOOPP server results for secondary structure matches of *Oryza sativa* SB serine protease target sequence with known PDB structures

PDB Name	<i>Oryza sativa</i> (MER054575)			Score	Sequence Identity (%)	Length (%)
	Secondary Structure					
	Helical structure (%)	Extended (%)	Loops /Other (%)			
Target	16.77	16.67	66.26	-	-	-
1XF1	21.24	17.38	61.37	6.259	20.45	93.05
1V6C	19.20	26.00	54.80	3.364	18.71	84.87
1EA7	23.36	15.13	61.51	2.589	17.94	48.26
1THM	28.78	17.99	53.24	2.196	20.44	35.38
1SUE	31.70	17.74	50.57	2.041	14.86	49.69
1EA7	22.98	18.45	58.58	1.892	19.49	55.62
1GCI	30.60	18.66	50.75	1.833	19.58	39.26
1BH6	22.34	13.55	64.10	1.823	15.94	51.74
1BH6	29.67	17.95	52.38	1.764	20.08	40.49
1MEE	29.93	17.15	52.92	1.717	18.18	39.47

**Table S5:** Backbone refinement of the modeled SB proteases from *Pyrococcus furiosus*, *Plasmodium falciparum*, and *Oryza sativa*:

Threading model	$\phi$ - $\psi$ distribution in the regions of Ramachandran's plot			
	Number of residues (percentage)			
	Most favoured	Additional allowed	Generously allowed	Disallowed
<b><i>Pyrococcus furiosus</i></b>				
Before backbone refinement	162(56.1%)	91(31.5%)	23(8.0%)	13(4.5%)
After backbone refinement	178(61.8%)	110(38.0%)	0 (0.0%)	0 (0.0%)
<b><i>Plasmodium falciparum</i></b>				
Before backbone refinement	172(65.9%)	69(26.4%)	10 (3.8%)	10 (3.8%)
After backbone refinement	176(67.4%)	85(32.6%)	0 (0.0%)	0 (0.0%)
<b><i>Oryza sativa</i></b>				
Before backbone refinement	159(70.0%)	65(34.5%)	1 (0.2%)	2 (0.9%)
After backbone refinement	159(65.5%)	68(34.8%)	0 (0.0%)	0 (0.0%)

**Table S6:** Structural validation of the modeled SB proteases from *Pyrococcus furiosus*, *Plasmodium falciparum*, and *Oryza sativa*

Structural Model	All atom Clashscore (No/1000 atoms)	Rotamer Outliers (%)	RMSD of Bond Length (Å)	RMSD of Bond Angle (Degree)
X-ray structure (PDB ID: 1S01)	5.48	0.49	0.020	2.89
Threading model of <i>Pyrococcus furiosus</i> protease	5.35	4.69	0.021	3.31
X-ray structure (PDB ID: 1BH6)	4.49	1.46	0.022	3.39
Threading model of <i>Plasmodium falciparum</i> protease	5.65	3.94	0.021	3.17
X-ray structure (PDB ID: 1THM)	4.28	4.33	0.014	2.44
Threading model of <i>Oryza sativa</i> protease	4.1	2.5	0.021	3.51

**Table S7:** Structural parameters of experimentally determined and predicted 3D structures of SB serine proteases:

ID	Taxa	Species	Superposed of AA %	RMSD Å	Distances between the catalytic triad Å		
					(D-S)	(H-S)	(S-D)
1MEE	Bacteria	<i>B. pumilus</i>	100	0.000	7.5	8.6	10.1
1AF4	Bacteria	<i>B. licheniformis</i>	95.2	0.423	7.3	8.5	10.0
1BH6	Bacteria	<i>B. licheniformis</i>	95.25	0.472	7.1	8.1	9.7
1DBI	Bacteria	<i>B. ak1</i>	61.07	0.909	7.3	8.7	9.9
1AH2	Bacteria	<i>B. alcalophilus</i>	54.18	0.920	7.8	8.6	10.9
1GCI	Bacteria	<i>B. lentus</i>	81.09	0.674	7.2	8.5	9.9
1MPT	Bacteria	<i>B. clausii</i>	89.96	0.703	7.1	8.6	9.9
1SBT	Bacteria	<i>B. amyloliquefaciens</i>	100	0.000	7.1	9.1	10.2
1THM	Bacteria	<i>T. vulguris</i>	62.72	0.793	7.3	8.4	10.0
1OT5	Fungi	<i>S. cerevisiae</i>	24.94	0.849	6.8	8.7	10.2
1IC6	Fungi	<i>T. album</i>	39.06	0.735	9.2	9.9	10.0
1PJ8	Animalia	<i>M. musculus</i>	19.14	0.760	6.8	8.2	10.2

Mean $\pm$ SD of the C $\alpha$ distances between the triad residues					7.4 $\pm 0.06$	8.7 $\pm 0.04$	10.0 $\pm 0.03$
MER017399	Archaea	<i>P. furiosus</i>	21.92	0.947	7.2	8.5	9.9
MER004690	Protozoa	<i>P. falciparum</i>	43.95	0.776	7.6	8.6	10.2
MER054575	Plantae	<i>O. sativa</i>	36.69	0.841	7.2	9.8	10.8
Mean $\pm$ SD of the C $\alpha$ distances between the triad residues					7.3 $\pm 0.09$	9.0 $\pm 0.30$	10.3 $\pm 0.19$