Hindawi Publishing Corporation
Molecular Biology International
Volume 2012, Article ID 580965, 9 pages
doi:10.1155/2012/580965

Research Article

Three-Dimensional Molecular Modeling of a Diverse Range of SC Clan Serine Proteases

Aparna Laskar,[1] Aniruddha Chatterjee,[2,3] Somnath Chatterjee,[1] and Euan J. Rodger[2]

[1] *Infectious Diseases and Immunology Division, CSIR-Indian Institute of Chemical Biology, West Bengal, Kolkata 700032, India*
[2] *Department of Pathology, Dunedin School of Medicine, University of Otago, P.O. Box 913, Dunedin 9054, New Zealand*
[3] *National Research Centre for Growth and Development, University of Auckland, Auckland 1142, New Zealand*

Correspondence should be addressed to Euan J. Rodger, euan.rodger@otago.ac.nz

Received 25 July 2012; Revised 17 October 2012; Accepted 17 October 2012

Academic Editor: Alessandro Desideri

Serine proteases are involved in a variety of biological processes and are classified into clans sharing structural homology. Although various three-dimensional structures of SC clan proteases have been experimentally determined, they are mostly bacterial and animal proteases, with some from archaea, plants, and fungi, and as yet no structures have been determined for protozoa. To bridge this gap, we have used molecular modeling techniques to investigate the structural properties of different SC clan serine proteases from a diverse range of taxa. Either SWISS-MODEL was used for homology-based structure prediction or the LOOPP server was used for threading-based structure prediction. The predicted models were refined using Insight II and SCRWL and validated against experimental structures. Investigation of secondary structures and electrostatic surface potential was performed using MOLMOL. The structural geometry of the catalytic core shows clear deviations between taxa, but the relative positions of the catalytic triad residues were conserved. Evolutionary divergence was also 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. Encompassing a wide range of taxa, our structural analysis provides an evolutionary perspective on SC clan serine proteases.

1. Introduction

Serine proteases account for over a third of all known proteolytic enzymes and are involved in a range of physiological processes including digestion, immunity, blood clotting, fibrinolysis, reproduction, and protein folding [1]. The proteolytic mechanism of these proteases involves nucleophilic attack of the carbonyl atom of the substrate peptide bond by a catalytic serine (Ser) residue in the active site of the enzyme. In addition to the nucleophilic Ser residue, this reaction is dependent on other critical amino acids in the catalytic site such as an Aspartate (Asp) and a Histidine (His) that together form what is referred to as the catalytic triad (or a dyad in some cases) [2]. The presence of this catalytic triad in at least four distinct protein folds indicates the same mechanism evolved four separate times during evolution [3].

The MEROPS classification system (http://merops.sanger.ac.uk/) has grouped proteases into families 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 [4]. The SC clan of serine proteases is widely distributed across all taxa, and in contrast to other clans, it includes both endopeptidases and exopeptidases. At the core of all SC clan proteases is an α/β hydrolase fold, which typically consists of an eight-stranded β-sheet flanked by two or more α-helices. The α-helices contribute to substrate specificity, and the curvature of the β-sheet may also affect interactions with the substrate [1]. The α/β hydrolase fold is a common hydrolytic enzyme structure and is found in many other enzymes such as lipases, peroxidases, and esterases [5]. The SC clan has the same classical serine protease catalytic triad residue formation as clans SB and PA, but with the amino acid sequence order of Ser, Asp, and His. Typically, these residues are confined to the C-terminal region within about 130 residues. The proteolytic mechanism is initiated by the nucleophilic Ser158 (standard serine carboxypeptidase 2 *CBP2* numbering) hydroxyl group transferring a proton to the carbonyl of the peptide substrate. This reaction is catalyzed by the His413 acting as a general base, which is thought to be supported by a hydrogen bond to Asp361. The resulting tetrahedral intermediate breaks down to an acylenzyme intermediate, followed by the formation of a second tetrahedral intermediate. With the protonation of Ser158 by His413, the second tetrahedral intermediate breaks down and the cleaved substrate is released [2].

The five main families in this clan have distinct specificities and have different peptidase activities as represented by the archetypes prolyl oligopeptidase (S9 family), carboxypeptidase Y (S10 family), Xaa-Pro dipeptidyl-peptidase (S15 family), lysosomal Pro-Xaa carboxypeptidase (S28 family), and prolyl aminopeptidase (S33 family).

Because of their abundance and biological significance, the S9 serine protease family has been the most intensively studied. The proteases of this family are up to three times larger than their classic serine protease counterparts, trypsin and subtilisin (25–30 kDa). Many members hydrolyze the peptide bond on the C-terminal side of proline, but the exceptions include oligopeptidase B, which recognizes arginine or lysine, and acylaminoacyl peptidase, which is a cytoplasmic omega-peptidase that releases an N-acylated amino acid [6]. Notably, the central tunnel of an unusual N-terminal β-propeller domain covers the catalytic site and selectively restricts access to oligopeptides of approximately 30 amino acids in length [7]. The S9 family appears to be important in the processing and degradation of peptide hormones, and, therefore, these proteases are important targets of drug design [8]. In humans, prolyl oligopeptidase is involved in several neurological conditions and control of blood pressure [9–11], dipeptidyl peptidase 4 in type 2 diabetes and cancer [12, 13], and acylaminoacyl peptidase in small-cell lung and renal cancer [14, 15]. Both prolyl oligopeptidase and oligopeptidase B seem to facilitate the virulence of protozoan parasites such as *Trypanosoma cruzi* and *Trypanosoma brucei*, which result in the trypanosome infections Chagas disease and sleeping sickness, respectively [16, 17]. Dipeptidyl peptidase 4 contributes to the pathogenicity of *Porphyromonas gingivalis*, the gram-negative bacteria associated with periodontitis [18].

Proteases in the S10 family are serine carboxypeptidases, which cleave C-terminal peptide bonds. They generally prefer hydrophobic amino acids but exhibit broad substrate specificity. In contrast with most other serine proteases, which are typically active at neutral/alkaline pH, family S10 proteases maintain catalytic activity in an acidic environment [19]. This family mostly contributes to proteolytic degradation and protein processing within specific cellular compartments such as vacuoles in fungi and plants (carboxypeptidase Y) and lysosomes in animals (serine carboxypeptidase A) [20, 21]. Members of the S15 family selectively cleave Xaa-Pro, in which Xaa is an N-terminal amino acid. In *Lactobacillus helveticus*, which is used for commercial cheese-making, Xaa-Pro dipeptidyl-peptidase is involved in the casein-degradation pathway, providing essential amino acids for the bacteria [22]. The S28 proteases are a distinct family of eukaryotic carboxypeptidases that selectively cleave a Pro-Xaa bond, in which Xaa is a C-terminal amino acid. The human lysosomal Pro-Xaa carboxypeptidase (*PRCP*) is thought to be involved in regulating blood pressure by inactivating angiotensin II [23]. Dipeptidyl peptidase 2, which has a similar substrate specificity to dipeptidyl peptidase 4 of the S9 family, is essential for maintaining lymphocytes and fibroblasts in a quiescent state [24]. Proteases in the S33 family are prolyl aminopeptidases, which preferentially cleave an N-terminal proline residue peptide bond. Many of the bacteria and fungi that produce prolyl aminopeptidases are pathogenic and have therefore been proposed as a viable drug target [25].

Table 1: Experimental structures and predicted structures of SC serine proteases across different taxa.

Species	Structure	MEROPS ID
Bacteria		
Xanthomonas campestris	PDB: 1AZW	MER002678
Myxococcus xanthus	PDB: 2BKL	MER005694
Streptomyces lividans	PDB: 1A88	MER026339
Archaea		
Thermoplasma acidophilum	PDB: 1MU0	MER003537
Aeropyrum pernix	PDB: 1VE7	MER005807
Fungi		
Saccharomyces cerevisiae	PDB: 1AC5	MER000413
Saccharomyces cerevisiae	PDB: 1WPX	MER002010
Animalia		
Homo sapiens	PDB: 1N1M	MER000401
	PDB: 1QFS	MER000392
Sus scrofa	PDB: 1ORW	MER028372
	PDB: 2BUC	MER028372
Plantae		
Arabidopsis thaliana	**PMDB: PM0078228**	MER045469
Protozoa		
Plasmodium falciparum	**PMDB: PM0078229**	MER035185

2. Material and Methods

Structural data for 3 bacterial, 2 archaeal, 2 fungal, and 4 animal SC clan serine protease structures (Table 1) were obtained from the Protein Data Bank (PDB, http://www.rcsb.org/pdb/). Our inhouse modeling software package MODELYN [26] was developed to perform customized molecular editing and *in silico* structural analysis. It has a set of powerful menus for batch processing commands leading to automated implementation of complicated tasks, including complete model building based on sequence homology and batch processing of replacement mutations. ANALYN [26] is an ancillary protein sequence analysis program that assists MODELYN by analyzing homologous sequences and formulating the strategy for model building. In addition to the experimental structures, amino acid sequences of SC serine proteases (Table 1) for 1 plant (*Arabidopsis thaliana*) and 1 protozoan (*Plasmodium falciparum*) were obtained from the MEROPS protease database (http://merops.sanger.ac.uk/) in

Table 2: SWISS-MODEL homology results of *Arabidopsis thaliana* SC serine protease target sequence with known PDB structures.

PDB ID	Resolution (Å)	*R*-value	Score (bits)	Expect value	AA identity (%)
2BKLB	1.50	0.161	47.4	9×10^{-7}	30
2BKLA	1.50	0.161	47.4	9×10^{-7}	30
1YR2A	1.80	0.162	40.8	9×10^{-6}	34
1QFSA	2.00	0.201	39.7	9×10^{-5}	30
1VZ2A	2.20	0.165	39.7	2×10^{-2}	30

FASTA format [27]. These sequences were initially submitted to SWISS-MODEL for homology-based structure prediction [28]. If a sequence had less than 25% sequence similarity with known experimental structures, these sequences were then submitted to the LOOPP server [29] for threading-based structure prediction as previously described [30, 31]. This analysis reported a ranked list of possible structure predictions 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 Cα atoms on the structure with the highest match score, and a suitable starting scaffold was determined using MODELYN. Root mean square deviation (RMSD) values helped to identify the common segments, corresponding to the structurally conserved regions. The initial structures were refined using the DISCOVER and ANALYSIS modules within the software package Insight II [32] through energy minimization and molecular dynamics. The side chains were regenerated using SCRWL [33], and the overall structure was energy minimized. The SCWRL software package was used for prediction of protein side-chains of a fixed backbone, using graph theory to solve the combinatorial problem (details of the structure refinement are given in the Supplementary Material available online at doi:10.1155/2012/580965). PROCHECK was used to check the distribution of φ-ψ dihedral angles and identify Ramachandran outliers [34]. The CHARMM module within Insight II was used to apply dihedral constraints in these segments. MOLPROBITY [35] and MODELYN were used to validate the structural models against experimental structure data. MOLPROBITY provides all-atom contact analysis and gives quantitative information on the steric interactions (H-bond and van der Waals contacts) at the interfaces between components. This program is widely used for quality validation of three-dimensional (3D) protein structures by measuring deviations of bond lengths, bond angles from standard values, overall atom clashscores, and rotamer outliers. MODELYN was used to analyze other structural parameters, including the distance between Cα atoms of the catalytic triad. Verify3D [36], ProSA [37], and ERRAT [38] were also used to further assess the quality of the protease models. Verify3D analyzes the compatibility of the model against its own amino acid sequence. The Verify3D score (the sum of scores for individual residues using a 21-residue sliding window) is normalized to the length of the sequence: $\log_2$(Verify3D score/L^2) [39]. ProSA calculates an overall quality score (Z score) of a model in comparison to a range of characteristics expected for native protein structures. ERRAT analyzes the statistics of nonbonded interactions between different atom types (9-residue sliding window) and provides an overall quality factor that is expressed as the percentage of the protein for which the calculated error value falls below the 95% threshold. The ribbon structure and electrostatic potential surface of the structures were determined by MOLMOL [40]. To determine sequence conservation between species, ClustalW [41] was used for multiple sequence alignment. For each sequence, PEPSTATS [42] was used to determine the molar percentage of each amino acid physicochemical class. A flowchart of the modeling and structure refinement strategy has been included as Supplementary Figure S1.

3. Results

3.1. Modeling of Protease Structures. The plant protease from *A. thaliana* had significant homology with proteases of known experimental structure for successful structure prediction using SWISS-MODEL. The homology model was essentially built on the structures 2BKL, 1YR2, 1VZ2 (prolyl oligopeptidases from *Myxococcus xanthus*, *Novosphingobium capsulatum*, and *Sus scrofa* resp.), and 1QGS (an spsA glycosyltransferase from *Bacillus subtilis*), with sequence identity ranging from 30% to 34% (Table 2). Homology-based structure prediction for the *P. falciparum* protease was unsuccessful due to insufficient sequence similarity with known experimental structures. The amino acid sequence was then submitted to the LOOPP server for threading-based structure prediction, which yielded a list of 14 different PDB experimental structures that matched the protease sequence. The matching structures showed good confidence scores ranging from 2.7 to 3.5, sequence identity ranging from 13% to 19%, with best length coverage between 86% and 100% (Table 3). The matched structures were superposed with respect to a selected set of *P. falciparum* protease Cα atoms (43% superposition), with the structure 1U2E (an MhpC C–C bond hydrolase from *Escherichia coli*) having the best score of 3.5 (RMSD values were between 0.332 and 0.564 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 1U2E. Structural refinement of the two models using Insight II and SCRWL is provided in detail as Supplementary Material (additional file 1). The overall backbone conformations of the predicted structures were measured, and Ramachandran outliers were corrected for by applying dihedral constraints in these segments (Table 4). The general structural parameters and the overall quality of the final refined model were compared to experimental structure data (Table 5). The physical parameters were

Table 3: LOOPP server results for secondary structure matches of *Plasmodium falciparum* SC serine protease target sequence with known PDB structures.

PDB ID	Secondary structure			Score	Sequence identity (%)	Length (%)
	Helical structure (%)	Extended (%)	Loops/other (%)			
Target	36.20	19.00	44.80	—	—	—
1U2E	39.44	18.31	42.25	3.547	19.46	100.00
1J1I	41.80	19.53	38.67	3.484	14.88	97.29
1UKS	40.15	17.84	42.01	3.353	18.35	98.64
1C4X	40.71	16.43	42.86	3.294	14.68	98.64
1CQW	32.98	18.85	48.17	3.207	15.67	85.97
1BN6	32.46	19.37	48.17	3.205	15.67	85.97
1MJS	32.53	18.69	48.79	3.158	18.64	99.55
1BN7	32.98	18.47	43.55	3.042	17.73	99.55
1B6G	33.44	14.75	51.80	3.036	15.84	100.00
1FJ2	29.65	19.91	50.44	2.999	19.71	94.12
1JJF	32.40	20.40	47.20	2.997	16.29	100.00
1EHY	38.43	17.08	44.48	2.937	14.55	99.55
1IMJ	30.73	21.95	47.32	2.933	12.79	92.31
1A85	42.28	15.81	41.91	2.688	17.06	95.48

Table 4: Backbone refinement of the modeled SC proteases from *Arabidopsis thaliana* and *Plasmodium falciparum*.

Structural model	φ-ψ distribution in the regions of Ramachandran plot Number of residues (percentage)			
	Most favoured	Additionally allowed	Generously allowed	Disallowed
A. thaliana				
Before backbone refinement	190 (78.8%)	43 (17.8%)	4 (1.7%)	4 (1.7%)
After backbone refinement	194 (80.8%)	46 (19.2%)	0 (0.0%)	0 (0.0%)
P. falciparum				
Before backbone refinement	132 (65.7%)	53 (26.4%)	10 (5.0%)	6 (3.0%)
After backbone refinement	129 (65.5%)	68 (34.5%)	0 (0.0%)	0 (0.0%)

Table 5: Structural validation of the modeled SC proteases from *Arabidopsis thaliana* and *Plasmodium falciparum.*

Structural model	All-atom clashscore (No/1000 atoms)	Rotamer outliers (%)	RMSD of bond length (Å)	RMSD of bond angle (degree)
X-ray structure (1YR2)	3.76	2.40	0.240	2.76
Homology model of *A. thaliana* protease	5.37	2.14	0.270	3.70
X-ray structure (1U2E)	23.50	9.62	0.013	2.27
Threading model of *P. falciparum* protease	11.30	4.50	0.028	3.26
	Average Verify3D-1D score	Normalized 3D profile score ($\log_2$(Verify3D/L^2)	ProSA *Z*-score	ERRAT quality factor (%)
X-ray structure (1YR2)	0.46	−10.53	−9.68	95.2
Homology model of *A. thaliana* protease	0.34	−9.66	−7.55	88.9
X-ray structure (1U2E)	0.42	−10.86	−9.88	93.7
Threading model of *P. falciparum* protease	0.29	−9.58	−4.53	75.5

Table 6: Structural parameters of experimentally determined and predicted 3D structures of SC serine proteases.

ID	Taxa	Species	Superposed of AA %	RMSD Å	Distances between the catalytic triad Å		
					(D-H)	(H-S)	(S-D)
1AZW	Bacteria	*X. campestris*	10.20	1.330	4.4	7.8	10.5
2BKL	Bacteria	*M. xanthus*	27.60	0.726	4.6	8.3	10.6
1A88	Bacteria	*S. lividans*	12.30	1.080	4.6	7.9	10.6
1VE7	Archaea	*A. pernix*	15.76	1.083	4.7	8.4	10.4
1MU0	Archaea	*T. acidophilum*	17.00	1.013	4.3	7.8	10.1
1AC5	Fungi	*S. cerevisiae*	17.86	1.441	4.8	7.7	10.8
1WPX	Fungi	*S. cerevisiae*	12.30	1.168	4.6	7.4	10.4
1N1M	Animalia	*H. sapiens*	16.29	1.163	4.8	7.7	10.7
1QFS	Animalia	*S. scrofa*	100	0.000	4.6	8.1	10.5
1ORW	Animalia	*S. scrofa*	18.29	1.092	4.8	7.7	10.7
2BUC	Animalia	*S. scrofa*	17.15	1.108	4.8	7.7	10.7
Mean ± SD of the Cα distances between the triad residues					4.6 ± 0.03	7.9 ± 0.06	10.5 ± 0.04
PM0078228	Plantae	*A. thaliana*	16.7	1.021	4.5	8.2	11.2
PM0078229	Protozoa	*P. falciparum*	69.3	0.686	4.6	8.1	10.5
Mean ± SD of the Cα distances between the triad residues					4.6 ± 0.01	8.2 ± 0.01	10.9 ± 0.25

comparable between the experimental and predicted structures. The good scores provided by Verify3D, ProSA, and ERRAT further validated the overall quality of the refined models from *A. thaliana* (PMDB: PM0078228) and *P. falciparum* (PMDB: PM0078229).

3.2. Catalytic Core Geometry. Superposition of the *A. thaliana* and *P. falciparum* proteases on the representative 1U2E protease structure found that 17% to 69% of the Cα atoms superposed with an RMSD below 1.1 Å (Table 6). In comparison, X-ray protease structures had 10% to 28% of the Cα atoms superposed with an RMSD below 1.5 Å (Table 6). The superposed structures have a common core structure with large variation in loops outside the core (Figure 1). The Cα atom distances of Asp to His, His to Ser, and Asp to Ser averaged over the experimentally determined structures were 4.6 ± 0.03, 7.9 ± 0.06, and 10.5 ± 0.04 Å, respectively (Table 6). The small standard deviations (SDs) indicated that the structural environment around the catalytic triad was highly conserved. Averaged over the predicted structures, the Cα atom distances between the catalytic triad residues were 4.6 ± 0.01, 8.2 ± 0.01, and 10.9 ± 0.25 Å, respectively, in good agreement with the values averaged over the experimental structures. Multiple sequence alignment (Figure 2) confirmed sequence conservation of the catalytic triad residues at Ser158 Asp361 His413 (serine carboxypeptidase 2 numbering). Among the sequences analyzed, the highly conserved amino acids Gly156 and Gly160 had the occupancy percentage of 70% and 77%, respectively, which has been previously described [43]. In addition, Gly161, Asp315, Val317, and Gly343 were all highly conserved with an occupancy percentage of 75% in the S9 family member sequences analyzed. As confirmed in other serine proteases, such residues may confer stabilization of the catalytic site via a hydrogen-bonding interaction [44, 45]. By incorporating an evolutionarily diverse range of SC serine proteases, our analysis indicates that although the core structures deviated considerably during evolution, the relative positions of the catalytic triad Cα atoms maintained very close relative distances and were potentially stabilized by other highly conserved residues.

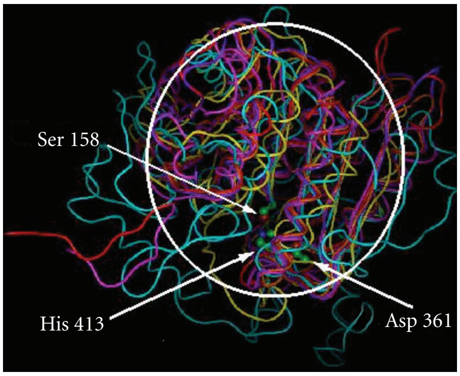
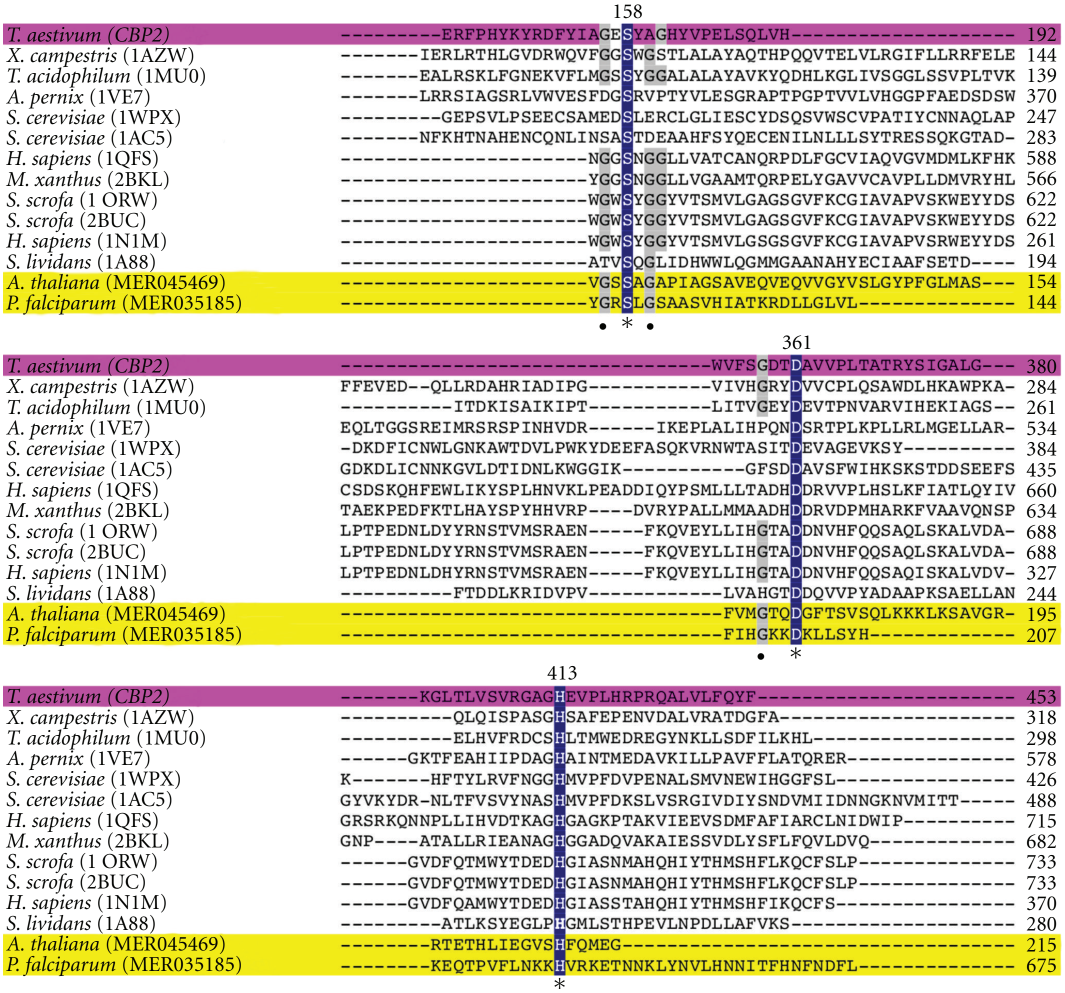


Figure 1: Superposed structures of X-ray and modeled structures of the selected proteases of the SC clan. Structures of the plant (PM0078228, *Arabidopsis thaliana*, purple) and protozoan (PM0078229, *Plasmodium falciparum*, yellow) SC proteases were superposed with the animal (1QFS, *Sus scrofa*, red), archaeon (1VE7, *Aeropyrum pernix*, magenta), fungal (1WPX, *Saccharomyces cerevisiae*, cyan), and bacterial (2BKL, *Myxococcus xanthus*, orange) X-ray structures. The catalytic triad residues (Ser, Asp, His; *CBP2* residue numbering used as a standard reference) are shown in ball and stick models, and the core regions of the structures are indicated by the white circle.

3.3. Structural Analysis. The catalytic core of all SC clan proteases bears an α/β hydrolase fold, which typically consists



```
                                                  158
T. aestivum (CBP2)        ---------ERFPHYKYRDFYIAGESYAGHYVPELSQLVH-------------------- 192
X. campestris (1AZW)      -------IERLRTHLGVDRWQVFGGSWGSTLALAYAQTHPQQVTELVLRGIFLLRRFELE 144
T. acidophilum (1MU0)     -------EALRSKLFGNEKVFLMGSSYGGALALAYAVKYQDHLKGLIVSGGLSSVPLTVK 139
A. pernix (1VE7)          -------LRRSIAGSRLVWVESFDGSRVPTYVLESGRAPTPGPTVVLVHGGPFAEDSDSW 370
S. cerevisiae (1WPX)      ---------GEPSVLPSEECSAMEDSLERCLGLIESCYDSQSVWSCVPATIYCNNAQLAP 247
S. cerevisiae (1AC5)      -------NFKHTNAHENCQNLINSASTDEAAHFSYQECENILNLLLSYTRESSQKGTAD- 283
H. sapiens (1QFS)         ----------------------NGGSNGGLLVATCANQRPDLFGCVIAQVGVMDMLKFHK 588
M. xanthus (2BKL)         ----------------------YGGSNGGLLVGAAMTQRPELYGAVVCAVPLLDMVRYHL 566
S. scrofa (1 ORW)         ----------------------WGWSYGGYVTSMVLGAGSGVFKCGIAVAPVSKWEYYDS 622
S. scrofa (2BUC)          ----------------------WGWSYGGYVTSMVLGAGSGVFKCGIAVAPVSKWEYYDS 622
H. sapiens (1N1M)         ----------------------WGWSYGGYVTSMVLGSGSGVFKCGIAVAPVSRWEYYDS 261
S. lividans (1A88)        ----------------------ATVSQGLIDHWWLQGMMGAANAHYECIAAFSETD---- 194
A. thaliana (MER045469)   ----------------------VGSSAGAPIAGSAVEQVEQVVGYVSLGYPFGLMAS--- 154
P. falciparum (MER035185) ----------------------YGRSLGSAASVHIATKRDLLGLVL-------------- 144
                                                 • * •
                                                                 361
T. aestivum (CBP2)        ---------------------------------WVFSGDTDAVVPLTATRYSIGALG--- 380
X. campestris (1AZW)      FFEVED--QLLRDAHRIADIPG-----------VIVHGRYDVVCPLQSAWDLHKAWPKA- 284
T. acidophilum (1MU0)     ----------ITDKISAIKIPT-----------LITVGEYDEVTPNVARVIHEKIAGS-- 261
A. pernix (1VE7)          EQLTGGSREIMRSRSPINHVDR------IKEPLALIHPQNDSRTPLKPLLRLMGELLAR- 534
S. cerevisiae (1WPX)      -DKDFICNWLGNKAWTDVLPWKYDEEFASQKVRNWTASITDEVAGEVKSY---------- 384
S. cerevisiae (1AC5)      GDKDLICNNKGVLDTIDNLKWGGIK-----------GFSDDAVSFWIHKSKSTDDSEEFS 435
H. sapiens (1QFS)         CSDSKQHFEWLIKYSPLHNVKLPEADDIQYPSMLLLTADHDDRVVPLHSLKFIATLQYIV 660
M. xanthus (2BKL)         TAEKPEDFKTLHAYSPYHHVRP----DVRYPALLMMAADHDDRVDPMHARKFVAAVQNSP 634
S. scrofa (1 ORW)         LPTPEDNLDYYRNSTVMSRAEN-----FKQVEYLLIHGTADDNVHFQQSAQLSKALVDA- 688
S. scrofa (2BUC)          LPTPEDNLDYYRNSTVMSRAEN-----FKQVEYLLIHGTADDNVHFQQSAQLSKALVDA- 688
H. sapiens (1N1M)         LPTPEDNLDHYRNSTVMSRAEN-----FKQVEYLLIHGTADDNVHFQQSAQISKALVDV- 327
S. lividans (1A88)        ----------FTDDLKRIDVPV------------LVAHGTDDQVVPYADAAPKSAELLAN 244
A. thaliana (MER045469)   ----------------------------------FVMGTQDGFTSVSQLKKKLKSAVGR- 195
P. falciparum (MER035185) ----------------------------------FIHGKKDKLLSYH------------- 207
                                                               •  *
                                            413
T. aestivum (CBP2)        -------KGLTLVSVRGAGHEVPLHRPQALVLFQYF------------------------ 453
X. campestris (1AZW)      ----------QLQISPASGHSAFEPENVDALVRATDGFA--------------------- 318
T. acidophilum (1MU0)     ----------ELHVFRDCSHLTMWEDREGYNKLLSDFILKHL------------------ 298
A. pernix (1VE7)          ------GKTFEAHIIPDAGHAINTMEDAVKILLPAVFFLATQRER--------------- 578
S. cerevisiae (1WPX)      K-------HFTYLRVFNGGHMVPFDVPENALSMVNEWIHGGFSL---------------- 426
S. cerevisiae (1AC5)      GYVKYDR-NLTFVSVYNASHMVPFDKSLVSRGIVDIYSNDVMIIDNNGKNVMITT----- 488
H. sapiens (1QFS)         GRSRKQNNPLLIHVDTKAGHGAGKPTAKVIEEVSDMFAFIARCLNIDWIP---------- 715
M. xanthus (2BKL)         GNP----ATALLRIEANAGHGGADQVAKAIESSVDLYSFLFQVLDVQ------------- 682
S. scrofa (1 ORW)         ------GVDFQTMWYTDEDHGIASNMAHQHIYTHMSHFLKQCFSLP-------------- 733
S. scrofa (2BUC)          ------GVDFQTMWYTDEDHGIASNMAHQHIYTHMSHFLKQCFSLP-------------- 733
H. sapiens (1N1M)         ------GVDFQAMWYTDEDHGIASSTAHQHIYTHMSHFIKQCFS---------------- 370
S. lividans (1A88)        ---------ATLKSYEGLPHGMLSTHPEVLNPDLLAFVKS-------------------- 280
A. thaliana (MER045469)   --------RTETHLIEGVSHFQMEG----------------------------------- 215
P. falciparum (MER035185) --------KEQTPVFLNKKHVRKETNNKLYNVLHNNITFHNFNDFL-------------- 675
                                             *
```

Figure 2: Multiple amino acid sequence alignment of SC serine proteases. ClustalW was used to align amino acid sequences of SC serine proteases for which their structures were determined experimentally or predicted computationally (highlighted in yellow). Wheat serine carboxypeptidase 2 (*CBP2*, highlighted in magenta) is used as a standard reference for residue numbering. Only the regions showing the conserved catalytic residues Ser (S), Asp (D), and His (H) are shown (highlighted in blue). Other residues showing medium (•) conservation are highlighted in gray.

of a β-sheet flanked by two or more α-helices. Figures 3(a) and 3(b) are representative X-ray structures of an animal SC protease (1QFS, prolyl oligopeptidase from *Sus scrofa*), comprising 8 β-sheets and 12 α-helices. Figures 3(c) and 3(d) are representative X-ray structures of an archaeon SC protease (1VE7, acylpeptidase hydrolase from *Aeropyrum pernix*), comprising 30 β-sheets and 14 α-helices. The *A. thaliana* SC protease model had 8 β-sheets and 12 α-helices, with Ser120, Asp176, and His198 in separate turn/coil structures (Figure 3(e)). The electrostatic potentials around the Asp and His catalytic residues were mostly electronegative, and there was a patch of electropositive potential around the Ser residue of the catalytic triad (Figure 3(f)). The electronegative region in the catalytic site of the modeled protease could facilitate specificity by favoring positively charged C-terminal amino acid side chains at specific sites within the binding pocket. The *A. thaliana* protease had a higher proportion (>SD of the mean) of aliphatic residues (32%, molar percentage), compared to other species (see Table S1), which could influence stability of the enzyme at a wide range of temperatures [46]. According to MEROPS annotation (MER045469), this protease has been assigned to the S9 family, but it has an unknown function. Our homology model was essentially built on the structures 2BKL, 1YR2, and 1VZ2, which are prolyl oligopeptidases (S9 family) from *Myxococcus xanthus*, *Novosphingobium capsulatum*, and *Sus scrofa,* respectively. There have been 23 genes encoding prolyl oligopeptidase-like proteins identified in *A. thaliana* [47]. Although the function of most of these is unknown, there is some evidence that prolyl oligopeptidase is involved in seed development [48]. *A. thaliana* is a highly studied model organism, and mutational analysis of this protease would be useful to explore these features.

The protease model from *P. falciparum* had 7 β-sheets and 7 α-helices, with Ser124, Asp188, and His217 in separate turn/coil structures (Figure 3(g)). The surface electrostatic

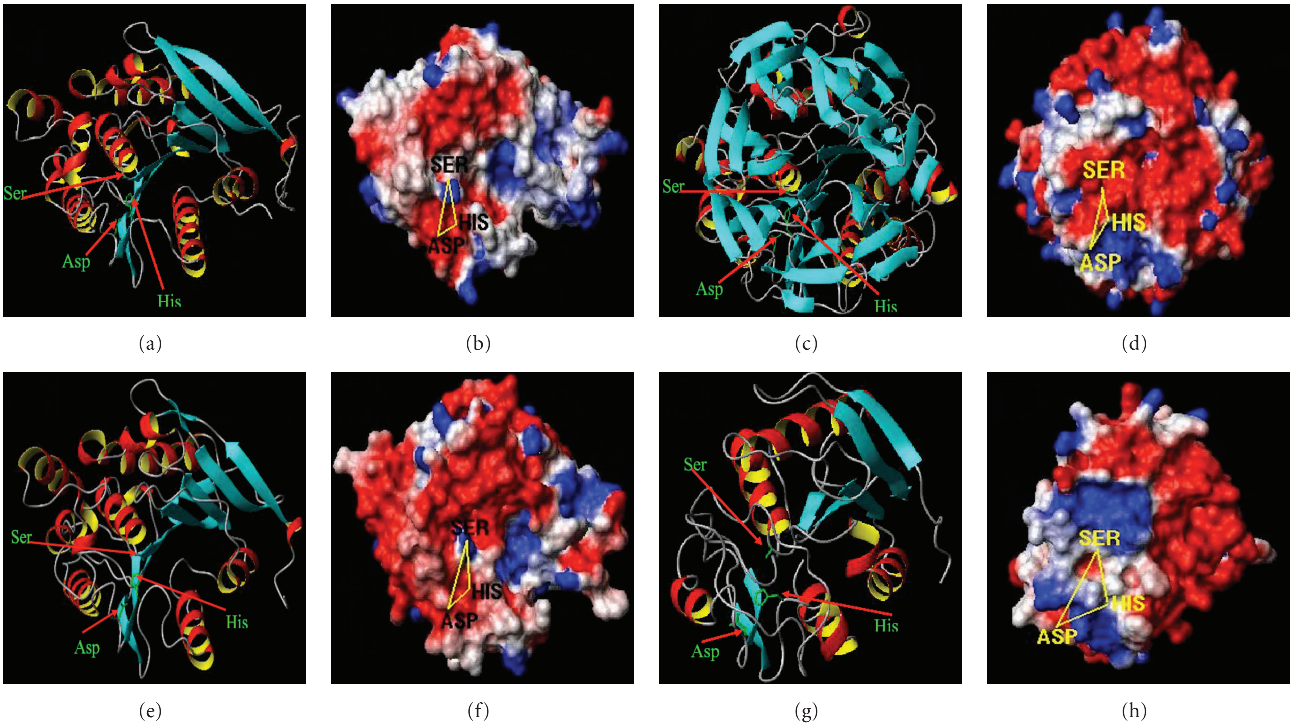


Figure 3: Representative X-ray SC protease structures and modeled SC protease structures from *Arabidopsis thaliana* and *Plasmodium falciparum*. Ribbon models of *S. scrofa,* 1QFS (a), *A. pernix,* 1VE7 (c), *A. thaliana* (e), and *P. falciparum* (g) SC protease structures show β-sheets with an arrow directed to the C-terminus (light blue), α-helices (red and yellow), turn/loops (gray), and catalytic triad residue side chains (green sticks). Surface electrostatic potential models of *S. scrofa,* 1QFS (b), *A. pernix,* 1VE7 (d), *A. thaliana,* PM0078228 (f), and *P. falciparum,* PM0078229 (h) SC protease structures show electronegative (red), electropositive (blue), and electroneutral (white) amino acid side chains. Electrostatic potential thresholds: $-1.4\,\mathrm{kT/e} < 0.0\,\mathrm{kT/e} < +1.4\,\mathrm{kT/e}$ (red → white → blue).

potentials around the catalytic site were very different to those of other clan members studied, with large patches of electropositive and electroneutral regions around the catalytic triad residues (Figure 3(h)). The largely electropositive catalytic site of this modeled protease suggests it favors a negatively charged substrate. The largely electroneutral regions possibly relax the stringency of the substrate binding, allowing for a number of different protein substrates. In comparison with the other species analyzed (see Table S1), the *P. falciparum* protease had a higher proportion (>SD of the mean) of polar residues (60%, molar percentage) and basic amino acids (16%), which indicates it could favor a more hydrophilic environment. Like the modeled protease from *A. thaliana*, this protease (MER035185) has also been assigned to the S9 family. Although the function of this protease is not known, it is of interest that both prolyl oligopeptidase and oligopeptidase B of the S9 family appear to facilitate the virulence of other protozoan parasites such as *Trypanosoma cruzi* and *Trypanosoma brucei* [16, 17]. Further investigation of substrate specificity and other properties contributing to it would be beneficial for functional analysis of this protease, as it could be a potential target for rational antimalarial drug design.

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

(1) SC serine protease from *Arabidopsis thaliana* (PMDB ID: PM0078228),

(2) SC serine protease from *Plasmodium falciparum* (PMDB ID: PM0078229).

4. Conclusion

In conjunction with 11 experimentally determined 3D protein structures, our analysis of predicted structures from a plant and a protozoan encompassed an evolutionarily diverse range of SC clan proteases. The structural geometry of the catalytic core clearly deviated considerably during evolution, but the relative positions of the catalytic triad residues were conserved, and other highly conserved residues possibly provide stabilization of the core. Evolutionary divergence was also 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 the diverse functions of the different protease orthologs. The modeled proteases from *A. thaliana* and *P. falciparum* appear to be prolyl oligopeptidases of the S9 family. Evidence indicates that prolyl oligopeptidase is involved in plant seed development [48] and facilitates the virulence of protozoan parasites [16, 17]. Further structural investigation of these proteases would be useful for protein engineering strategies and for rational drug design in the case of the *P. falciparum* protease.

Acknowledgments

A. Laskar and S. Chatterjee are grateful for the funding and infrastructural support provided by the Indian Institute of Chemical Biology, Kolkata, West Bengal, India. E. J. Rodger and A. Chatterjee gratefully acknowledge the support provided by Professor Ian Morison and the Department of Pathology, University of Otago, Dunedin, the Health Research Council (E. J. Rodger), and the National Research Centre for Growth and Development (A. Chatterjee), New Zealand. These entities did not have a role in study design, collection, analysis, or interpretation of data, writing the paper, or decision to submit paper for publication.

References

[1] M. J. Page and E. Di Cera, "Serine peptidases: classification, structure and function," *Cellular and Molecular Life Sciences*, vol. 65, no. 7-8, pp. 1220–1236, 2008.

[2] L. Hedstrom, "Serine protease mechanism and specificity," *Chemical Reviews*, vol. 102, no. 12, pp. 4501–4523, 2002.

[3] M. J. Page and E. Di Cera, "Evolution of peptidase diversity," *Journal of Biological Chemistry*, vol. 283, no. 44, pp. 30010–30014, 2008.

[4] N. D. Rawlings, A. J. Barrett, and A. Bateman, "MEROPS: the database of proteolytic enzymes, their substrates and inhibitors," *Nucleic Acids Research*, vol. 40, no. D1, pp. D343–D350, 2012.

[5] M. Holmquist, "Alpha/Beta-hydrolase fold enzymes: structures, functions and mechanisms," *Current Protein and Peptide Science*, vol. 1, no. 2, pp. 209–235, 2000.

[6] D. Rea and V. Fülöp, "Structure-function properties of prolyl oligopeptidase family enzymes," *Cell Biochemistry and Biophysics*, vol. 44, no. 3, pp. 349–365, 2006.

[7] V. Fülöp, Z. Böcskei, and L. Polgár, "Prolyl oligopeptidase: an unusual β-propeller domain regulates proteolysis," *Cell*, vol. 94, no. 2, pp. 161–170, 1998.

[8] L. Polgár, "The prolyl oligopeptidase family," *Cellular and Molecular Life Sciences*, vol. 59, no. 2, pp. 349–362, 2002.

[9] M. Maes, "Alterations in plasma prolyl endopeptidase activity in depression, mania, and schizophrenia: effects of antidepressants, mood stabilizers, and antipsychotic drugs," *Psychiatry Research*, vol. 58, no. 3, pp. 217–225, 1995.

[10] M. Maes, P. Monteleone, R. Bencivenga et al., "Lower serum activity of prolyl endopeptidase in anorexia and bulimia nervosa," *Psychoneuroendocrinology*, vol. 26, no. 1, pp. 17–26, 2001.

[11] W. R. Welches, K. B. Brosnihan, and C. M. Ferrario, "A comparison of the properties and enzymatic activities of three angiotensin processing enzymes: angiotensin converting enzyme, prolyl endopeptidase and neutral endopeptidase 24.11," *Life Sciences*, vol. 52, no. 18, pp. 1461–1480, 1993.

[12] T. Vilsbøll, T. Krarup, C. F. Deacon, S. Madsbad, and J. J. Holst, "Reduced postprandial concentrations of intact biologically active glucagon-like peptide 1 in type 2 diabetic patients," *Diabetes*, vol. 50, no. 3, pp. 609–613, 2001.

[13] B. Pro and N. H. Dang, "CD26/dipeptidyl peptidase IV and its role in cancer," *Histology and Histopathology*, vol. 19, no. 4, pp. 1345–1351, 2004.

[14] S. L. Naylor, A. Marshall, C. Hensel, P. F. Martinez, B. Holley, and A. Y. Sakaguchi, "The DNF15S2 locus at 3p21 is transcribed in normal lung and small cell lung cancer," *Genomics*, vol. 4, no. 3, pp. 355–361, 1989.

[15] R. Erlandsson, F. Boldog, B. Persson et al., "The gene from the short arm of chromosome 3, at D3F15S2, frequently deleted in renal cell carcinoma, encodes acylpeptide hydrolase," *Oncogene*, vol. 6, no. 7, pp. 1293–1295, 1991.

[16] J. M. Santana, P. Grellier, J. Schrével, and A. R. L. Teixeira, "A Trypanosoma cruzi-secreted 80 kDa proteinase with specificity for human collagen types I and IV," *Biochemical Journal*, vol. 325, no. 1, pp. 129–137, 1997.

[17] R. E. Morty, J. D. Lonsdale-Eccles, J. Morehead et al., "Oligopeptidase B from Trypanosoma brucei, a new member of an emerging subgroup of serine oligopeptidases," *Journal of Biological Chemistry*, vol. 274, no. 37, pp. 26149–26156, 1999.

[18] Y. Kumagai, K. Konishi, T. Gomi, H. Yagishita, A. Yajima, and M. Yoshikawa, "Enzymatic properties of dipeptidyl aminopeptidase IV produced by the periodontal pathogen Porphyromonas gingivalis and its participation in virulence," *Infection and Immunity*, vol. 68, no. 2, pp. 716–724, 2000.

[19] K. Breddam, "Serine carboxypeptidases. A review," *Carlsberg Research Communications*, vol. 51, no. 2, pp. 83–128, 1986.

[20] J. A. Endrizzi, "2.8-Å structure of yeast serine carboxypeptidase," *Biochemistry®*, vol. 33, no. 37, pp. 11106–11120, 1994.

[21] D. I. Liao and S. J. Remington, "Structure of wheat serine carboxypeptidase II at 3.5-Å resolution. A new class of serine proteinase," *Journal of Biological Chemistry*, vol. 265, no. 12, pp. 6528–6531, 1990.

[22] G. U. Yüksel and J. L. Steele, "DNA sequence analysis, expression, distribution, and physiological role of the Xaa-prolyldipeptidyl aminopeptidase gene from Lactobacillus helveticus CNRZ32," *Applied Microbiology and Biotechnology*, vol. 44, no. 6, pp. 766–773, 1996.

[23] C. E. Odya, D. V. Marinkovic, and K. J. Hammon, "Purification and properties of prolylcarboxypeptidase (angiotensinase C) from human kidney," *Journal of Biological Chemistry*, vol. 253, no. 17, pp. 5927–5931, 1978.

[24] D. A. Mele, P. Bista, D. V. Baez, and B. T. Huber, "Dipeptidyl peptidase 2 is an essential survival factor in the regulation of cell quiescence," *Cell Cycle*, vol. 8, no. 15, pp. 2425–2434, 2009.

[25] L. Zhang, Y. Jia, L. Wang, and R. Fang, "A proline iminopeptidase gene upregulated in planta by a LuxR homologue is essential for pathogenicity of Xanthomonas campestris pv. campestris," *Molecular Microbiology*, vol. 65, no. 1, pp. 121–136, 2007.

[26] C. Mandal, "MODELYN: a molecular modelling program, version PC-1.0. Indian copyright No. 9/98," *Copyright Office, Government of India*, 1998.

[27] N. D. Rawlings, A. J. Barrett, and A. Bateman, "MEROPS: the peptidase database," *Nucleic Acids Research*, vol. 38, no. 1, Article ID gkp971, pp. D227–D233, 2009.

[28] T. Schwede, J. Kopp, N. Guex, and M. C. Peitsch, "SWISS-MODEL: an automated protein homology-modeling server," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3381–3385, 2003.

[29] J. Meller and R. Elber, "Linear programming optimization and a double statistical filter for protein threading protocols," *Proteins*, vol. 45, no. 3, pp. 241–261, 2001.

[30] A. Laskar, E. Rodger, A. Chatterjee, and C. Mandal, "Modeling and structural analysis of evolutionarily diverse S8 family

serine proteases," *Bioinformation*, vol. 7, no. 5, pp. 239–245, 2011.

[31] A. Laskar, E. Rodger, A. Chatterjee, and C. Mandal, "Modeling and structural analysis of PA clan serine proteases," *BMC Research Notes*, vol. 5, no. 1, p. 256, 2012.

[32] *Insight II Modeling Environment*, Molecular Simulations, San Diego, Calif, USA, 2005.

[33] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack Jr, "A graph-theory algorithm for rapid protein side-chain prediction," *Protein Science*, vol. 12, no. 9, pp. 2001–2014, 2003.

[34] R. A. Laskowski, "PROCHECK: a program to check the stereochemical quality of protein structures," *Journal of Applied Crystallography*, vol. 26, pp. 283–291, 1993.

[35] I. W. Davis, L. W. Murray, J. S. Richardson, and D. C. Richardson, "MolProbity: structure validation and all-atom contact analysis for nucleic acids and their complexes," *Nucleic Acids Research*, vol. 32, pp. W615–W619, 2004.

[36] R. Luthy, J. U. Bowie, and D. Eisenberg, "Assesment of protein models with three-dimensional profiles," *Nature*, vol. 356, no. 6364, pp. 83–85, 1992.

[37] M. Wiederstein and M. J. Sippl, "ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins," *Nucleic acids research*, vol. 35, pp. W407–410, 2007.

[38] C. Colovos and T. O. Yeates, "Verification of protein structures: patterns of nonbonded atomic interactions," *Protein Science*, vol. 2, no. 9, pp. 1511–1519, 1993.

[39] Y. D. Yang, P. Spratt, H. Chen, C. Park, and D. Kihara, "Sub-AQUA: real-value quality assessment of protein structure models," *Protein Engineering, Design and Selection*, vol. 23, no. 8, pp. 617–632, 2010.

[40] R. Koradi, M. Billeter, and K. Wüthrich, "MOLMOL: a program for display and analysis of macromolecular structures," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 51–55, 1996.

[41] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[42] P. Rice, L. Longden, and A. Bleasby, "EMBOSS: the European molecular biology open software suite," *Trends in Genetics*, vol. 16, no. 6, pp. 276–277, 2000.

[43] M. M. Krem and E. Di Cera, "Molecular markers of serine protease evolution," *EMBO Journal*, vol. 20, no. 12, pp. 3036–3045, 2001.

[44] R. J. Siezen and J. A. M. Leunissen, "Subtilases: the superfamily of subtilisin-like serine proteases," *Protein Science*, vol. 6, no. 3, pp. 501–523, 1997.

[45] L. B. Smillie and B. S. Hartley, "Histidine sequences in the active centres of some 'serine' proteinases," *Biochemical Journal*, vol. 101, no. 1, pp. 232–241, 1966.

[46] S. Roy, N. Maheshwari, R. Chauhan, N. K. Sen, and A. Sharma, "Structure prediction and functional characterization of secondary metabolite proteins of Ocimum," *Bioinformation*, vol. 6, no. 8, pp. 315–319, 2011.

[47] L. P. Tripathi and R. Sowdhamini, "Cross genome comparisons of serine proteases in Arabidopsis and rice," *BMC Genomics*, vol. 7, p. 200, 2006.

[48] L. Gutierrez, M. Castelain, J. L. Verdeil, G. Conejero, and O. Van Wuytswinkel, "A possible role of prolyl oligopeptidase during Linum usitatissimum (flax) seed development," *Plant Biology*, vol. 10, no. 3, pp. 398–402, 2008.