ISSN 1996-0700

Asian Journal of Biotechnology

Knowledgia SCIENTIFIC
A Place to Publish Outstanding Research

http://knowledgiascientific.com

Asian Journal of Biotechnology 3 (5): 435-448, 2011
ISSN 1996-0700 / DOI: 10.3923/ajbkr.2011.435.448

Molecular Modeling and Structural Analysis of Five SE Clan (S12 Family) Serine Proteases

[1,2]Aparna Laskar, [1]Sirshendu Chatterjee, [1]Anirban Roy, [2]Sumit Kumar Dey and [3]Chhabinath Mandal

[1]Indian Institute of Chemical Biology (A unit of CSIR), Division of Structural Biology and Bioinformatics, 4, Raja S.C. Mullick Road, Jadavpur, Kolkata-700 032, India

[2]Indian Institute of Chemical Biology (A unit of CSIR), Division of Immunology and Infectious Diseases, 4, Raja S.C. Mullick Road, Jadavpur, Kolkata-700 032, India

[3]National Institute of Pharmaceutical Education and Research, Indian Institute of Chemical Biology, (A unit of CSIR), 4, Raja S.C. Mullick Road, Jadavpur, Kolkata-700 032, India

Corresponding Author: Aparna Laskar, Indian Institute of Chemical Biology (A unit of CSIR), Division of Structural Biology and Bioinformatics, 4, Raja S.C. Mullick Road, Jadavpur, Kolkata-700 032, India Tel: +91-33-2499-5789 Fax: +91-33-24735197

ABSTRACT

SE clan of serine proteases have Lysine, Serine and Tyrosine residues as catalytic triad in their active site which is an exception to all other serine proteases having conserved residues Serine, Aspartic acid and Histidine in their active sites. The aim of this study was to explore the structural distinctiveness of different S12 family (D-Ala-D-Ala carboxypeptidase B family) serine peptidases from diverse groups of species belonging to SE clan, using molecular modeling techniques. Amino acid sequences of those proteases from each group of organisms, i.e., archean, protozoan, fungal, plant and human were taken from the MEROPS database in FASTA format. Homology and threading modeling approach were used to construct all the structures by SWISS MODEL and LOOPP server respectively. Structural quality assessing validation programs (PROCHECK, MODELYN, MOLPROBITY and PROSA) of the final model have demonstrated its reliability for further studies. A full structural database was constructed starting from bacteria up to human, focusing on the catalytic triad of the proteases. MODELYN study showed statistically that the region of the catalytic triad is highly conserved from bacteria to human. Phylogenetic analysis also supports their evolutionary connection. MOLMOL study showed that the catalytic sites of all the SE Clan proteases exhibit acidic regions in the surface electrostatic potential maps but a few of them contain both large patches of positive and neutral potentials. Hence some of the proteases with predominant positive and neutral regions around the catalytic site can be used as potential drug target against bacterial and protozoan pathogens.

Key words: SE clan, MEROPS, homology modeling, threading, electrostatic potential, catalytic triad

INTRODUCTION

Proteases (proteinases, peptidases or proteolytic enzymes) are enzymes that can split a protein into peptides and break peptide bonds between amino acids of proteins. Proteases work by lowering the activation barrier for peptide hydrolysis (Barrett *et al*., 2003; Rawlings and Barrett, 1993). The

activated state is usually a tetrahedral intermediate. The process is called proteolytic cleavage, a common mechanism of activation or inactivation of enzymes especially involved in blood coagulation or digestion. Peptidases and their natural, protein inhibitors are of great relevance to biology, medicine and biotechnology (Perona and Craik, 1995; Esmon, 2000).

A clan contains all the modern-day peptidases (Barrett *et al*., 2001) that have arisen from a single evolutionary origin of peptidases. It represents one or more families that show evidence of their evolutionary relationship by their similar tertiary structures, or when structures are not available, by the order of catalytic-site residues in the polypeptide chain. The MEROPS database (Rawlings *et al*., 2006) aims to fulfill the need for an integrated source of information about the proteins. The organizational principle of the database is a hierarchical classification in which homologous sets of proteins of interest are grouped into families and the homologous families are grouped in clans. There are different clans like, PA/SA, SB, SC and SE clans.

In this study, SE clan is considered that contains peptidases which have specialized for roles in bacterial cell wall metabolism, for the structure prediction and analysis using advanced molecular modeling techniques. The homology of the three families S11 (Englebert *et al*., 1994), S12 (Kelly *et al*., 1985) and S13 (Sauvage *et al*., 2005) of SE clan is supported by similarities in their sequence, protein fold, catalytic mechanism and biological functions. For this study, attention was focused mainly to the members of the S12 Family/D-Ala-D-Ala carboxypeptidase B family, contains serine-type D-Ala-D-Ala carboxypeptidases (Silvaggi *et al*., 2003). The active site residues Ser and Lys form the catalytic dyad and are found in the motif Ser-Xaa-Thr-Lys. There is a catalytic Tyr residue further toward the C-terminus in a conserved Tyr-Xaa-Asn motif. The tyrosine is replaced by serine in other families of SE clan. The aim of this study was to explore the structural characteristics of different S12 family serine peptidases belongs to SE clan from diverse groups of the living kingdom using molecular modeling techniques. Huge availability of protein sequences and computational approaches for identifying target proteins/ genes has been a great support for further experimental work (Shakyawar *et al*., 2011).

Conserved residues in serine proteases (Rawlings and Barrett, 1994; Krem and Cera, 2001) with a wide evolutionary span (Rawlings and Barrett, 1993) ranging from bacteria to mammals on the catalytic sites and their interactions are mainly highlighted for evolutionary insight and structure based drug designing (Anand *et al*., 2003; Turk, 2006; Tambunan and Parikesit, 2010). The present study aims at exploring the structural uniqueness of different S12 family (D-Ala-D-Ala carboxypeptidase B family) serine peptidases from different species belonging to SE clan which have specialized roles in bacterial cell wall metabolism, using molecular modeling techniques.

MATERIALS AND METHODS

Computational tools: In this study, the computational studies were performed using the following software packages. The study materials were collected from MEROPS database which aims to fulfill the need for an integrated source of information about the peptidases. The database contains a hierarchical classification of peptidases in which homologous sets of proteins of interest are grouped into clans and families among kingdoms of organisms, giving concise text annotations for each peptidase family. Homology (Floudas, 2007) and threading modeling were carried out using SWISS-MODEL server (Schwede *et al*., 2003) and LOOPP server (Meller and Elber, 2001) on PC.

SWISS-MODEL, PROSITE, LOOPP, CLUSTALW, MOLPROBITY and MOLMOL were used from the respective web-servers available through Internet at sites:

- http://swissmodel.expasy.org//SWISS-MODEL.html
- http://www.expasy.org/prosite
- http://cbsuapps.tc.cornell.edu/loopp.aspx
- http://www.ebi.ac.uk/clustalw/ and
- http://molprobity.biochem.duke.edu

Model construction and refinement: Amino acid sequence of SE clan-serine protease from each of the groups of archea (*Pyrococcus abyssi*, No MER026262), protozoan (*Dictyostellium discoideum*, MER035017), fungal (*Neurospora crassa*, MER033421), plant (*Arabidopsis thaliana*, No. MER004158) and human (*Homo sapiens*, MER017071) were taken from the MEROPS database. The amino acid sequences of the three proteases were subjected to NCBI PSI BLAST (Altschul *et al*., 1997) search in the Protein Structure Database (PDB) for selection of suitable template for comparative modeling. The challenge of template based modeling lies in the recognition of correct templates and generation of accurate sequence-template alignments. When suitable template was found to be absent in the protein BLAST search result, the initial models of the proteases were predicted using two web-based molecular modeling servers e.g. SWISS-MODEL server and LOOPP v4.0 and v3.0 server. Actually, protein modeling is the only way to obtain structural information if experimental techniques fail (Prajapat *et al*., 2010). The sequences of all the predicted structures both homologous and non homologous were aligned using the multiple alignment server, CLUSTALW and the structures were aligned using ABGEN (Mandal *et al*., 1996) to select the core structure of the stating scaffold of the target protein for subsequent refinement. The starting structure was refined using the Insight II 2005 of Accelrys (San Diego, CA) equipped with DISCOVER as the energy minimization and molecular dynamics module. Structural optimization involved energy minimization (100 steps each of steepest descent and conjugate gradient methods) using cff91 force-field followed by dynamics simulations. A typical dynamics run consisted of 10000 steps of one femto-second (10 picoseconds) after 1000 steps of equilibration with a conformational sampling of 1 in 10 steps at 300 K. However, dynamics simulation of 100 picoseconds was also applied to certain bigger loops for proper regularization. At the end of the dynamics simulation, the conformation with lowest potential energy was picked for the next cycle of refinement using the ANALYSIS module of Insight II. This combination of minimization and dynamics were repeated until satisfactory conformational parameters were obtained. Each loop was separately regularized applying position constraints to the rest of the atoms of the protein which were two amino acids away from the desired loop by energy minimization and molecular dynamics followed by evaluation of the structural parameters. SCWRL 3.0 (Canutescu *et al*., 2003) was used to regenerate the sidechains (Summers and Karplus, 1991) of the modeled protein. The final structure was energy minimized 100 steps each with steepest descent and conjugate gradient methods keeping all the atoms of the protein free.

Structure validation: Structural parameters and quality of the final 3D structures were determined using PROCHECK (Laskowski *et al*., 1993; Amir *et al*., 2011), MOLPROBITY (Davis *et al*., 2004), PROSA (Wiederstein and Sippl, 2007) and MODELYN (Mandal, 1998; Indian Copyright No. 9/98).

3D structure analysis: The ribbon structure and electrostatic potential surface of the model were determined by MOLMOL (Koradi *et al*., 1996).

Multiple sequence alignment and phylogenetic analysis of serine proteases: Amino acid sequences of serine proteases of archea (*Pyrococcus abyssi*, No. MER026262), protozoan (*Dictyostellium discoideum*, MER035017), fungal (*Neurospora crassa*, MER033421), plant (*Arabidopsis thaliana*, No MER004158) and human (*Homo sapiens*, MER017071) were taken from the MEROPS database in FASTA format and subjected to multiple sequence alignment using CLUSTALW (Thompson *et al*., 1994). Phylogenetic Analysis was also done by PHYLIP 3.5 program package while distance matrix was computed using PROTDIST with 100 BOOTSTRAP analysis. Phylogeny tree searching was performed by KITSH program (Chatterjee *et al*., 2011; Tambunan *et al*., 2008).

RESULTS

Model construction: There were X-ray structures of bacterial peptidases for this S12 of SE clan was present in the PDB; one of the X-ray structures, 1ONH.pdb (MER000463) was downloaded for structural analysis. For the rest of the groups i.e. archea, protozoa, fungi, plant and animal, no experimentally determined structures were available. Hence we selected one sequence from each group from the MEROPS database to predict their structures by molecular modeling techniques.

The sequences of the proteases were looked for significant sequence identity (>30%) suitable for homology based structure prediction, finds and compares regions of local similarity between sequences of proteins and calculates the statistical significance of matches. Homology modeling is based on the fact that a structure of a protein can be reliably modeled when its sequence is sufficiently similar to a protein sequence with known 3D structure. As the 3D structure models of target proteins are built upon its alignment to the template, 1KVM.pdb and 1CEG.pdb were considered as templates for homology modeling of archean and plant proteases respectively. Initial models were developed using SWISS-MODEL server which is a fully automated protein structure homology-modeling web server (Fig. 1). The purpose of this server is to make protein modeling accessible to all biochemists and molecular biologists worldwide.

The species of protozoa, fungi and animal as those have <30% sequence identity were considered for threading based modeling. Sequences from protozoan (*Dictyostellium discoideum*, MER035017), fungal (*Neurospora crassa*, MER033421) human (*Homo sapiens*, MER017071) proteases were submitted to the LOOPP server for threading based structure prediction and the search results for secondary structures of the target sequence has the best match with the X-ray structure PDB IDs: 1RGZ, 1SDE and 1CI9, respectively. The percent sequence identities of the templates are 18.96, 25.15 and 18.29%, respectively which shows that they are fit for threading models. The template modeling match scores and the extent of sequence coverage are given in Table 1. Template modeling match score is a measure of similarity between two protein structures, i.e., the modeled structure and the experimentally determined structure (X-ray structure) and they are independent of protein lengths. Here, all the match scores are above 0.5 which interprets that they have roughly assumed the same fold with template proteins. 3D structures of *Dictyostellium discoideum, Neurospora crassa* and *Homo sapiens* based on 1RGZ.pdb, 1SDE.pdb and 1CI9.pdb respectively were chosen as initial models.

Refinement of the modeled structures: Quality of the backbone of the modeled structure was assessed with PROCHECK for reliability. PROCHECK is widely used to scan a model for unlikely bonds, angles and dihedral values. It was observed that although most of the **φ-Ψ** pairs were distributed in the most favored and additional allowed regions of the Ramachandran's plot, the backbone conformation of some of the Amino Acids (AA) were in the generously allowed and disallowed regions as shown in Fig. 2 a and b and values are given in Table 2. For a model to

Table 1: The match scores, percentage sequence identity and the extent of sequence coverage of protozoan (*Dictyostellium discoideum*), fungal (*Neurospora crassa*) and animal (*Homo sapiens*) proteases using LOOPP server

	Initial sequence	PDB ID	Score	Sequence identity (%)	Length (%)
Dictyostellium discoideum	MER035017	1RGZ	3.786	18.96	92.45
Neurospora crassa	MER033421	1SDE	2.787	25.15	88.30
Homo sapiens	MER017071	1CI9	4.226	18.29	96.62

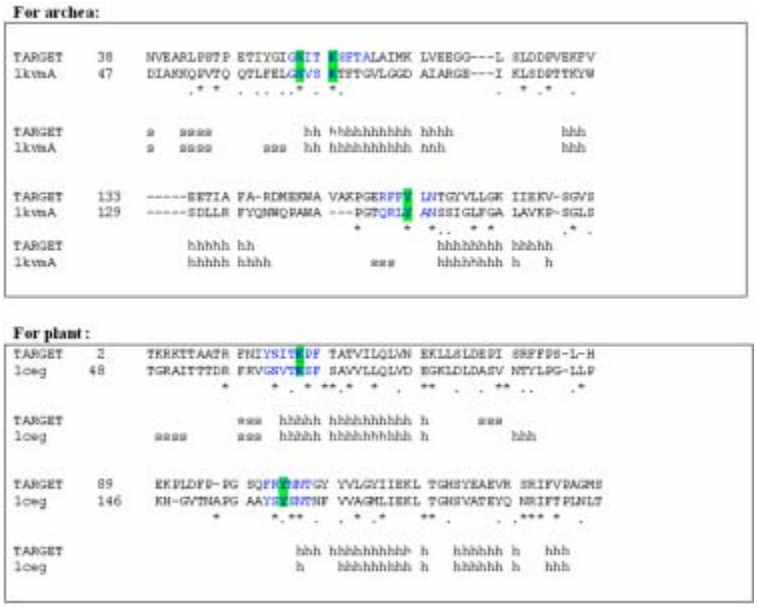

Fig. 1: Pairwise alignment of the amino acid sequences of the target with those of known structures/templates: Predicted secondary structures, s for β-sheets and h for α-helices, are also shown for each sequence. Columns containing highly conserved AA are marked with '*' and partially conserved sequences are marked with '.' along the alignments with the color blue and green

be validated based on quality, a good quality protein model should have >80% or more residues in the most favored regions of quadrangle in the Ramachandran's plot. In the cases of archean and plant protease models, 81.8 and 88.8% of residues respectively fall in the most favored regions after refinement. Initially, there were some residues which didn't fall under the most favored region in Ramachandran's plot. Then using CHARMm module of Insight II, the dihedral angles (φ-Ψ) of all residues were fixed so that they fall under the most favored region. The same protocol was followed for each residue and for each model. The percentage residues of other models of protozoan, fungal and human proteases are very close to 80% which fall in the most favored region in the Ramachandran's plot (Table 2). These were grouped into different segments of the structure and refined by energy minimization and molecular dynamics until all the backbone conformations fell in the desired regions. The right panels of Fig. 2 show the Ramachandran's plot of the refined model.

Validation of the structure: The side chains of the modeled structure were regenerated using SCRWL and the overall structure was energy minimized to get the refined homology model as well as threading models of the proteases from SE clan of different species. Overall atom clashscores and

Table 2: Results of the backbone refinement of the modeled proteases from 5 different species of the SE clan as analyzed checking the distribution of φ-Ψ dihedral angles in the Ramachandran's plot also in percentage

Serine Proteases										
Species with (MEROPS No.)	*Pyrococcus abyssi* (MER026262)		*Arabidopsis thaliana* (MER004158)		*Dictyostellium discoideum* (MER035017)		*Neurospora crassa* (MER033421)		*Homo sapiens* (MER017071)	
Model type	Homology		Homology		Threading		Threading		Threading	
Server used	SWISSMODEL		SWISSMODEL		LOOPP		LOOPP		LOOPP	
Procheck streochemical quality	Initial model	Final model	Initial model	Finalmodel	Initial model	Final model	Initial model	Final model	Initial model	Final model
Residues inmost favored region (%)	81.3	81.8	89.2	88.8	73.5	76.6	73.6	76.9	64.5	69.3
Residues in additional allowed region (%)	14.0	18.2	8.6	11.4	21.7	24.4	19.4	23.1	29.0	30.7
Residues in generously allowed region (%)	3.3	0.0	1.6	0.0	3.2	0.0	5.6	0.0	4.3	0.0
Residues in disallowed region (%)	1.4	0.0	0.5	0.0	1.6	0.0	1.4	0.0	2.2	0.0

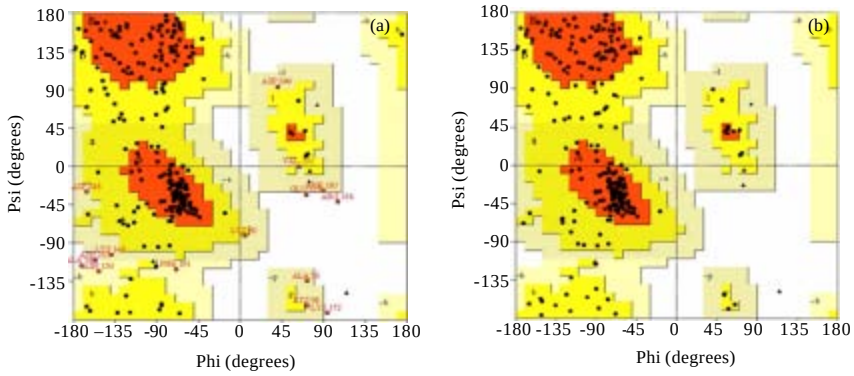

Fig. 2 (a-b): Ramachandran's plot of the **φ-Ψ** dihedral angles of the modeled structure of the protease from *Homo sapiens* (No.MER017071) (a) before and (b) after refinement of the backbone. Residues whose **φ-Ψ** pairs fell outside the desired most favorable and additional allowed zones are marked in red. Plot for a single species is shown for brevity

rotamer outliers were measured using MOLPROBITY. For structure validation and error correction purposes, the clash overlaps are very much dominant issue. It is considered that a serious clash to occur where two incompatible atoms overlap by 0.4 Å or more. The overall clashscore of a structure is the number of serious clashes per 1000 atoms (Gu and Bourne, 2009). It is seen from the listed values in the Table 3 that the all atom clashscore of predicted models of SE clan proteases of archean, animal and plant are quite close to the X-ray structures. Again, for X-ray targets, the target rotamer set consists of all residues for which a valid rotamer name could be assigned. If that score is = 1%, it assigns the name of the local rotamer peak and if = 1%, it declares an outlier. In this work, except for the fungal and animal proteases, all the calculated rotamer outliers for archean, plant and protozoan proteases are very close to their X-ray templates. The other structural parameters like deviation of bond lengths and bond angles from standard values were measured using MODELYN and Z score of the model were measured using PROSA (Fig. 3). These values were also compared with the relevant x-ray structures of the modeled structures used in homology modeling and threading modeling as shown in Table 3. It may be noted that these structural parameters are comparable to those of the X-ray structure indicating that our modeled structure has good general structural parameters.

Analysis of secondary structure and electrostatic potential surface of the protease of the SE clan: Most of the members of this clan exhibit extensive ordered secondary structures of **α**-helix and **β**-sheet except for fungal protease from *Neurospora crassa* and plant protease from *Arabidopsis thaliana*. The lack of extensive ordered secondary structures is probably due to less sequence coverage in the modeled structures of these proteases. All the structures of the proteases of this clan, experimental or modeled, contained two catalytic triad residues, Ser and Lys, are present on a single α-helical segment as they are present in a short and highly conserved motif, characteristics of this clan. But, the third catalytic triad residue is either tyrosine or serine and is well separated from the other two showing large variation in sequence and structures around it. The Cα atom RMSD of the triad residues of the modeled protease are within 1 Å which is found among serine proteases.

Table 3: Results of structural validity tests in terms of clashscores (overlaps >0.4 Å) and rotamer outliers (first two χ angles by >20° from its nearest associated rotamer) calculated using MOLPROBITY and root mean square deviations from standard bond lengths and bond angles, measured using MODELYN

S. No.	Structural model	All atom clashscore (No/1000 atoms)	Rotamer outliers (%)	RMSD of bond length (Å)	RMSD of bond angle (degree)	PROSA Z Score overall model quality
A1	X-ray structure with PDB ID 1KVM	3.50	2.62	0.015	2.60	-
A2	Homology model of *Pyrococcus abyssi* protease	8.00	1.46	0.015	2.64	-6.03
B1	X-ray structure with PDB ID 1SDE	3.76	2.40	0.018	2.59	-
B2	Homology model of *Arabidopsis thaliana* protease	3.25	1.13	0.029	2.94	-7.41
C1	X-ray structure with PDB ID 1RGZ	6.56	2.51	0.016	2.86	-
C2	Threading model of *Dictyostellium discoideum* protease	6.57	3.65	0.025	3.52	-5.27
D1	X-ray structure with PDB ID 1SDE	7.78	1.77	0.018	2.59	-
D2	Threading model of *Neurospora crassa* protease	10.10	4.96	0.017	3.24	-2.86
E1	X-ray structure with PDB ID 1CI9	11.00	6.12	0.014	2.56	-
E2	Threading model of *Homo sapiens* protease	11.30	3.51	0.021	3.83	-3.54

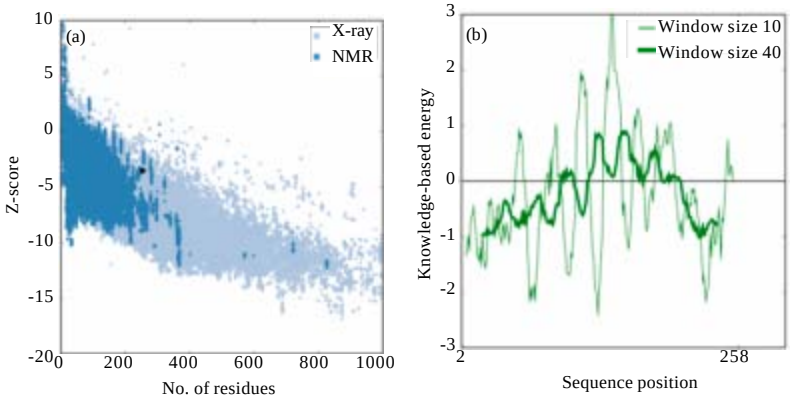

Fig. 3 (a-b): (a) Overall model quality (Z score) and (b) local model quality (residue interaction energy profile) of *Homo sapiens* (No. MER017071) after refinement of the backbone are determined using PROSA. Plot for a single species is shown for brevity

The Electrostatic Surface Potential (ESP) of the X-ray structure (PDB ID: 1ONH) of the bacterial protease of the SE clan showed that the electrostatic potential surface around the catalytic serine residue is mostly neutral along with patches of an intense positive regions.

The ESP of the archean protease from *Pyrococcus abyssi* of the SE clan (homology modeled structure) showed that the surface electrostatic potentials of this protein around the catalytic triad are mostly positive with a patch of neutral potentials. There is a highly intense negatively charged region near the catalytic triad.

ESP of the threading-modeled structure of the protein from *Dictyostellium discoideum* like the first two members of this clan showed the surface electrostatic potentials around the catalytic site are mostly neutral mixed with a patch of positive potentials. Thus, it has very different electrostatic environment around the catalytic site making it useful as a target for drug design (Fig. 4).

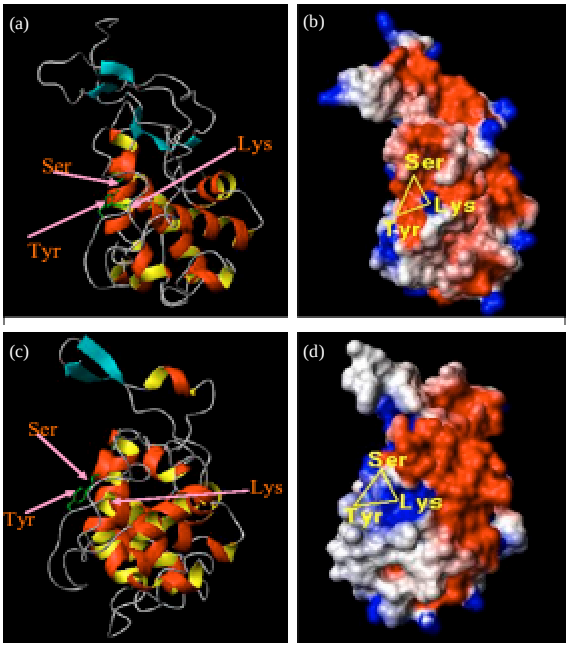

Fig. 4 (a-d): Secondary structure (left panel) and surface electrostatic potentials (right panel) of the X-ray structure of (a) plant protease from *Arabidopsis thaliana* of the SE clan (MEROP No. MER004158), (c) protozoan protease from *Dictyostellium discoideum*(MEROP No. MER035017) of the SE clan. In secondary structures β-sheets are shown in light blue with an arrow pointing to C-terminus, α-helices are shown in red and yellow, turn/loops are shown in gray and side-chains of catalytic triad residues are shown in green sticks. In surface electrostatic potentials (right panel) of *Arabidopsis thaliana*, (b) and Dictyostellium *discoideum* and (d) red, blue and white colors represent negative, positive and neutral charges respectively. Both the figures are generated in the same orientation using MOLMOL. Plots for only two species are shown for brevity

In the same way, ESP of the threading model of the fungal protease from *Neurospora crassa* was constructed. The surface electrostatic potentials around the catalytic site are a combination of negative, positive and neutral regions which is different from the other proteases of this SE clan.

The ESP of the homology model of the plant protease from *Arabidopsis thaliana* showed that the catalytic triad residues Ser-16 and Tyr-102 are present in the negative electrostatic potentials and Lys-19 is in a positive zone (Fig. 4).

Secondary structure (left panel) and surface electrostatic potential map (right panel) of the threading model of the human (*Homo sapiens*) protease of the SE clan were also prepared. The surface electrostatic potentials around the catalytic site are a combination of negative, positive and neutral regions. In many human proteases the surface electrostatic potential is highly negative but in this bacterial serine protease is very different and may be useful in drug design.

Analysis of the core structure and active site geometry of the modeled structures of the proteases of the SE clan in reference to the experimental structures: In order to analyze the structural environment of the modeled structures and to compare them with those of the experimental structures, the distances between the Cα atoms of catalytic triad residues Ser to Lys,

Lys to Tyr (/Ser) and Tyr (/Ser) to Ser in the refined structures were measured using MODELYN. The mean values of distances between the Cα atoms of Ser to Lys, Lys to Tyr and Tyr to Ser of the triads in the modeled structures are 5.7±0.07, 11.5±0.06 and 10.6±0.07, respectively. The deviations of values from mean distances between the Cα atoms of the triad residues are much smaller than those of the X-ray structures (Table 4). This is because of the fact that all the triad residues are the same (i.e., the third residue is Tyr and not Ser).

The 3-D structures of the selected proteases of the SE clan were superpose with the representative x-ray structure (1ONH) of the bacterial protease with respect to a selected set of Cα atoms in order to examine the deviations of the core of the structures of the modeled proteins of the SE clan. It is found that 14 to 37% of the Cα atoms superposed with an RMS deviation below 1 Å. while only 2 to 7% of Cα atoms of the X-ray structures of the bacterial proteases of this clan superposed with an RMS deviation below 2 Å.

Phylogenetic analysis: The consensus phylogenetic tree of all the five S12 family members of our interest is shown in Fig. 5. Although, the core remained relatively conserved some segments of the proteases varied a lot during evolution.

Table 4: Analysis of the core structure and active site geometry of the modeled structures of the proteases of the SE clan in reference to the experimental structures. All the modeled structures were superposed on the x-ray structure (1ONH) of the bacterial protease with respect to a selected set of Cα atoms

				Active site residues			Cα distance of triad (Å)		
MEROP No.	Group	Species	Peptidase unit	S	K	Y	(S-K)	(K-Y)	(Y-S)
MER000463	Bacteria	*Enterobacter cloacae*	30-381	84	87	170	5.5	11.3	10.7
MER026262	Archea	*Pyrococcus abyssi*	4-353	59	62	160	5.4	11.2	10.4
MER004158	Plant	*Arabidopsis thaliana*	2-208	17	20	103	5.6	11.4	10.8
MER035017	Protozoa	*Dictyostellium discoideum*	48-444	111	114	217	5.8	11.9	10.2
MER033421	Fungi	*Neurospora crassa*	9-302	118	121	179	5.4	11.3	10.7
MER017071	Animal	*Homo sapiens*	102-547	164	167	323	6.1	11.5	10.9
Mean±SD of the Ca distances between the triad residues→							5.7±0.07	11.5±0.06	10.6±0.07

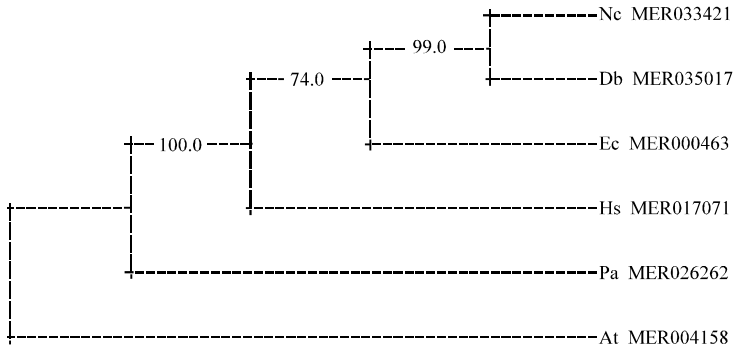

Fig. 5: Phylogenetic Analysis of SE clan proteases, One from each of the groups of bacteria (*Enterobacter cloacae* No MER000463), archea (*Pyrococcus abyssi*, No MER026262), protozoan (*Dictyostellium discoideum*, MER035017), fungal (*Neurospora crassa*, MER033421), plant (*Arabidopsis thaliana*, No MER004158) and human (*Homo sapiens*, MER017071) were carried out using Phylip web server

DISCUSSION

Conceptually, there are three basic approaches to protein structure prediction, e.g., homology modeling, threading and ab initio or template free methods. Comparative modeling alone or in conjunction with fold recognition analysis provides relatively accurate structures when the target sequence of a protein is evolutionarily related to a template which has an experimentally determined structure in the Protein Data Bank (PDB). (Chatterjee *et al*., 2011). In this study, serine proteases from the diverse groups were selected of the living kingdom based on MEROP classification of the SE clan for 3-D structural modeling. Five structures of only bacterial proteases of this clan have been determined by experimental method; we have included them as representative of structures of the proteases of this clan for comparison with our predicted structures of proteases from other groups. Most of the studies in this context has predicted a 3D model of a protein and established it by further refinement and validation. There are ample numbers of literatures which have predicted 3D structures of different proteins (Rajesh *et al*., 2008; Prabhavathi *et al*., 2011; Ginalski and Rychlewski, 2003). Majority of the studies have been reported about the identification of drug targets by experimental as well as *in silico* methods as targeting proteases is going to be very useful in drug design; especially the non-homologous proteins are very effective (Smith, 2004; Chong *et al*., 2006; Suthar *et al*., 2009; Shakyawar *et al*., 2011; Singh *et al*., 2006) but till date, no studies have been reported on computer-based 3D modeling of SE clan serine proteases. The prepared models of this clan using experimental data available in the literature are also submitted to PDB. From that point of view, the present study is unique and much more virgin in nature as this has covered a wide arena of model prediction to evolutionary analysis.

CONCLUSION

In the present study S12 family of serine peptidases of SE clan were chosen for structural study. No 3D structures of the serine proteases of SE clan were available in databases except only the X-ray structures of bacteria, as SE clan is a lower species belonging clan. Therefore, 3D models for the archea, protozoa, fungi, plant and human were built using comparative modeling techniques. The models were evaluated and validated using Discovery Studio visualizer (Accelrys) and PROCHECK. Further refinement and validation were done by Ramachandran plot analysis.

The structural models, both experimental and predicted, are used to analyze various structural properties of the serine proteases of the SE clan mainly focusing on the common core harboring the catalytic site of this very important class of enzymes which have been used and hold high potential in the structure based rational drug design. Examination of the geometry of the triad residues reveal that the modeled enzyme can catalyze the peptide hydrolyzing activity. In general, the catalytic site of all the proteases exhibit acidic regions in the surface electrostatic potential maps but a few of them contain both large patches of positive and neutral potentials. Hence some of the proteases with predominant positive and neutral regions around the catalytic site can be used as drug target against bacterial and protozoan pathogenic organisms as the interactions of inhibitors are strongly influenced by the nature of electrostatic potential surface near the substrate/inhibitor binding sites.

However, the essence of this study has not only prediction of the 3D structures of the serine proteases; but also the establishment an evolutionary relationship between them. The residues of the catalytic triad in all the proteases remain same but they differ in their position in each case. The

kinship amongst the proteases, though belonging in different species inspite of having big phylogenetic distances have been established by phylogenetic analysis; that's why they belong to a single "clan".

Most of the studies in this context has predicted a 3D model of a protein and established it by further refinement and validation. From that point of view, the present study is unique and much more virgin in nature as this has covered a wide arena of model prediction to evolutionary analysis.

ACKNOWLEDGMENTS

The authors thank Council for Scientific and Industrial Research (CSIR), New Delhi, India, for providing financial grant for this project. Authors also thank Indian Institute of Chemical Biology (IICB), Kolkata, India, for providing the laboratory support and other infrastructures.

REFERENCES

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

Amir, A., M.A. Siddiqui, N. Kapoor, A. Arya and H. Kumar, 2011. *In silico* molecular docking of influenza virus (PB2) protein to check the drug efficacy. Trends Bioinform., (In Press).

Anand, K., J. Ziebuhr, P. Wadhwani, J.R. Mesters and R. Hilgenfeld, 2003. Coronavirus main proteinase (3CLpro) structure: Basis for design of anti-SARS drugs. Sci., 300: 1763-1767.

Barrett, A.J., N.D. Rawlings and E.A. O'Brien, 2001. The MEROPS database as a protease information system. J. Struct. Biol., 134: 95-102.

Barrett, A.J., N.D. Rawlings and J.F. Woessner, 2003. The Handbook of Proteolytic Enzymes. 2nd Edn., Academic Press, USA.

Canutescu, A.A., A.A. Shelenkov and R.L. Dunbrack Jr., 2003. A graph theory algorithm for protein side-chain prediction. Protein Sci., 12: 2001-2014.

Chatterjee, S., A. Laskar, A. Chatterjee, C. Mandal and S. Chaudhuri, 2011. Insilico structural analysis of an immunotherapeutic glycoprotein T11TS (sheep CD58). Int. J. Biol. Med. Res., 2: 346-359.

Chong, C.E., B.S. Lim, S. Nathan and R. Mohamed, 2006. *In Silico* analysis of *Burkholderia pseudomallei* genome sequence for potential drug targets. *In Silico* Biol., 6: 341-346.

Davis, I.W., L.W. Murray, J.S. Richardson and D.C. Richardson, 2004. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res., 32: W615-W619.

Englebert, S., P. Charlier, E. Fonze, Y. To'th, M. Vermeire *et al.*, 1994. Crystallization and X-ray diffraction study of the Streptomyces K15 penicillin-binding DD-transpeptidase. J. Mol. Biol., 241: 295-297.

Esmon, C.T., 2000. Regulation of blood coagulation. Biochim. Biophys. Acta, 1477: 349-360.

Floudas, C.A., 2007. Computational methods in protein structure prediction. Biotechnol. Bioeng, 97: 207-213.

Ginalski, K. and L. Rychlewski, 2003. Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. Proteins, 53: 410-417.

Gu, J. and P.E. Bourne, 2009. Structural Bioinformatics. 2nd Edn., Wiley-Blackwell Publication, USA.

Kelly, J.A., J.R. Knox, P.C. Moews, G.J. Hite and J.B. Bartolone *et al.*, 1985. 2.8-A Structure of penicillin-sensitive D-alanyl carboxypeptidase-transpeptidase from Streptomyces R61 and complexes with beta-lactams. J. Biol. Chem., 260: 6449-6458.

Koradi, R., M. Billeter and K. Wuthrich, 1996. MOLMOL: A program for display and analysis of macromolecular structures. J. Mol. Graph., 14: 51-55.

Krem, M.M., and D.E. Cera, 2001. Molecular markers of serine protease evolution. EMBO J., 20: 3036-3045.

Laskowski, R.A., M.W. MacArthur, D.S. Moss and J.M. Thornton, 1993. PROCHECK: A program to check the stereochemical quality of protein structures. J. Applied Cryst., 26: 283-291.

Mandal, C., Kingery B.D., J.M. Anchin, S. Subramanium and D.S. Linthicum, 1996. ABGEN: A knowledge based automated approach for antibody structure modeling. Nat. Biotechnol., 14: 323-328.

Mandal, C., 1998. MODELYN: A molecular modelling programversion, PC-1.0. Indian Copyright No 9/98.

Meller, J. and R. Elber, 2001. Linear programming optimization and a double statistical filter for protein threading protocols. Proteins, 45: 241-261.

Perona, J.J. and C.S. Craik, 1995. Structural basis of substrate specificity in the serine proteases. Protein Sci., 4: 337-360.

Prabhavathi, M., K. Ashokkumar, N. Geetha and K.M. Saradha Devi, 2011. Homology modeling and structure prediction of thioredoxin (TRX) protein in wheat (*Triticum aestivum* L.). Int. J. Biosci., 1: 20-32.

Prajapat, R., R.K. Gaur, R. Raizada and V.K. Gupta, 2010. *In silico* analysis of genetic diversity of begomovirus using homology modelling. J. Biol. Sci., 10: 217-223.

Rajesh, S., M. Raveendran and A. Manickam, 2008. Comparative modeling and analysis of 3-D structure of EMV2, a late embryogenesis abundant protein of *Vigna Radiata* (Wilczek). J. Proteomics Bioinform., 1: 401-407.

Rawlings, N.D. and A.J. Barrett, 1993. Evolutionary families of peptidases. Biochem. J., 290: 205-218.

Rawlings, N.D. and A.J. Barret, 1994. Families of serine peptidases. Methods Enzymol., 244: 19-61.

Rawlings, N.D., F.R. Morton and A.J. Barrett, 2006. MEROPS: The peptidase database. Nucleic Acids Res., 34: D270-D272.

Sauvage, E., R. Herman, S. Petrella, C. Duez, F. Bouillenne, J.M. Frere and P. Charlier, 2005. Crystal structure of the Actinomadura R39 DD-peptidase reveals new domains in penicillin-binding proteins. J. Biol. Chem., 280: 31249-31256.

Schwede, T., J. Kopp, N. Guex and M.C. Peitsch, 2003. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res., 31: 3381-3385.

Shakyawar, S.K., A. Goyal and V.K. Dubey, 2011. Database of *in silico* predicted potential drug target proteins in common bacterial human pathogens. Am. J. Drug Discovery Dev., 1: 70-74.

Silvaggi, N.R., J.W. Anderson, S.R. Brinsmade, R.F. Pratt and J.A. Kelly, 2003. The crystal structure of phosphonate-inhibited D-Ala-D-Ala peptidase reveals an analogue of a tetrahedral transition state. Biochem., 42: 1199-1208.

Singh, S., B.K. Malik and D.K. Sharma, 2006. Molecular drug targets and structure based drug design: A holistic approach. Bioinformation, 1: 314-320.

Smith, C., 2004. Drug target identification: A question of biology. Nature, 428: 225-231.

Summers, N.L. and M. Karplus, 1991. Modeling of side chains, loops and insertions in proteins Methods Enzymol., 202: 156-204.

Suthar, N., A. Goyal and V.K. Dubey, 2009. Identification of potential drug targets of *Leishmania Infantum* by in-silico genome analysis. Lett. Drug Des. Discovery, 6: 620-622.

Tambunan, U.S.F., O. Hikmawan and T.A. Tockary, 2008. *In silico* mutation study of haemagglutinin and neuraminidase on banten province strain influenza a H5N1 virus. Trends Bioinform., 1: 18-24.

Tambunan, U.S.F. and A.A. Parikesit, 2010. *In silico* design of drugs and vaccines for dengue disease. Trends Bioinform., (In Press).

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

Turk, B., 2006. Targeting proteases: Successes, failures and future prospects. Nat. Rev. Drug Discov., 5: 785-799.

Wiederstein, M. and M.J. Sippl, 2007. PROSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res., 35: 407-410.