Gene 593 (2016) 154–161

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Identification and in silico characterization of two novel genes encoding peptidases S8 found by functional screening in a metagenomic library of Yucatán underground water

Max M. Apolinar-Hernández [a], Yuri J. Peña-Ramírez [b], Ernesto Pérez-Rueda [c,1], Blondy B. Canto-Canché [a], César De los Santos-Briones [a], Aileen O'Connor-Sánchez [a,*]

[a] Unidad de Biotecnología, Centro de Investigación Científica de Yucatán A.C., Calle 43 No. 130, Chuburná de Hidalgo, Mérida, Yucatán CP 97200, Mexico
[b] El Colegio de la Frontera Sur (ECOSUR) Unidad Campeche, Avenida Rancho Polígono 2A, Ciudad Industrial Lerma, Campeche, Campeche CP 24500, Mexico
[c] Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, UNAM, Cuernavaca, Morelos CP 62210, Mexico

ARTICLE INFO

Article history:
Received 26 April 2016
Received in revised form 19 July 2016
Accepted 4 August 2016
Available online 10 August 2016

Keywords:
Culture-independent analysis
Enzymes
Metagenomics
Proteases
Serine proteases
Yucatán aquifer

ABSTRACT

Metagenomics is a culture-independent technology that allows access to novel and potentially useful genetic resources from a wide range of unknown microorganisms. In this study, a fosmid metagenomic library of tropical underground water was constructed, and clones were functionally screened for extracellular proteolytic activity. One of the positive clones, containing a 41,614-bp insert, had two genes with 60% and 68% identity respectively with a peptidase S8 of *Chitinimonas koreensis*. When these genes were individually sub-cloned, in both cases their sub-clones showed proteolytic phenotype, confirming that they both encode functional proteases. These genes – named *PrAY5* and *PrAY6*- are next to each other. They are similar in size (1845 bp and 1824 bp respectively) and share 66.5% identity. An extensive in silico characterization showed that their ORFs encode complex zymogens having a signal peptide at their 5′ end, followed by a pro-peptide, a catalytic region, and a PPC domain at their 3′ end. Their translated sequences were classified as peptidases S8A by sequence comparisons against the non-redundant database and corroborated by Pfam and MEROPS. Phylogenetic analysis of the catalytic region showed that they encode novel proteases that clustered with the sub-family S8_13, which according to the CDD database at NCBI, is an uncharacterized subfamily. They clustered in a clade different from the other three proteases S8 found so far by functional metagenomics, and also different from proteases S8 found in sequenced environmental samples, thereby expanding the range of potentially useful proteases that have been identified by metagenomics. I-TASSER modeling corroborated that they may be subtilases, thus possibly they participate in the hydrolysis of proteins with broad specificity for peptide bonds, and have a preference for a large uncharged residue in P1.

1. Introduction

Proteases (synonym of peptidases) are hydrolytic enzymes which catalyze the cleavage of peptide bonds of other proteins, and in some cases they even have the capacity to cleave themselves (Zheng et al., 2002). They are produced by all living beings, in which they play important physiological roles, e.g. in catabolism, protein turnover, and cell division. In the market of biotechnological products, proteases constitute the most important group of commercial enzymes. They are used for a wide range of applications, especially in detergent, food and pharmaceutical industries (Sawant and Nagendran, 2014), which are eager for novel proteases that can better fulfill their process requirements. Specifically, peptidases S8 (also called subtilases) have an outstanding relevance in the detergent industry (Niehaus et al., 2011), which has the largest share of the enzyme market (Gupta et al., 2002).

Although proteases can be obtained from plants and animals, microbial proteases have conquered two thirds of the global protease market, mainly because they are easier to extract and face less climatic and ethical issues (Rao et al., 1998). Until some years ago, the search for novel microbial proteases was conducted exclusively by traditional microbiology procedures based on the cultivation of microorganisms. Nevertheless, nowadays we know that these procedures exclude up to the 99% of the total bacteria present in a given environmental sample (Torsvik et al., 1990; Amann et al., 1995). Functional metagenomics arose as the most promising alternative to search for novel biological enzymes with potential industrial relevance from uncultivated microorganism (Chistoserdova, 2009). It basically consists in extracting and cloning the total microbial DNA contained in an environmental sample

* Corresponding author.
E-mail address: aileen@cicy.mx (A. O'Connor-Sánchez).
[1] Current address: Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM. Unidad Académica Yucatán. Carretera Sierra Papacal, Chuburná Km 5. Mérida, Yucatán, CP 97302, Mexico.

http://dx.doi.org/10.1016/j.gene.2016.08.009
0378-1119/

(constructing a metagenomic library), and screening the clones by appropriate functional assays to detect those with a desired phenotype. Thus, it is an approach that does not depend on any conserved gene sequences, thereby permitting the discovery of unknown genes. This way of searching has allowed the discovery of several novel enzymes, such as lipases, esterases, amylases, chitinases and proteases (e.g. Hjort et al., 2014; Privé et al., 2015; Vester et al., 2015).

For as we know, there are eleven papers reporting novel proteases discovered by functional metagenomics: five from DNA in soil samples (Waschkowitz et al., 2009; Berlemont et al., 2011; Niehaus et al., 2011; Biver et al., 2013; Purohit and Singh, 2013), one from mud (Lee et al., 2007), one from sand (Neveu et al., 2011), one from goat skin (Pushpam et al., 2011), one from sediment (Zhang et al., 2011), one from plankton and rhizosphere (Guazzaroni et al., 2013), and one from activated sludge (Morris and Marchesi, 2015). To our knowledge, no water body has ever been investigated to search for novel proteases by means of functional metagenomics. The underground aquifer in Yucatán is an especially interesting fresh water environment because, as the ground in the state of Yucatán consists of a highly permeable limestone, the rain water easily percolates into the subsoil (Bauer-Gottwein et al., 2011), collecting and carrying a wide assortment of microorganisms. Furthermore, as the Aquifer of Yucatán constitutes an extensive net of underground rivers, its streams can carry a mixture of microorganisms from different distant places. Thus, although it should be a process of natural selection that favors some of them and eliminates some others, it may contain a representation of microorganisms of the whole zone.

The aim of the present work was to find novel genes encoding secreted proteases, for their potential to be biotechnologically useful, and to learn the characteristics of these genes, in order to enable their further overexpression and purification of their products. This paper reports the construction of a large-insert metagenomic library constructed with DNA isolated from Yucatán aquifer water, the identification and characterization of two novel protease S8 ORFs found by functional assays, and the modeling of the enzymes they encode.

2. Materials and methods

2.1. Environmental sample and metagenomic DNA extraction

Sixty liters of underground water from the Yucatán aquifer were pumped from a well (21° 01′ 44″ N/89° 38′ 18″ W) and immediately sequentially filtered, first through a 5 μm Millipore® filter (Cat No: SVGV010RS) (Merck-Millipore. Darmstadt, DE), which was discarded, and then through a 0.22 μm Millipore Sterivex – GV® filter (Cat No: SVGPB1010), where the prokaryote biomass was trapped. Previously, the internal surface of all plastic tubes used for pumping had been disinfected by passing a 10% (v/v) commercial bleach (6% free chlorine) solution containing 100 $\mu L \cdot L^{-1}$ of Tween 20 (Sigma, St. Louis, US) for 10 min, followed by passing sterile water for another 10 min. Metagenomic DNA was extracted from the 0.22 μm filter using the "Metagenomic DNA isolation kit for water" (Epicentre-Illumina. Madison, US) following the manufacturer's protocol.

2.2. Library construction and functional screening

A large-insert metagenomic DNA was constructed using the previously extracted DNA and the "CopyControl®™ HTP Fosmid Library Production Kit with pCC2FOS®™ Vector" (Epicentre-Illumina) according to the manufacturer's instructions, using *Escherichia coli* EPI300 (Epicentre-Illumina) as host. The quality of the library was evaluated analyzing the *Bam*HI restriction pattern of inserts from different clones. Functional screening (for protease activity detection) was performed by inducing multicopy fosmids with 0.01% L(+)-arabinose, and growing the recombinant clones of the library on Luria-Bertani (LB) agar medium supplemented with 2% (w/v) lactose-free skim milk. As selective agent 12.5 $\mu L \cdot L^{-1}$ chloramphenicol were used. After three days of incubation at 37 °C, protease activity was detected by the presence of a clear halo around the positive clones.

2.3. Analysis of positive clones

Positive clones were grown overnight at 37 °C and 250 rpm, in 1.5 mL of liquid LB medium containing 12.5 $\mu g \cdot mL^{-1}$ chloramphenicol and 0.01% arabinose. Recombinant fosmids were extracted from these cultures using the "FosmidMAX®™ DNA Purification Kit" (Epicentre-Illumina) following the manufacturer's protocol. The *Bam*HI restriction patterns of positive clones were analyzed in a 0.8% agarose gel, using "GeneRuler 1 Kb Plus DNA Ladder" (Thermo-Fisher Scientific. Waltham, US) as molecular-weight marker.

2.4. DNA sequencing and in silico analysis

Recombinant fosmids from positive clones were sequenced and assembled at the "USU Center for Integrated Biosystems" of Utah State University using the Ion Torrent chip 314 (Thermo-Fisher Scientific). ORFs in the contigs were predicted by using the Glimmer3 system (with default parameters) (Salzberg et al., 1998), and results were verified with fgenesB (Solovyev and Salamov, 2011) (http://www.softberry.com/berry.phtml?topic=fgenesb). Annotation of these ORFs was by BLAST analysis against the Non-Redundant NCBI database (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Translation was performed using the genetic code of Bacteria and Archaea. Identity of the sequences acknowledged as peptidases was subsequently verified by using Pfam (http://pfam.xfam.org/), UniProt (http://www.uniprot.org/), and HMMER (http://www.ebi.ac.uk/Tools/hmmer/) databases. Classification of the identified proteases was accomplished by using MEROPS–the peptidase database (http://merops.sanger.ac.uk/). Searching of conserved domains was conducted with Conserved Domains and Protein Classification (CDD) (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). SignalP 4.1 server (http://www.cbs.dtu.dk/services/SignalP/) was used for signal peptides prediction, and TMHMM 2.0 server (http://www.cbs.dtu.dk/services/TMHMM/) for trans-membrane helixes. Protease sequences were then aligned with the MUSCLE tool included in the MEGA6 software (Tamura et al., 2013) (http://www.megasoftware.net/), completed with manual refinements to correct misalignments. A phylogenetic tree was constructed by using the neighbor-joining method (Saitou and Nei, 1987) with 1000 replicates (Felsenstein, 1985). Evolutionary distances were computed by the Poisson correction method (Zuckerkandl and Pauling, 1965).

2.5. Subcloning of PrAY5 and PrAY6 ORFs

PrAY5 and *PrAY6* ORFs were individually subcloned in the expression vector pLATE52 by using specific primers (Table S1). The PCR reaction was carried out using MAX10 fosmid DNA as template and the enzyme Phusion Green Hot Star II High-Fidelity (Thermo scientific), following the manufacturer's specifications. The constructions were transformed into the *E. coli* host strain BL21 (DE3) (New England Biolabs). As selective agent 100 $\mu g \cdot mL^{-1}$ Ampicillin was used. Once subcloned, the ORF sequences were confirmed by Primer walking sequencing (Macrogen Inc.).

2.6. Structural modeling

A computational model of the three-dimensional structure of PrAY5 and PrAY6 was constructed by using I-TASSER V4.4 Suite (Yang et al., 2015). Overlapping of the 3D models was made with Mustang-MR Structural Sieving Server (Konagurthu et al., 2010) (http://pxgrid.med.monash.edu.au:8080/mustangserver/). Results were visualized in the PyMOL Molecular Graphics System (DeLano, 2002).

2.7. Accession numbers

PrAY5 and *PrAY6* ORF sequences were deposited in the NCBI Sequence Read Archive (SRA) under the accession numbers KU198630 and KU198631, respectively.

3. Results and discussion

3.1. Library construction and functional screening

A fosmid metagenomic library was constructed with DNA extracted from prokaryotic biomass of the Yucatán underground water. This library consisted of ~250,000 clones harboring DNA inserts of about 30–40 Kb length (Fig. S1).

Screening of clones with extracellular proteolytic activity was based on the appearance of hyaline halos around the positive colonies (caused by the hydrolysis of milk in the growing medium) (Fig. S2). 21,000 randomly chosen clones of the library (approximately 735 Mb of metagenomic DNA) were screened, out of which 23 resulted positive for the assay (MAX1 to MAX23). When these 23 positive clones were transferred into fresh indicator medium, all of them retained the proteolytic phenotype for at least three subculture cycles, suggesting that their insert-bearing-fosmids, as well as the phenotype they originate, are stable. When the inserts of these 23 colonies were analyzed by *Bam*HI restriction, it was observed that all of them showed unique digestion patterns, with inserts ranging from 15 Kb to 47 Kb (Fig. S3), meaning that all of them were different from each other.

3.2. Sequence analysis

The inserts of clones MAX1, MAX10 and MAX19, which produced more hyaline and wider halos than the other positive clones, were sequenced. The insert of MAX1 was 35,616 bp, whereas the inserts of MAX10 and MAX19 were 41,614 bp and 30,538 bp respectively. GLIMMER3 gene prediction analysis showed 38 genes in MAX1, 50 in MAX10, and 36 in MAX19. When all these genes were analyzed by BLAST in the non-redundant NCBI database, only two ORFs in MAX10 showed significant identity with known protease-encoding-genes. Therefore, we focused on the analysis of clone MAX10 to go forward.

3.3. Analysis of the cloned DNA in MAX10

Nineteen out of the 50 predicted genes in MAX10 insert were encoded in the sense (positive) DNA strand and 31 were located in the antisense (negative) DNA strand. The general GC content of the insert was 56.2%. Fifty percent of the ORFs (25/50) were similar to proteins with a known nearest homolog when searched by BLAST against the non-redundant NCBI database, 42% of the genes encoded proteins with a hypothetical or an uncharacterized protein as nearest homolog, and 8% percent encoded proteins with no significant similarity (Table S2). These results suggest that the cloned DNA belongs to a bacterial genome that has not been sequenced until now (or at least is not publicly available in the NCBI database).

When zooming into the neighborhood of the two protease genes in MAX10 (Fig. 1, Table 1), verifying the genes annotation with fgenesB suite, it was observed that nine out of the ten adjacent ORFs encode proteins with nearest homologs belonging to the phylum Proteobacteria, with identities ranging from 33% to 82%. Only in one case the nearest homolog belonged to the phylum Acidobacteria, with 46% identity. Of the Proteobacteria cases, seven belonged to Betaproteobacteria, and two to Gammaproteobacteria. In four out of the ten cases, the genus of the nearest homolog was *Chitinimonas*, including the two protease genes. These results reinforce the supposition that the cloned DNA in MAX10 belongs to an unknown bacterial genome, possibly a Proteobacterium.

Specifically, the translated ORF sequences of gene 5 and gene 6 (named hereafter PrAY5 and PrAY6 – from the Spanish "Proteasas del Acuífero de Yucatán") shared 60% and 68% identity, respectively, with a peptidase S8 of *Chitinimonas koreensis* (phylum Proteobacteria, class Betaproteobacteria, order Burkholderiales), which has been reported only once, as a Gram-negative motile bacterium isolated from greenhouse soil in Korea (Kim et al., 2006). The query cover for PrAY5 and PrAY6 was 94% and 100% respectively, and in both cases the e-value was zero, meaning that it is likely that the identity percentage is statistical and biologically significant.

3.4. Subcloning of PrAY5 and PrAY6 ORFs

To verify whether *PrAY5* and *PrAY6* ORFs were indeed responsible of the hydrolytic phenotype, each of them was individually sub-cloned in a pLATE52 vector. Randomly taken sub-clones were subjected to the skim-milk-medium-test, being in all cases able to produce a clear halo around them (Fig. S4). This was experimental evidence showing that both of these genes certainly encode functional proteases. The halo around *PrAY5* sub-clones was smaller and less hyaline than the halo formed around *PrAY6*. One possible explanation for this is that the protease encoded by *PrAY5* is less stable or less active under the tested conditions than the protease encoded by *PrAY6*. Another is that PrAY5 is not as efficiently folded or secreted as PrAY6 in *E. coli*. Nevertheless, this result shows that each of the ORFs is able to produce a proteolytic phenotype when separated from the other, suggesting that none of them depends on the other for its expression, and none of the proteases they encode depends on the other to be secreted and active.

3.5. In silico characterization of PrAY5 and PrAY6

As can be observed in Fig. 1, *PrAY5* and *PrAY6* are next to each other; they are only separated by 718 bp. In both cases, the ORF starts with the codon ATG and ends with the codon TAA, which are common initiation and termination codons. They exhibit similar length, *PrAY5* is 1845 bp (614 aa) and *PrAY6* is 1824 bp (607 aa) long, and they are also similar at sequence level (66.5% identity); perhaps they encode isozymes. It is likely that they are paralogs (Mira et al., 2006), and one of these genes arose by duplication of the other in the evolution of the genome of the organism they belong to. It was also noticed that *PrAY5* and *PrAY6* are much larger than their eight adjacent genes (Fig. 1, Table 1).

When PrAY5 and PrAY6 were aligned to each other with MUSCLE, it was observed that they share 66.9% identity, and that their differences were distributed all through their ORFs, and not concentrated only in one region, suggesting that they are the result of parsimonious changes. When analyzed by TMHMM Server v. 2.0 (Krogh et al., 2001), none of the two translated ORFs showed transmembrane helices, suggesting

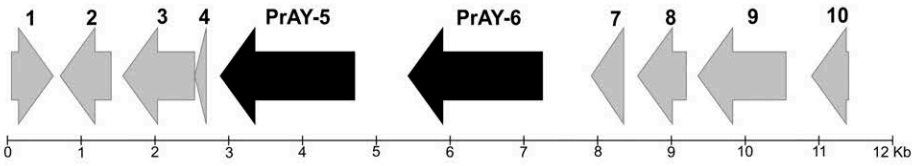


Fig. 1. Genomic neighborhood of two protease S8A genes found in an underground water metagenomic library. Schematic representation of the eight genes adjacent to *PrAY5* and *PrAY6*. Arrows represent the direction and length of genes.

Table 1
Identity of genes in the vicinity of two peptidase S8 genes found in an underground water metagenomic library, as predicted by fgenesB suite. Abbreviations: LTO, length of the translated open reading frame (amino acids); Q cover, query cover; E-value, expected value; Accession No., accession number in the NCBI. *PrAY5* and *PrAY6* are in bold letters.

LTO (aa)	Nearest homolog protein	Organism	Q cover (%)	E-value	Identity %	Accession No.
195	Hypothetical protein	*Algicola sagamiensis*	97%	4e−29	42%	WP_018692576.1
231	Hypothetical protein	*Chitinimonas koreensis*	96%	4e−44	45%	WP_051319001.1
332	Uncharacterized protein	*Achromobacter* sp.	92%	9e−117	56%	CUJ34213.1
61	Uncharacterized protein	*Achromobacter* sp.	98%	2e−22	82%	CUJ34213.1
614	**Peptidase S8**	***Chitinimonas koreensis***	**94%**	**0.0**	**60%**	**WP_028445624.1**
607	**Peptidase S8**	***Chitinimonas koreensis***	**100%**	**0.0**	**68%**	**WP_028445624.1**
150	GNAT family acetyltransferase	*Acidobacterium* sp. *PMMR2*	100%	5e−42	46%	WP_026441689.1
218	Hypothetical protein	*Chitinimonas koreensis*	91%	9e−83	60%	WP_051319279.1
397	Phosphatidylinositol-specific phospholipase C	*Photobacterium marinum*	84%	3e−117	54%	WP_007463772.1
166	Multispecies: hypothetical protein	*Acidovorax*	96%	1e−56	60%	WP_019703938.1

that they are not trans-membrane proteases. But by using the SignalP 4.1 server, it was noticed that both encode a signal peptide at their 5′ end, which likely guides the secretion of the enzymes out of the cell. The predicted cleavage position in PrAY5 is between amino acids 31 and 32, in the sequence AHA|EQ, and the predicted cleavage position in PrAY6 it is between amino acids 30 and 31, in the sequence AQA|DQ. When the translated sequences of *PrAY5* and *PrAY6* were analyzed by BLAST in the NCBI, both ORFs also showed a 101 aa C-terminal pro-peptidase (containing a PPC domain) at their 3′ end (Fig. 2), which is usually found in the immature stage of secreted peptidases. All these results suggest that the proteases encoded by *PrAY5* and *PrAY6* are secreted peptidases, consistent with the observed phenotype in clone MAX10. This also shows that *E. coli* has the molecular means for the proper folding, secretion, and processing, of these enzymes.

According to the NCBI dataset of conserved domains, the proteins encoded by *PrAY5* and *PrAY6* have a peptidase S8 domain, which contains the typical triad in serine proteases, formed by Asp/His/Ser located within the motifs D(S/T)G, HGTH, GTSMAXP, as well as the Asn oxyanion hole, in its catalytic domain (Fig. 2). This was corroborated by using the search tools HMMER, Pfam, and UniProt. When the two proteases were classified by using the specialized peptidase database MEROPS, they grouped within serine peptidases of the family S8, subfamily S8A, whose type enzyme is the Subtilisin Carlsberg (from the bacterium *Bacillus licheniformis*), which is an important commercial protease.

3.6. Phylogenetic analysis

A neighbor-joining phylogenetic tree comprising 26 peptidase sequences representative of the S8 family (subset cd00306 - NCBI conserved domain dataset), the three S8 peptidase sequences discovered hitherto by functional metagenomics, thirteen sequences homologous to PrAY5 and PrAY6 found in environmental sequences, PrAY5, and PrAY6, was constructed (Fig. 3). This NJ tree shows that PrAY5 and PrAY6 cluster together within the sub-family S8_13 (cd07496), which, according to the NCBI, is an uncharacterized subfamily. Interestingly, the other three S8 peptidases found by functional metagenomics spread in different S8 subfamilies, suggesting that this technology does not bias towards proteases of a specific subfamily. Although it is worth keeping in mind that the methodology we followed allows detecting only genes encoding proteases that are properly expressed/processed/secreted in *E. coli*. Thus, it is likely that there is some bias disfavoring genes encoding proteases with different requirements than those that *E. coli* can provide. PrAY5 and PrAY6 grouped also in different clades than all the homolog metagenomic sequences deposited so far in the NCBI database, strongly suggesting that the proteases identified in this work are different from those previously found by metagenomics. These results reinforce the hypothesis that the proteases identified in this work represent novel members of family S8, and indicate moreover that the premise about the underground water of the Yucatán Aquifer being a good source of an assortment of unknown microbial genomes is plausible, although further work would be necessary to be conclusive about this point.

3.7. Structural modeling

The best template found by I-TASSER to elaborate the 3D modeling for both, PrAY5 and PrAY6, was a protease of *Thermococcus kodakaraensis* -PDB file 3AFG- (Foophow et al., 2010), a species that interestingly belongs to archaea and not to bacteria. PrAY5 and 3AFG had 78% coverage, 25% identity, 2.40 Norm Z-score, and −0.48 C-score; while PrAY6 and 3AFG had 79% coverage, 28% identity, 2.42 Norm Z-score, and −0.66 C-score. Thus, in both cases, the coverage was good and the scores indicate that there is an acceptable alignment between the structures, meaning that the model is appropriate. Fig. 2 shows the models with the best C-score for each case.

To have a reference, the S8 protease of *Chitinimonas koreensis* (Ck) (WP_028445624.1), which according to the NCBI was the nearest homolog to both PrAY5 and to PrAY6, was also modeled with I-TASSER. In this case, the results were very similar to those found for PrAY5 and PrAY6: the best template for Ck was also 3AFG, and they had 79% coverage, 28% identity, 2.43 Norm Z-score, and −0.46 C-score. These results reinforce the conclusion that PDB: 3AFG is the best possible template existing so far in I-TASSER to model PrAY5 and PrAY6, even when it is not a bacterial protease, but an archaeal one. And they suggest that some bacterial and archaeal proteases have similar 3D structures, although their sequences may have a relatively low identity, as it has been discussed by some other authors (e.g. Siezen and Leunissen, 1997).

PrAY5 and PrAY6 modeling allowed furthermore to realize that both enzymes have a pro-peptide (between the signal peptide and the catalytic region), PrAY5 at Glu32-Leu113, and PrAY6 at Asp31-Leu109 (Fig. 2). It is common that peptidases S8 are synthesized as inactive zymogens (also called pre-pro-enzymes), which are guided by the signal peptide to the cell exterior, where they become active by release of the signal peptide and the pro-peptides. The pro-peptides have two roles, on one hand, they functions as chaperones (assisting in the correct folding of the enzyme); and on the other, they block the catalytic site (preventing the peptidase from damaging other proteins on its way out of the cell).

Modeling also shows that PrAY5 has 9 α-helices and 17 β-sheets, out of which 6 α-helices and 6 β-sheets are in the active site (Fig. S5A), and PrAY6 has 10 α-helices and 21 β-sheets, out of which 7 α-helices and 8 β-sheets are in its active site (Fig. S5B). According to the SCOPe structural classification of proteins, they also have the three-layer structure: α/β/α characteristic of peptidases S8 (Laskar et al., 2011).

Alignment between the template and the modeled sequences (Fig. S6) shows that the amino acids forming the catalytic triad are quite conserved in the three enzymes; being D186, H246, and S429 for PrAY5; D181, H242, and S423 for PrAY6; and D170, H203, and S382 for Ck; as well as the Asn in the oxyanion hole, corroborating that PrAY5 and PrAY6 are indeed proteases S8.

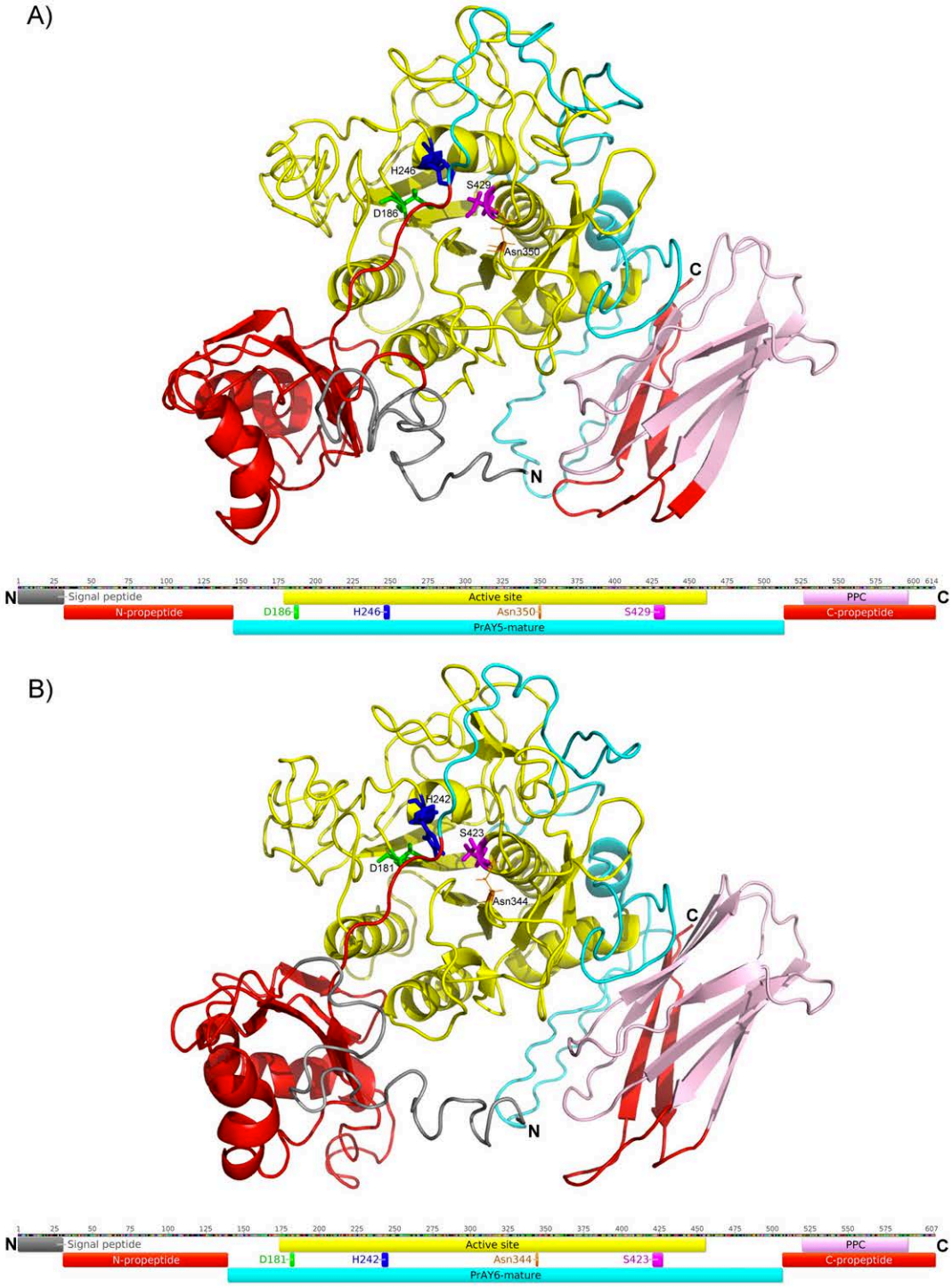
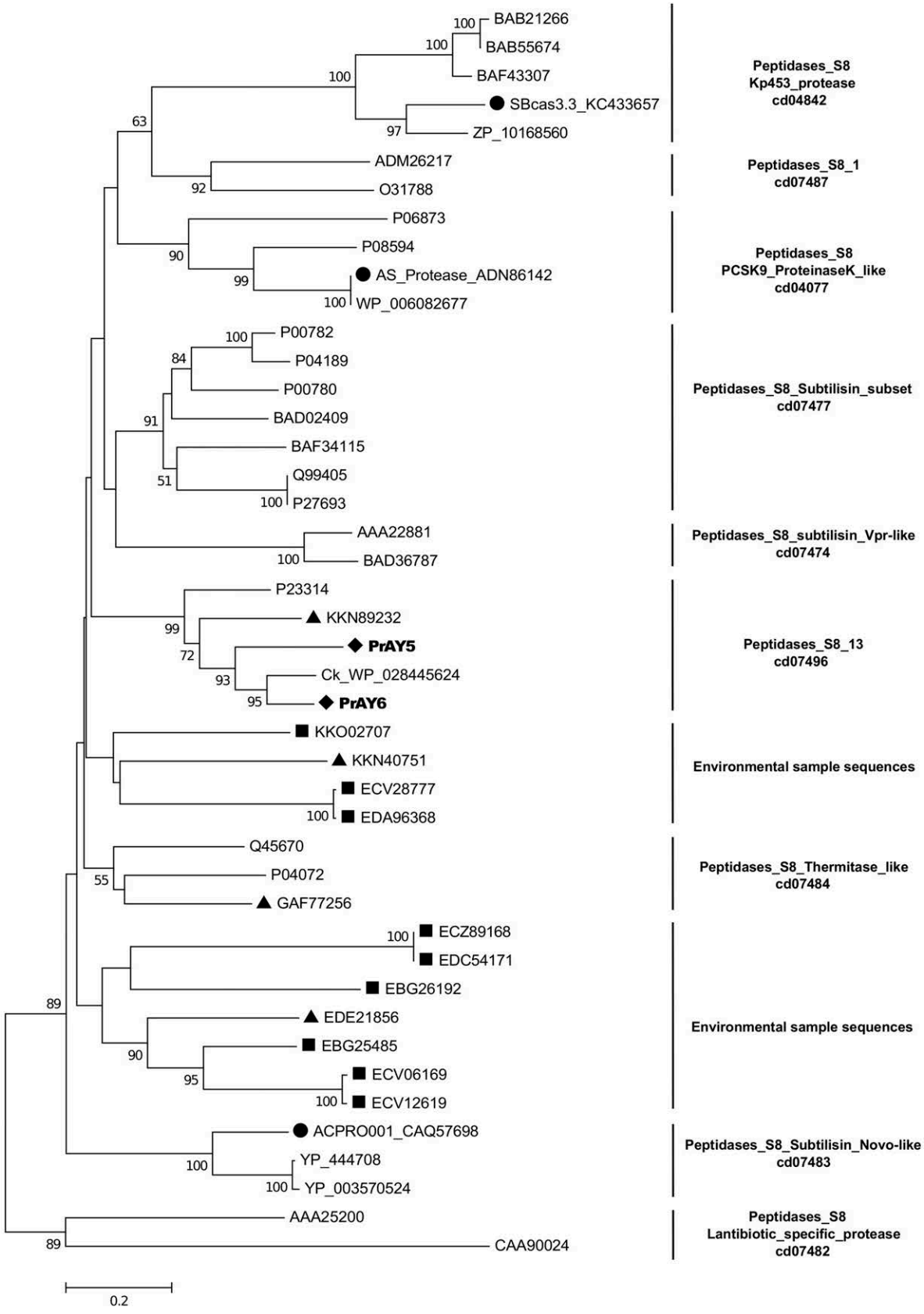


Fig. 2. PrAY5 and PrAY6 zymogens 3D modeling. Symbols: A. PrAY5-mature (145–513), cian; signal peptide (1−31), gray; N-pro-peptide (32–144) and C-pro-peptide (514–614), red; PPC domain (527–569), pink; active site (179–461), yellow. Amino acids forming the catalytic triad (D186, H246, S429) and the oxyanion hole (Asn350) are in green, blue, magenta and orange respectively. B. PrAY6 has the same code of colors as PrAY5. PrAY6-mature (140–506); Signal peptide (1−30); N-propeptide (31–139) and C-propeptide (507–607); PPC domain (520–589); Active site (174–455); catalytic triad (D181, H242, S423); oxyanion hole (Asn344). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 3. Phylogenetic relation between PrAY5, PrAY6, members of the nine major S8 peptidase subfamilies (according to CDD), the three peptidase S8 genes found so far by functional metagenomics, and environmental sequences annotated as putative Subtilase S8 genes. Major S8 peptidase subfamilies were taken from the conserved protein domain family subset cd00306 (Siezen and Leunissen, 1997). Only those peptidase S8 genes found by functional metagenomics which are well characterized were considered for the analysis (●). PrAY5 and PrAY6 sequences (♦) were used as query for BLASTp against all the metagenomic sequences reported in the NCBI-Metagenomic protein (env_nr) dataset until now. Results were filtered, picking only those with >70% cover, E-value <10e−6, and having the catalytic amino-acid triad (D/H/S) which is characteristic of the peptidases S8. When PrAY5 was used as query, four peptidase sequences fulfilling these criteria were found (▲), and when PrAY6 was used, there were nine peptidases (■). The tree was built using the neighbor-joining method with a Poisson correction model. Bootstrap values are expressed as percentages of 1000 replications and are shown at the nodes. Only bootstrap values higher than 50% are indicated. Analysis was made with MEGA6 (Tamura et al., 2013).

Overlap of 3D models of these three sequences (Fig. S7A, B) shows that their structure at the N-terminus has important differences, while their structure at the C-terminus is quite conserved, and the active site is practically identical (0.4 Å in 64 residues). This suggests that at structural level, there is a high selective pressure favoring the conservation of the active site, but at the lateral regions more variation is permitted. When zooming in the catalytic triad (Fig. S7C), it can be observed that this, as well as the oxyanion hole, is entirely conserved. This structural conservation at the catalytic triad in proteases S8 has been reported previously by some authors (e.g. Rawlings and Barrett, 2004; Laskar et al.,



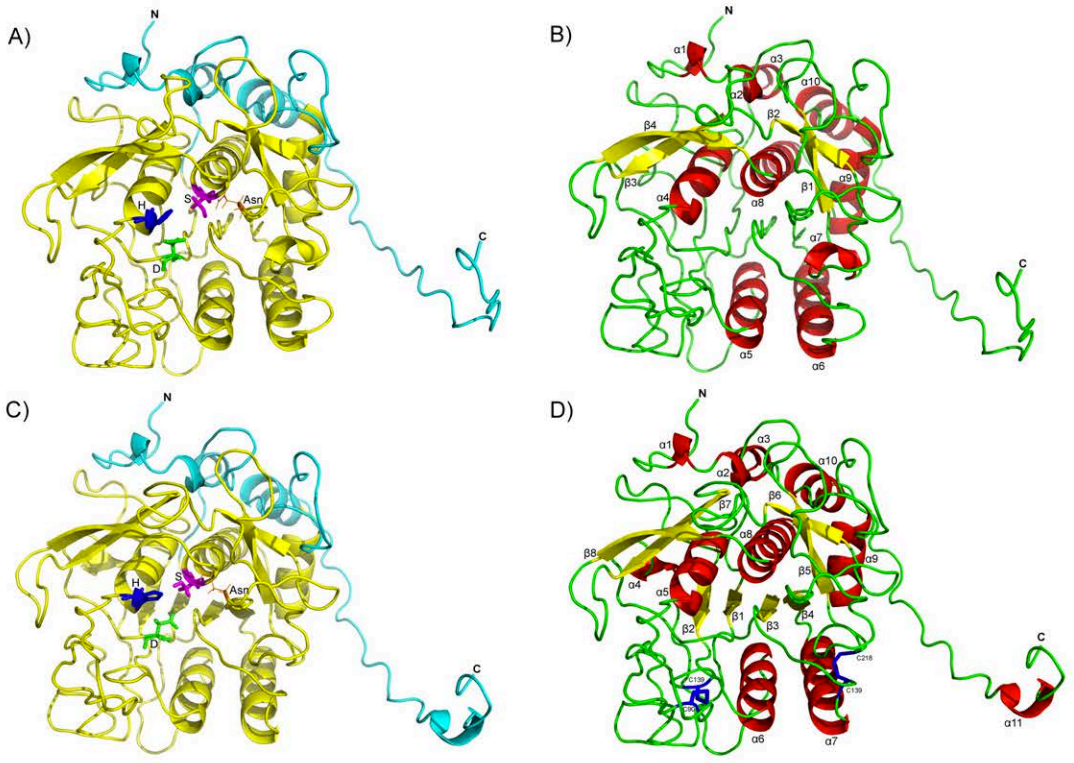


Fig. 4. PrAY5 and PrAY6 mature enzymes 3D modeling. A. PrAY5 catalytic triad and oxyanion hole. B. PrAY5 β-sheets, yellow; and α-helices, red. C. PrAY6 catalytic triad and oxyanion hole. D. PrAY6 β-sheets, yellow; α-helices, red; and disulphide-tethered (C90–C139 and C181–C218), blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2011, Zhang et al., 2015), thus this is an additional evidence of PrAY5 and PrAY6 being proteases S8. An interesting difference between PrAY5 and PrAY6, in comparison with other proteases S8, is that while the catalytic site of proteases S8 typically comprises three layers of seven-stranded β-sheets between two layers of α-helices (Rawlings and Barrett, 2004; Kennan et al., 2010; Laskar et al., 2011), PrAY5 modeling showed three layers of six-stranded β-sheets and PrAY6 showed three layers of eight-stranded β-sheets between the two layers of α helices. However, it is worth keeping in mind that there can be some artifacts of modeling, because although 3AFG is the best template available so far, it is likely that it is not totally faultless. A crystallographic analysis would be necessary to be totally certain about the number of β-sheets in PrAY5 and PrAY6.

The nearest homolog PDBs found by I-TASSER showed that possibly the EC number for PrAY5 and PrAY6 is 3.4.21.62 (with a C-score of 0.207 and 0.200 respectively), meaning that they are subtilisins, and therefore, it is likely that they participate in the hydrolysis of proteins with broad specificity for peptide bonds, and have a preference for a large uncharged residue in P1 (amino acid residues in a substrate undergoing cleavage are designated P1, P2, P3, etc. in the N-terminal direction from the cleaved bond; and likewise, the residues in C-terminal direction are designated P1′, P2′, P3′, etc. –Schechter and Berger, 1967-). It is worth mentioning that the protease of *Thermococcus kodakaraensis*, which resulted to be the best template found by I-TASSER to elaborate the 3D modeling for PrAY5 and PrAY6, is a hyperthermostable peptidase S8, which has two propeptides –one in the N-terminus and one in the C-terminus- very similar to those found in PrAY5 and PrAY6.

When PraY5 and PraY6 core sequences were 3D modeled (Fig. 4), the best template found by I-TASSER was a subtilisin-like protease of *Dichelobacter nodosus*, named AprV2 (PDB file 3LPA), a Gram-negative, anaerobic Gammaproteobacteria, which is the principal causative agent of ovine footrot, a debilitating disease of the hoof of ruminants (Kennan et al., 2010). AprV2 is synthesized as an inactive precursor very similar to those of PrAY5 and PrAY6, having an N-terminal pre-pro-region, a core, and a C-terminal domain. PrAY5 and 3LPA had 89% coverage, 51% identity, 3.30 Norm Z-score, and 1.40 C-score; while PrAY6 and 3LPA had 90% coverage, 48% identity, 3.42 Norm Z-score, and 1.03 C-score. Thus, in both cases, the coverage and scores were even better that those obtained when PraY5 and PraY6 were compared with 3AFG, meaning that also in this case the models are acceptable.

AprV2 adopts a subtilisin-like fold, consisting of a curved six-stranded parallel β-sheets sandwiched between two α-helices at one side and five α-helices at the other, and it has a two stranded anti-parallel β hairpin, which runs perpendicular to the plane of the central β-sheets. It contains a catalytic triad formed by D41, H105, and S277; and furthermore, AprV2 has two unusual extended disulphide-tethered loops, which function as exosites, mediating effective enzyme-substrate interactions (Kennan et al., 2010). When PraY5 and PraY6 core sequence models were compared with 3LPA, on one hand it was observed that the catalytic triad was very similar in the three cases (Fig. 4A–C); and on the other hand, it was noticed that while PraY6 and 3LPA had the six-stranded parallel β-sheets sandwiched between two and five α-helices, PraY5 had only two parallel β-sheets at that place. Both, PraY5 and PraY6, had the two stranded anti-parallel β hairpin present in AprV2. Another interesting observation was that PrAY6 model shows two disulphide-tethered loops almost in the same positions as those present in AprV2 (Fig. 4D), while PrAY5 lacks these exosites. Possibly this difference between PrAY5 and PrAY6 might explain why PrAY6 clones show a higher activity than PrAY5 clones in the milk test; although several experiments would be necessary to prove this hypothesis.

4. Conclusion

In the work described here, two novel genes encoding secreted proteases S8 were discovered by functional metagenomic techniques, and they and the enzymes they encode were characterized in silico.

This work provides the necessary knowledge to support and guide further work to isolate and biochemically characterize these proteases;

a task that could not be accomplished without the information about the structure of the genes and the enzymes they encode, obtained in this study.

The results can also be used as a reference for further work in which genes and enzymes with similar characteristics are found, considering the complexity of their zymogens.

It is worth mentioning that the relatively far phylogenetic distance between PrAY5 and PrAY6, and all the other proteases S8 found hitherto by metagenomic techniques, suggests that the tropical underground water of the Yucatán Aquifer may contain microbial communities considerably different –and therefore interesting- from the other environments explored so far –mainly at higher latitudes-. Thus, it might be worth to conduct further bioprospecting studies searching for genes encoding novel enzymes, using the microbial biomass of this until now neglected water body.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2016.08.009.

Conflicts of interest

The authors have not any conflict of interest.

Acknowledgments

The authors would like to thank Reynaldo C Pless-Elling for his suggestions and edits to the manuscript, and to MC Jorge Arturo Dominguez-Maldonado for his technical assistance. The work for this publication was partially supported by the CONACYT-Gobierno del Estado de Yucatán Grant 165026, CONACYT Grant 269833, and CICY Fiscal Fund 1039200015.

References

Amann, R.I., Ludwig, W., Schleifer, K.H., 1995. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. Microbiol. Rev. 59, 143–169.

Bauer-Gottwein, P., Gondwe, B.R.N., Charvet, G., Marín, L.E., Rebolledo-Vieyra, M., Merediz-Alonso, G., 2011. Review: the Yucatán Peninsula karst aquifer, Mexico. Hydrogeol J. 19 (3), 507–524. http://dx.doi.org/10.1007/s10040-010-0699-5.

Berlemont, R., Pipers, D., Delsaute, M., Angiono, F., Feller, G., Galleni, M., Power, P., 2011. Exploring the Antarctic soil metagenome as a source of novel cold-adapted enzymes and genetic mobile elements. Rev. Argent. Microbiol. 43 (2), 94–103. http://dx.doi.org/10.1590/S0325-75412011000200005.

Biver, S., Portetelle, D., Vandenbol, M., 2013. Characterization of a new oxidant-stable serine protease isolated by functional metagenomics. Springerplus 2 (410), 1–10.

Chistoserdova, 2009. Functional metagenomics: recent advances and future challenges. Biotechnol. Genet. Eng. Rev. 26 (335), 352.

DeLano, W.L., 2002. The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, USA.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Foophow, T., Tanaka, S., Angkawidjaja, C., Koga, Y., Takano, K., Kanaya, S., 2010. Crystal structure of a subtilisin homologue, Tk-SP, from *Thermococcus kodakaraensis*: requirement of a C-terminal beta-jelly roll domain for hyperstability. J. Mol. Biol. 400 (4), 865–877. http://dx.doi.org/10.1016/j.jmb.2010.05.064 (23).

Guazzaroni, M.E., Morgante, V., Mirete, S., González-Pastor, J.E., 2013. Novel acid resistance genes from the metagenome of the Tinto River, an extremely acidic environment. Environ. Microbiol. 15 (4), 1088–1102. http://dx.doi.org/10.1111/1462-2920.12021.

Gupta, R., Beg, Q.K., Lorenz, P., 2002. Bacterial alkaline proteases: molecular approaches and industrial applications. Appl. Microbiol. Biotechnol. 59, 15–32. http://dx.doi.org/10.1007/s00253-002-0975-y.

Hjort, K., Presti, I., Elväng, A., Marinelli, F., Sjöling, S., 2014. Bacterial chitinase with phytopathogen control capacity from suppressive soil revealed by functional metagenomics. Appl. Microbiol. Biotechnol. 98, 2819–2828. http://dx.doi.org/10.1007/s00253-013-5287-x.

Kennan, R.M., Wong, W., Dhungyel, O.P., Han, X., Wong, D., et al., 2010. The subtilisin-like protease AprV2 is required for virulence and uses a novel disulphide-tethered exosite to bind substrates. PLoS Pathog. 6 (11), e1001210. http://dx.doi.org/10.1371/journal.ppat.1001210.

Kim, B.Y., Weon, H.Y., Yoo, S.H., Chen, W.M., Kwon, S.W., Go, S.J., Stackebrandt, E., 2006. *Chitinimonas koreensis* sp. nov., isolated from greenhouse soil in Korea. Int. J. Syst. Evol. Microbiol. 56, 1761–1764. http://dx.doi.org/10.1099/ijs.0.64163-0.

Konagurthu, A.S., Reboul, C.F., Schmidberger, J.S., Irving, J.A., Lesk, A.M., Stuckey, P.J., Whisstock, J.C., Buckle, A.M., 2010. MUSTANG-MR structural sieving server: applications in protein structural analysis and crystallography. PLoS One 5 (4), e10048 (Apr 6).

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305 (3), 567–580. http://dx.doi.org/10.1006/jmbi.2000.4315.

Laskar, M., James, R.E., Chatterjee, A., Mandal, C., 2011. Modeling and structural analysis of evolutionarily diverse S8 family serine proteases. Bioinformation 7 (5), 239–245.

Lee, D.G., Jeon, J.H., Jang, M.K., Kim, N.Y., Lee, J.H., Kim, S.J., Kim, G.D., Lee, S.H., 2007. Screening and characterization of a novel fibrinolytic metalloprotease from a metagenomic library. Biotechnol. Lett. 29 (3), 465–472. http://dx.doi.org/10.1007/s10529-006-9263-8.

Mira, A., Pushker, R., Rodriguez-Valera, F., 2006. The neolithic revolution of bacterial genomes. Trends Microbiol. 14 (5), 200–206.

Morris, L.S., Marchesi, J.R., 2015. Current functional metagenomic approaches only expand the existing protease sequence space, but does not presently add any novelty to it. Curr. Microbiol. 70 (1), 19–26. http://dx.doi.org/10.1007/s00284-014-0677-6.

Neveu, J., Regeard, C., DuBow, M.S., 2011. Isolation and characterization of two serine proteases from metagenomic libraries of the Gobi and Death Valley deserts. Appl. Microbiol. Biotechnol. 91 (3), 635–644. http://dx.doi.org/10.1007/s00253-011-3256-9.

Niehaus, F., Gabor, E., Wieland, S., Siegert, P., Maurer, K.H., Eck, J., 2011. Enzymes for the laundry industries: tapping the vast metagenomic pool of alkaline proteases. Microb. Biotechnol. 4 (6), 767–776. http://dx.doi.org/10.1111/j.1751-7915.2011.00279.x.

Privé, F., Newbold, C.J., Kaderbhai, N.N., Girdwood, S.G., Golyshina, O.V., GolyshinPN, S.N.D., Huws, S.A., 2015. Isolation and characterization of novel lipases/esterases from a bovine rumen metagenome. Appl. Microbiol. Biotechnol. 99, 5475–5485.

Purohit, M.K., Singh, S.P., 2013. A metagenomic alkaline protease from saline habitat: cloning, over-expression and functional attributes. Int. J. Biol. Macromol. 53, 138–143.

Pushpam, P.L., Rajesh, T., Gunasekaran, P., 2011. Identification and characterization of alkaline serine protease from goat skin surface metagenome. AMB Express 1 (3), 1–10.

Rao, M.B., Tanksale, A.M., Ghatge, M., Deshpande, V.V., 1998. Molecular and biotechnological aspects of microbial proteases. Microbiol. Mol. Biol. Rev. 62 (3), 597–635.

Rawlings, N.D., Barrett, A.J., 2004. Introduction: serine peptides and their clans. In: Barrett, A.J., Rawlings, N.D., Woessner, J.F. (Eds.), Handbook of Proteolytic Enzymes, second ed. Elsevier, London, United Kingdom, pp. 1425–1427.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Salzberg, S.L., Delcher, A.L., Kasif, S., White, O., 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 26 (2), 544–548. http://dx.doi.org/10.1093/nar/26.2.544.

Sawant, R., Nagendran, S., 2014. Protease: an enzyme with multiple industrial applications. World J. Pharm. Sci. 3 (6), 568–579.

Schechter, I., Berger, A., 1967. On the size of the active site in proteases. I. Papain. Biochem. Biophys. Res. Commun. 27 (2), 157–162. http://dx.doi.org/10.1016/S0006-291X(67)80055-X.

Siezen, R.J., Leunissen, J.A., 1997. Subtilases: the superfamily of subtilisin-like serine proteases. Protein Sci. 6 (3), 501–523. http://dx.doi.org/10.1002/pro.5560060301.

Solovyev, V., Salamov, A., 2011. Automatic annotation of microbial genomes and metagenomic sequences. In: Li, R.W. (Ed.), Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies. Nova Science Publishers, pp. 61–78.

Tamura, K., Stecher, G., Peterson, D., Alan Filipski, A., Kuma, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. http://dx.doi.org/10.1093/molbev/mst197.

Torsvik, V., Goksøyr, J., Daae, F.L., 1990. High diversity in DNA of soil bacteria. Appl. Environ. Microbiol. 56 (3), 782–787.

Vester, J.K., Glaring, M.A., Stougaard, P., 2015. An exceptionally cold-adapted alpha-amylase from a metagenomic library of a cold and alkaline environment. Appl. Microbiol. Biotechnol. 99, 717–727. http://dx.doi.org/10.1007/s00253-014-5931-0.

Waschkowitz, T., Rockstroh, S., Daniel, R., 2009. Isolation and characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries. Appl. Microbiol. Biotechnol. 75 (8), 2506–2516. http://dx.doi.org/10.1128/AEM.02136-08.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., Zhang, Y., 2015. The I-TASSER Suite: protein structure and function prediction. Nat. Methods 12, 7–8.

Zhang, Y., Zhao, J., Zeng, R., 2011. Expression and characterization of a novel mesophilic protease from metagenomic library derived from Antarctic coastal sediment. Extremophiles 15 (1), 23–29. http://dx.doi.org/10.1007/s00792-010-0332-5.

Zhang, Y., Ran, L., Li, C., Chen, X., 2015. Diversity, structures, and collagen-degrading mechanisms of bacterial collagenolytic proteases. Appl. Environ. Microbiol. 81, 6098–6107. http://dx.doi.org/10.1128/AEM.00883-15.

Zheng, F., Angleton, E.L., Lu, J., Peng, S.B., 2002. *In vitro* and *in vivo* self-cleavage of *Streptococcus pneumonia* signal peptidase I. Eur. J. Biochem. 269, 3969–3977. http://dx.doi.org/10.1046/j.1432-1033.2002.03083.x.

Zuckerkandl, E., Pauling, L., 1965. Evolutionary divergence and convergence in proteins. In: Bryson, V., Vogel, H.J. (Eds.), Edited in Evolving Genes and Proteins. Academic Press, New York, pp. 97–166.