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).