Data mining of metagenomes to find novel enzymes: a non-computationally intensive method

Elsa Góngora-Castillo¹ · Luisa A. López-Ochoa² · Max M. Apolinar-Hernández^{3,4} · Aldo M. Caamal-Pech³ · Perla A. Contreras-de la Rosa³ · Adriana Quiroz-Moreno³ · Jorge H. Ramírez-Prado³ · Aileen O'Connor-Sánchez³

Received: 16 August 2019 / Accepted: 28 December 2019 / Published online: 30 January 2020
© King Abdulaziz City for Science and Technology 2020

Abstract

Currently, there is a need of non-computationally-intensive bioinformatics tools to cope with the increase of large datasets produced by Next Generation Sequencing technologies. We present a simple and robust bioinformatics pipeline to search for novel enzymes in metagenomic sequences. The strategy is based on pattern searching using as reference conserved motifs coded as regular expressions. As a case study, we applied this scheme to search for novel proteases S8A in a publicly available metagenome. Briefly, (1) the metagenome was assembled and translated into amino acids; (2) patterns were matched using regular expressions; (3) retrieved sequences were annotated; and (4) diversity analyses were conducted. Following this pipeline, we were able to identify nine sequences containing an S8 catalytic triad, starting from a metagenome containing 9,921,136 Illumina reads. Identity of these nine sequences was confirmed by BLASTp against databases at NCBI and MEROPS. Identities ranged from 62 to 89% to their respective nearest ortholog, which belonged to phyla Proteobacteria, Actinobacteria, Planctomycetes, Bacteroidetes, and Cyanobacteria, consistent with the most abundant phyla reported for this metagenome. All these results support the idea that they all are novel S8 sequences and strongly suggest that our methodology is robust and suitable to detect novel enzymes.

Keywords Proteases · NGS · Bioinformatics pipeline · Pattern matching

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13205-019-2044-6>) contains supplementary material, which is available to authorized users.

✉ Jorge H. Ramírez-Prado
jhramirez@cicy.mx

✉ Aileen O'Connor-Sánchez
aileen@cicy.mx

¹ Unidad de Biotecnología, CONACYT, Centro de Investigación Científica de Yucatán, A. C., Mérida, Yucatán, Mexico

² Unidad de Bioquímica y Biología Molecular, Centro de Investigación Científica de Yucatán, A. C., Mérida, Yucatán, Mexico

³ Unidad de Biotecnología, Centro de Investigación Científica de Yucatán, A. C., Mérida, Yucatán, Mexico

⁴ Present Address: Instituto de Biotecnología, Universidad Autónoma de Nuevo León, San Nicolás de los Garza, Nuevo León, Mexico

Introduction

In recent years, Next Generation Sequencing (NGS) technologies have revolutionized the field of metagenomics by sequencing millions of DNA fragments at low cost and high accuracy. Consequently, large amounts of data have been generated at a remarkable rate and there is a prevailing need for bioinformatics tools to handle and manage these large datasets (Shendure et al. 2017).

The advent of NGS technologies made a significant contribution to metagenomics providing huge amounts of sequences that can be used as indicators of the structure, diversity, and possible functions of microbial communities. Thus, NGS opens a range of possibilities to infer the relation of a microbiome and its environment, e.g., studies of plant–microbe interactions for a better interpretation of the dynamics in rhizospheres and phyllospheres, studies in water quality to better follow pollution cycles, or the human microbiome project that has allowed to open new frontiers for a better understanding of epidemics outbreaks

such as Zika virus in South and Central America in 2015 (Thézé et al. 2018).

Despite the advantages offered by NGS technologies, and the valuable and abundant information generated, there are still many challenges to overcome. “Doing metagenomics” requires specialized hardware infrastructure and knowledge to manage enormous amounts of sequences. This has become a bottleneck for many research groups, thus there is an urgent need to develop tools and bioinformatics strategies that allow researchers to exploit the full potential of all these datasets. Initiatives such as MG-RAST (Keegan et al. 2016), EBI Metagenomics (Mitchell et al. 2018), IMG/MER (Chen et al. 2019), etc., provide computational facilities to perform analysis of metagenomic samples, mainly for Quality Control (QC) and filtering of the sequences, grouping, taxonomic identification, functional annotation, storage, and sharing. However, bioinformatics tools for downstream analysis, including identification of novel genes and their diversity which help to reduce the number of orphan sequences or unknown genes, are not available.

In this work, we present a simple and robust bioinformatics strategy to identify putative enzymes using a pattern recognition approach. As a case study we applied the pipeline to search for protease S8A genes in a set of metagenomic sequences derived from an aquatic environment. The strategy is based on pattern searching using as reference the three conserved motifs that protease S8A genes have. Briefly, proteases S8A, also called subtilisin-like peptidases, belong to a proteases family which has a wide range of important and useful applications, mainly for pharmaceutical and chemical industries, and have an important commercial value due to their wide catalytic activity (Jisha et al. 2013). These enzymes are characterized by three conserved motifs known as “catalytic triad” (CT). Each motif length range from 7 to 12 amino acids (AA) and contains a crucial AA for the catalytic activity [Aspartate (D), Histidine (H), and Serine (S)]. Two regions variable in sequence and size link the three conserved motifs. The region between D and H is shorter than the one between H and S, such characteristic allows the architecture of the catalytic domain formed by the α -helix, β -sheets and loops (Laskar et al. 2011).

Important efforts have been made to find new Proteases S8A with the potential to better fit some specific industrial applications. However, most of these efforts have been limited to cultivable bacterial strains, and as the majority of bacteria are not cultivable, the search has been very restricted (Amann et al. 1995). Experimental functional metagenomic approaches may overcome this restriction, allowing cloning and expression of genes from uncultivable microorganisms. Nevertheless, most experimental functional metagenomic procedures involve expensive and laborious methodologies that are not accessible to many research groups.

In this study, we present an original non-computationally-intensive bioinformatics data-mining strategy as an alternative to explore the diversity of novel enzymes in metagenomes, showcased with the identification of putative new protease S8 genes from an aquatic sample.

Methods

Informatics equipment

Bioinformatics analyses described in this study were carried out in a 64 bit desktop machine with an Intel Core i7 processor, 16 Gb of RAM, and the Ubuntu (v18.04) Linux Operating System.

Metagenomic data

Metagenomic sequences used in this work were obtained from the publicly available MG-RAST database (Keegan et al. 2016), ID 4536384.3, and name X10-JUL09, which were obtained from mesotrophic water of a Yucatan cenote (20°90'95.67" N/88°86'69.47" W).

Metagenome assembling and translation into amino acids

Illumina raw reads were checked for quality using FastQC (v 0.11.5) (Andrews 2010) and assembled into contigs using Newbler (v2.9) (Margulies et al. 2005) or Meta-Hit (v 1.14) (Nielsen et al. 2014) software. To account for possible assembly artifacts (e.g., chimeras) inherent to the algorithm used, two independent assemblies were generated using either the Newbler algorithm with default parameters or Meta-Hit algorithm with default parameters and a minimum contig length of 500 bp. Downstream analysis were carried out independently for each resulting assembly.

The assembled contigs were translated to amino acids (AA) into the six reading frames (+1, +2, +3, −1, −2, −3) using the “transeq” tool from EMBOSS Toolkit (v 6.6.0.0) (Rice et al. 2000) by typing the following command line in a terminal window:

```
$transeq -sequence metagenome_contigs.fna -outseq metagenome_aa_file.faa -frame 6 -Table 11 -nomethionine
```

The key parameters were: *-sequence* flag that specifies the input file containing the metagenomic sequences in fasta format; *-outseq* flag specifies the output file; *-frame* flag specifies the reading frame (6: all forward and reverse frames); *-table* flag specifies the table number for codon use (11: Bacterial); and *-nomethionine* flag avoids to always translate the first codon to methionine. The output is a multi-fasta file containing the contigs sequences translated into all

six reading frames. Depending on the assembly coverage, the translated contigs may contain partial protein sequences, one single full protein, or groups of varying numbers of partial and full proteins joined by translated intergenic regions. Authors note: Readers are advised to type (or copy from supplemental files) the command lines since due to type-setting non-standard characters may have been introduced.

Before pattern searching of translated contigs, “line breaks or new lines” in the multi-fasta file were removed using the following *perl* command:

```
$cat metagenome_aa_file.faa | perl
-ne 'if (/^>/) {print "\n", $_; next;}
else {chomp; print;}' > lineal_metagenome_
aa_file.faa
```

Briefly, the “cat” command passes the content of the metagenome_aa_file.faa into perl command, which removes all new lines from the end of the lines that do not begin with the “>” symbol. The resulting file is written to lineal_metagenome_aa_file.faa.

Pattern matching using regular expressions

As the three motifs which shape the characteristic proteases S8A catalytic triad (CT) are well conserved, patterns for the most conserved positions were designed according to the diversity alignments reported in the conserved domains database of NCBI (<https://www.ncbi.nlm.nih.gov/>) and from thoroughly curated proteases S8 sequences (34 type S8A and 3 type S8B sequences) from the MEROPS database aided by the Motif discovery tool of the MEME suite (Fig. 1a) (Bailey et al. 2015; Rawlings et al. 2018). These patterns, also known as regular expressions (RegExp), were built for each motif:

Aspartic acid (D) motif: [VIA][^T][VIFGL][LIVAF][D][TSADG][GDPS]

Histidine (H) motif: [H][GIA][TSDNCM][^R][VCTLIA][AISTG][GSHAL]

Serine (S) motif: [G][TN][S][^A][ASG][STAVCLG][PAG]

Characters inside square brackets match any (but only one) of the corresponding AA (one letter IUPAC code) in that position. Square brackets containing a single character preceded by the metacharacter “^” indicate that any AA but that one can occupy that position and are intended to restrict the matching to S8A proteases.

The regular expressions for the two non-conserved regions of variable size (VR) separating the motifs between D and H, and H and S (henceforth referred to as VR1 and VR2), are constructed as ‘[A-Z]{10,100}’ for VR1 and ‘[A-Z]{100,250}’ for VR2. The characters inside the brackets represents any AA and numbers indicate that

VR1 and VR2 should not be shorter than 10 and 100 AA, respectively, and not larger than 100 and 250 AA in length, respectively.

A single regular expression was created using the three motifs joined by VR1 and VR2 to identify putative protease S8A sequences. Searches in the metagenome based on pattern matching of the regular expression, were performed using the Linux command “grep” typing the following command line in a terminal window:

```
$grep -E -B 1 -no-group-separator
'[VIA][^T][VIFGL][LIVAF][D][TSADG]
[GDPS][A-Z]{10,100}[H][GIA][TSDNCM][^R]
[VCTLIA][AISTG][GSHAL][A-Z]{100,250}[G]
[TN][S][^A][ASG][STAVCLG][PAG]' lineal_
metagenome_aa_file.faa > contigs_with_
S8A_proteases.faa
```

The previous command line recovers any translated contig that contains the complete sequence of the CT of proteases S8A. Since the contigs may contain more than one single coding sequence (CDS), the following *perl* command was used to trim the sequences and extract only the S8A conserved region of the CT. The regular expression identifies the first amino acid for the ‘D’ motif ([VIA]) and the last amino acid for the ‘S’ motif ([PAG]). The resulting subsequences were written into a file that includes the corresponding FASTA header identifier for each subsequence. If a contig contains multiple proteases S8A, each CT will be individually trimmed and a unique counter will be added to its corresponding FASTA header:

```
$perl -lane '{ $c=0 } if (/>/) { $h=$_;
$h=~s/>\/ } else { while (/[VIA][^T]
[VIFGL][LIVAF][D][TSADG][GDPS][A-Z]
{10,100}[H][GIA][TSDNCM][^R][VCTLIA]
[AISTG][GSHAL][A-Z]{100,250}[G][TN][S]
[^A][ASG][STAVCLG][PAG])/g) { $c++
print ">$c\_h\n$1" } } <contigs_with_
S8A_proteases.faa > trimmed_S8A_triads.
faa
```

A bash file including all the commands in sequential order of execution, from translation of contigs to retrieval of the S8A CT subsequences, is provided (Supplementary File 1).

Functional annotation

To identify the nearest orthologs of the putative proteases S8A CT, a BLASTp search was performed aligning the CT subsequences against the NR database of NCBI using the online BLAST web interface. Additionally, the CT subsequences were aligned against the family S8A from the specialized database MEROPS (<https://www.ebi.ac.uk/merops/index.shtml>) (Rawlings et al. 2018).

The sequences from MEROPS database were downloaded in fasta format and converted to a blast database using the “makeblastdb” command in the terminal window:

```
$makeblastdb -in family_S8A_sequences.fasta -input_type fasta -dbtype prot -out database_name
```

The key options were: *-in* flag specifies the input fasta file; *-input_type* flag specifies the input file format; *-dbtype* flag specifies whether is a nucleotide or protein database; *-out* flag specifies the base name for the output file.

The local “blastp” search was executed using the following options in the command line: *-db* flag that specifies

Fig. 1 Sequence analysis of the nine S8A catalytic triad subsequences. **a** Sequence logos showing the conservation or variability of amino acids at the “D”, “H”, and “S” motifs (generated at MEME suite online from MEROPS proteases S8A collection) (Bailey et al. 2015). Underneath each logo is a corresponding regular expression designed in this study (see “Methods”). **b** Fragments of the Multiple Sequence Alignment (MSA) depicting similarity of the three motifs of the catalytic triad of protease S8 to their closest reported homologs. X 100% similar, X 80 to 100% similar, X 60 to 80% similar, X less than 60% similar. **c** Maximum Likelihood phylogenetic inference of the relationships between recovered putative proteases and reported proteases S8. Black circle putative proteases from metagenome; black diamond best hits against NR database; black square best hits against MEROPS database; black down-pointing triangle subtilisin Carlsberg protease enzyme; Homologs of Subtilisin Carlsberg protease enzyme (unmarked); white diamond Proteases S8B. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site (Bold numbers above branches: aLRT SH-like node support. The analysis involved 56 amino acid sequences. There were a total of 581 positions in the final dataset. Starting tree, BIONJ; Type of tree improvement, SPR; Branch support, aLRT SH-like. Model, Blosom62 +G +I +F; Gamma shape parameter, 1.568; Proportion of invariable sites, 0.027; Equilibrium of frequencies, empirical; Number of substitutions rate categories, 4

the database name; *-query* flag that specifies the query sequences, in this case, the putative proteases sequences; *-out* flag that specifies the output file.

```
$ blastp -db base_name -query S8A_
triad.faa -out output_file.txt
```

Sequence pairwise alignments that showed an identity percentage equal or greater than 50% were selected. For sequence alignments with multiple hits, only the best match was selected.

Diversity analyses

A phylogenetic inference was calculated to analyze the diversity among the recovered putative proteases S8A and its relationships to known proteases S8 (from NCBI's NR and MEROPS databases). To minimize the computational burdens, the following steps were carried out at publicly available online servers: (1) a multiple sequence alignment (MSA) was generated from the trimmed CT of the putative proteases S8A (MGR4536384.3_01 to MGR4536384.3_09), the closest hits from the NR database (KRO47350.1, KRO47990.1, OYW49734.1, PHY00594.1, PKP34958.1, WP_073617898.1, and WP_092886438.1), the closest hits from the MEROPS database (MER0062435, MER0240645, MER0401707, MER0967697, MER0971069, MER0973676, MER0974284, and MER0979041), and the subtilisin Carlsberg sequence (*Bacillus licheniformis*; MER0000309) (DeLange and Smith 1967). The MSA was constructed using the online version of the MAFFT program (<https://mafft.cbrc.jp/alignment/server/>) (Katoh and Standley 2013) with the default parameters plus its E-INS-i iterative method option, recommended for regions with multiple conserved

domains/motifs interspersed by long varying gaps. (2) A second MSA was constructed using the above sequences plus homologs sequences to the subtilisin Carlsberg derived from several phyla (CUU34058.1, OHD15664.1, PKK94069.1, PLX70644.1, PYQ12951.1, RLD94087.1, RMH56527.1, RPJ38856.1, SLB85922.1, TET56434.1, TFG76767.1, TFG78317.1, TFH13182.1, TMP97196.1, WP_018963760.1, WP_099072880.1, WP_099771610.1, WP_131353865.1, WP_133958236.1, WP_1371681.1, WP_144840858.1, MQA92622.1, HBH61597.1, HAO98718.1, and HAH31981.1), and sequences from S8B family as controls (P13134.1, P09958.2, and P29120.2). Both MSAs were visually inspected to confirm that the “D”, “H”, and “S” motifs were correctly aligned. (3) The file of the second MSA was exported to PHYLIP format. (4) A Maximum Likelihood phylogenetic tree was inferred from the MSA using the online server for PhyML (<http://www.atgc-montpellier.fr/phyml/>) (Guindon et al. 2010). The best-fit substitution model was calculated using the “Smart Model Selection” tool included on the PhyML server (chosen by the Akaike Information Criterion). The phylogenetic inference was carried out using the following parameters: Starting tree, BIONJ; Type of tree improvement, SPR; Branch support, aLRT SH-like.

Results and discussion

Metagenomic sequences assembly and pattern searching analysis

The metagenome obtained from MG-RAST on line server showed a total of 9,921,136 high-quality Illumina sequences. The sequences were assembled into 33,307 contigs (> 500 bp) using the Newbler algorithm (Margulies et al. 2005) and into 42,800 (> 500 bp) contigs using the Meta-Hit algorithm (Nielsen et al. 2014). Contigs lengths ranged from 500 to 131,803 bp and from 500 to 210,782 bp with Newbler and Meta-Hit assemblers, respectively. Both, Newbler and Meta-Hit assemblies, were translated into the six reading frames, resulting in 199,842 AA sequences ranging from 166 to 43,934 AA for Newbler, and 256,800 AA sequences ranging from 166 to 70,260 AA, for Meta-Hit.

A regular expression (RegExp) was built to search for the active domain of proteases S8A known as the catalytic triad (CT) (Fig. 1a). The RegExp used to identify the three motifs of the CT was able to recover 12 translated contigs from the Newbler assembly and 14 translated contigs from the Meta-Hit assembly. These contigs ranged from 299 to 4005 AA, and from 228 to 26,835 AA, respectively. The contigs were trimmed to recover only the CT subsequences due to translated contigs may contain multiple proteins coding regions (or even full bacterial genomes). To further reduce

the possibility of chimeras, the recovered CT subsequences from the Newbler and Metahit assemblies were compared to each other, identifying 9 CT subsequences that were identical for both assemblies. These 9 shared sequences, with a size range from 196 to 238 AA, were used for downstream analyses (Supplementary Table 1).

Functional annotation of the putative novel proteases S8

To identify the nearest orthologous of the nine putative proteases S8A, a BLASTp search of their CT subsequences was performed against the NR database at NCBI and on the other hand against family S8A sequences from MEROPS database, using the local BLAST algorithm (Altschul et al. 1990; NCBI Resource Coordinators 2012; Rawlings et al. 2018). Results revealed that the putative proteases S8A had best hits to five unique phyla, with identities ranging from 62 to 89% (Table 1). Additionally, a local BLASTp was performed against the entire MEROPS database to test for false positives, and similar results were obtained (data not shown).

To analyze the diversity among all nine recovered putative proteases S8A, a multiple sequence alignment (MSA) was generated using the CT subsequences of all nine putative proteases S8A, the seven closest hits from the NR database, the eight closest hits from the MEROPS database, plus the subtilisin Carlsberg (Fig. 1b, Table 1). Interestingly, although the three motifs for all 9 CT subsequences are conserved, AA variations can be observed; for instance, the motifs “H” of MGR4536384.3_01 and MGR4536384.3_08 have the sequences HGTQVAL and HGNQVAL, respectively, while the motif most commonly found was HGTHVAG (Fig. 1b). For the 9 CT subsequences, VR1 and VR2 span 30 to 59 AA and 145 to 179 AA, respectively. Subsequences that have variable regions of the same size do not necessarily share the exact same sequence, e.g., MGR4536384.3_01 shares identical D, H, and S motifs with OYW49734.1 and MER0974284, separated by a 31 AA VR1 and a 179 AA VR2 (Fig. 1b). However, in pairwise alignments the identity observed was 76% and 75%, respectively (Table 1).

A second MSA was generated and a phylogenetic tree was built including the closest hits from the NR and MEROPS

Table 1 Nearest homologs to the metagenomic putative protease S8 sequences

Seq ID	QL (aa)	Nearest Homologue Protein	Organism	E value	Identity (%)	Accession no.
MGR4536384.3_01	231	Hypothetical protein	<i>Novosphingobium</i> sp. (phylum Proteobacteria)	3e-114	76	OYW49734.1
		Subfamily S8A	<i>Novosphingobium subterraneum</i>	1e-116	75	MER0974284
MGR4536384.3_02	230	Hypothetical protein	Planctomycetaceae (phylum Planctomycetes)	3e-110	78	PHY00594.1
		Subfamily S8A	<i>Blastopirellula marina</i>	7e-95	67	MER0062435
MGR4536384.3_03	221	peptidase S8	Bacteroidetes (phylum Bacteroidetes)	2e-77	63	PKP34958.1
		Subfamily S8A	<i>Novosphingobium subterraneum</i>	8e-88	67	MER0967697
MGR4536384.3_04	214	Hypothetical protein	<i>Acidimicrobium</i> sp. (phylum Actinobacteria)	2e-131	89	KRO47990.1
		Subfamily S8A	<i>Actinobacterium acidi</i>	3e-64	53	MER0973676
MGR4536384.3_05	196	Hypothetical protein	<i>Acidimicrobium</i> sp. (phylum Actinobacteria)	6e-99	87	KRO46506.1
		Subfamily S8A	<i>Micromonospora carbonacea</i>	5e-90	65	MER0971069
MGR4536384.3_06	238	Hypothetical protein	<i>Calothrix</i> sp. (phylum Cyanobacteria)	1e-67	51	WP_073617898.1
		Subfamily S8A	<i>Oscillatoriales cyanobacterium</i>	8e-70	67	MER0979041
MGR4536384.3_07	196	Hypothetical protein	<i>Acidimicrobium</i> sp. (phylum Actinobacteria)	4e-109	87	KRO47350.1
		Subfamily S8A	<i>Micromonospora</i> sp.	1e-89	65	MER0401707
MGR4536384.3_08	231	Hypothetical protein	<i>Novosphingobium</i> sp. (phylum Proteobacteria)	3e-110	73	OYW49734.1
		Subfamily S8A	<i>Novosphingobium subterraneum</i>	1e-111	73	MER0974284
MGR4536384.3_09	197	S8 family peptidase	<i>Actinopolymorpha cephalotaxi</i> (phylum Actinobacteria)	3e-72	62	WP_092886438.1
		Subfamily S8A	<i>Arthrobacter phenanthrenivorans</i>	5e-76	60	MER0240645

Putative proteases S8 were searched against NR and MEROPS databases using BLASTp

Seq ID Sequence identifier, QL query sequence length (amino acids), E value expected value, Accession No. accession number in the NCBI

database, the homologs sequences to *Bacillus licheniformis* subtilisin Carlsberg derived from several phyla, the subtilisin Carlsberg, and sequences from S8B family as controls. A Maximum Likelihood phylogenetic tree was inferred from the MSA. The tree with the highest Log-likelihood ($-18,253.09782$) has a total branch length of 29.11407 units, measured in the number of substitutions per site (Fig. 1c). The putative proteases S8A CT subsequence recovered by our method show a clear close relationship to the proteases reported in NR and MEROPS databases, and to some homologs sequences of the subtilisin Carlsberg (MGR453684.3_06, MGR453684.3_02). Moreover, the S8B family was grouped on the basal branch of the tree and all 9 putative proteases S8A CT were grouped into a clade exclusively formed by proteases from S8A family (Fig. 1c). The identities of the nine recovered CT ranged from 62 to 89% to known proteases that belong to Proteobacteria, Actinobacteria, Planctomycetes, Bacteroidetes, and Cyanobacteria, consistent with the most abundant phyla reported for this metagenome (Keegan et al. 2016). These results strongly suggest that the sequences obtained with our method are proteases S8A (Table 1, Fig. 1c). It is worth mentioning that the identity and phylogenetic relationships are inferred from the CT subsequence and the tree represents the divergence path of the gen S8A only, since these enzymes might be subject to different selective pressures as those acting on taxonomy marker genes like 16S.

We were able to recover 3 putative full CDS out of the nine proteases S8A. These were found in three of the longest contigs: MGR453684.3_01 (80,506 NT), MGR453684.3_02 (5877 NT), and MGR453684.3_03 (18,085 NT). The identities of their predicted full proteins to the closest reported proteins were 62%, 59%, and 56%, respectively. Out of the six remaining putative proteases S8, two had truncated C-terminal ends, one had a truncated N-terminal end, and three were truncated at both N- and C-terminal ends.

Thus, this evidence support the idea that these nine genes can be considered as novel proteases S8A, and that our method is robust and capable to detect proteases S8A genes from a wide range of bacteria.

Customizing and automatization of the method

The first stage for customizing our methodology to any protein of interest is the design of the RegExp. Our RegExp was built from an analysis of the CT of proteases S8A. To aid in the identification of the conserved and non-conserved AA, a MSA of thoroughly curated proteases S8 sequences (34 type S8A and 3 type S8B sequences) from the MEROPS database was analyzed using the Motif discovery tool of the MEME suite (Bailey et al. 2015) and the proposed RegExp was manually refined (Fig. 1a). Thus, our straightforward

method allows to easily customize it to perform searches of any enzyme of interest with highly conserved regions. Nevertheless, although the RegExp are easy to use these are not the only strategy to search for novel proteins. The Hidden Markov Models (HMM) area powerful method that have been widely used to detect conserved and semi-conserved motifs and domains. However, using an HMM strategy to search for motifs in enzymes with multiple apparently independent motifs, such as proteases S8A, would increase the amount of false positive, since proteins with incomplete or unordered sets of motifs could be recovered. Nevertheless, if the aim is a broader study of any enzyme, the method described in this study could be used as a basis to perform more robust HMM searches, starting from an MSA of the sequences recovered by a RegExp to build a HMM profile.

Additionally, the automatization of this method will allow to analyze larger metagenomic datasets. Our methodology is easily scalable; using “for in” loops the commands can be iterated across all files present in a folder effectively mining multiple metagenomes consecutively (a bash file for this purpose is included as Supplementary File 2). However, it is important to highlight that: (1) the “spirit” of our method is to be able to mine metagenomic data using computational resources available at a typical lab and web tools for downstream analysis. (2) Yet, it should be pointed out that it is possible to run MAFFT (Katoh and Standley 2013) and PhyML (Guindon et al. 2010) locally to obtain the MSA and phylogenetic tree of the putative proteases S8A derived from the pattern searched with RegExp. When the MSA analysis is embedded in an automated pipeline, the MSA will show only the diversity of the recovered putative proteases S8A. (3) Despite the command lines execute at an acceptable rate on a desktop computer (3 seg/50 Mb approx., Supplementary Table 2), the bottleneck will be the MSA and the phylogenetic tree due to the number of sequences increasing exponentially the processing time.

Finally, as users of next generation sequencing technologies (NGS), we have been witness to how these have revolutionized genomic sciences allowing the access to the genetic content of any organism. Consequently, the amount of genomic data has exponentially increased during the last decade (NCBI Resource Coordinators 2012). However, we are aware that there are challenges associated with the development of the NGS technologies. For one, hardware infrastructure and specialized tools are required to manage and analyze “Big Data”, and concomitantly, very specialized knowledge (bioinformatics) is required for these purposes. This has resulted in a bottleneck for many research groups, although important efforts have been made to overcome these limitations by providing access to different servers and online tools [e.g., Galaxy (Afgan et al. 2018), MG-RAST (Keegan et al. 2016), MEME suite (Bailey et al. 2015), etc.]. Unfortunately, managing a large amount of data is not an

easy or cheap task, thus many of these very useful online resources are file-size limited and custom analysis are not allowed. In the light of the above, in our own experience having access to straightforward and robust strategies such as those described in this study becomes essential to accelerate the data analysis derived from NGS technologies.

Conclusion

The bioinformatics methodologies presented in this study can be used as a suitable tool to easily mine the huge amount of publicly available metagenomic sequences, increasing dramatically the possibility of identifying novel genes such as protease S8A and other different enzymes of potential interest for industrial applications.

Acknowledgements The authors wish to express their gratitude to National Science and Technology Council, Mexico for providing the financial support for this research (Project No. INFR-2016-01-269833). The authors thank César de los Santos-Briones and Mildred R. Carrillo-Pech for their technical assistance.

Authors contribution All the authors contributed to this work. Góngora-Castillo and Ramirez-Prado designed and performed the experiments and analyzed the data; Caamal-Pech, Contreras-De la Rosa and Apolinar-Hernández participated in performing the experiments and the data analysis. López-Ochoa and Quiroz-Moreno participated in drafting the paper and discussing results. O'Connor-Sanchez, Ramirez-Prado and Góngora-Castillo conceived and designed the research and wrote the paper.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical statement Each of the authors confirms that this manuscript is original, has not been previously published and is not currently under consideration by any other journal. Additionally, all of the authors have approved the contents of this paper and have agreed to the 3 Biotech's submission policies. The manuscript has two corresponding authors, who are Dr. Jorge H Ramírez-Prado and Dr. Aileen O'Connor-Sánchez.

References

- Afgan E, Baker D, Batut B et al (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 46:W537–W544. <https://doi.org/10.1093/nar/gky379>
- Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Amann R, Ludwig W, Schleifer K (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 59:143–169
- Andrews S (2010) FastQC. In: Qual. Control Tool High Throughput Seq. Data. www.bioinformatics.babraham.ac.uk/projects/fastqc. Accessed 4 Oct 2018
- Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME Suite. *Nucleic Acids Res* 43:W39–W49. <https://doi.org/10.1093/nar/gkv416>
- Chen I-MA, Chu K, Palaniappan K et al (2019) IMG/M vol 5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res* 47:D666–D677. <https://doi.org/10.1093/nar/gky901>
- DeLange R, Smith E (1967) Subtilisin Carlsberg. Amino acid composition; isolation and composition of peptides from the tryptic hydrolysate. *J Biol Chem* 243:2134–2142
- Guindon S, Dufayard J-F, Lefort V et al (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59:307–321. <https://doi.org/10.1093/sysbio/syq010>
- Jisha VN, Smitha RB, Pradeep S et al (2013) Versatility of microbial proteases. *Adv Enzyme Res* 01:39–51. <https://doi.org/10.4236/aer.2013.13005>
- Katoh K, Standley DM (2013) MAFFT multiple alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30:772–780. <https://doi.org/10.1093/molbev/mst010>
- Keegan KP, Glass EM, Meyer F (2016) MG-RAST, a metagenomics service for analysis of microbial community structure and function. In: Martin F, Uroz S (eds) *Microbial environmental genomics (MEG)*. Springer, New York, pp 207–233
- Laskar M, James RE, Chatterjee A et al (2011) Modeling and structural analysis of evolutionarily diverse S8 family serine proteases. *Bioinformation* 7(5):239–245
- Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. <https://doi.org/10.1038/nature03959>
- Mitchell AL, Scheremetjew M, Denise H et al (2018) EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res* 46:D726–D735. <https://doi.org/10.1093/nar/gkx967>
- NCBI Resource Coordinators (2012) Database resources of the national center for biotechnology information. *Nucleic Acids Res* 41:D8–D20. <https://doi.org/10.1093/nar/gks1189>
- Nielsen HB, Almeida M, Juncker AS et al (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* 32:822
- Rawlings ND, Barrett AJ, Thomas PD et al (2018) The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res* 46:D624–D632. <https://doi.org/10.1093/nar/gkx1134>
- Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet* 16:276–277. [https://doi.org/10.1016/S0168-9525\(00\)00204-2](https://doi.org/10.1016/S0168-9525(00)00204-2)
- Shendure J, Balasubramanian S, Church GM et al (2017) DNA sequencing at 40: past, present and future. *Nature* 550:345–353. <https://doi.org/10.1038/nature24286>
- Thézé J, Li T, du Plessis L et al (2018) Genomic epidemiology reconstructs the introduction and spread of Zika virus in Central America and Mexico. *Cell Host Microbe* 23:855–864.e7. <https://doi.org/10.1016/j.chom.2018.04.017>