Representative Candidate of Gelatinase Encoded Gene in *Enterobacter aerogenes* (Strain EA1) for Hydrolyzing Porcine Gelatin

(Calon Wakil Gen yang Mengekod Gelatinase di dalam *Enterobacter aerogenes* (Strain EA1) bagi Hidrolisis Gelatin Babi)

Safiyyah Shahimi, Sahilah Abd. Mutalib*, Rozida Mohd. Khalid, Mohd Fadly Lamri, Maaruf Abdul Ghani, Mohd. Faizal Abu Bakar & Mohd. Noor Mat Isa

ABSTRACT

Enterobacter aerogenes *(Strain EA1) was characterised as gelatinase species-specific bacterium to porcine and fish gelatin. This bacterium offers the possibility of enzymes production to each species gelatin. In this study, a bioinformatic analysis toward genome sequence of* E. aerogenes *was perfomed to determine genes encoded for gelatinase.* E. aerogenes *were partially genome sequenced resulting in 5.0 mega basepair total size of sequence. From pre-process pipeline, 78.58% of high quality reads was obtained for* E. aerogenes. *Genome assembly produced 120 contigs with 55.17% of GC base content. Protein prediction analysis determined 4705 genes from* E. aerogenes. *Two gelatinase candidate target genes selected from* E. aerogenes *were NODE_9_length_26866_cov_148.013245_12 containing 1029 bp sequence and NODE_24_length_155103_cov_177.082458_62 containing 717 bp sequence with highest similarity identity percentage against gelatinase enzyme available in Swiss-Prot and NCBI online database. Then, primers were designed for each target genes based on the open reading frame of selected genes. The expected gene candidate showed, 1029 bp and 717 bp amplicons for gen NODE_9_12 and gene NODE_24_62, respectively. Thus, gelatinase candidate encoded genes of* E. aerogenes *were identified for hydrolyzing porcine gelatin.*

Keywords: Bioinformatic analysis; Enterobacter aerogenes; gelatinase; hydrolysis; porcine gelatin

ABSTRAK

Enterobacter aerogenes *(Strain EA1) ialah sejenis bakteria yang menghasilkan gelatinase khusus-spesies terhadap gelatin babi dan ikan. Bakteria ini menawarkan kemungkinan penghasilan enzim yang mencernakan gelatin spesies tertentu. Dalam kajian ini, satu analisis bioinformatik terhadap jujukan genom* E. aerogenes *telah dilakukan bagi mengenal pasti gen yang mengekod gelatinase. Penjujukan separa genom* E. Aerogenes *telah menghasilkan saiz keseluruhan sebanyak 5.0 mega pasangan bes. Daripada pemprosesan awal, 78.58% jujukan berkualiti tinggi telah diperoleh bagi* E. aerogenes. *Penghimpunan genom telah menghasilkan 120 kontig dengan 55.17% kandungan bes. Analisis peramalan protein telah mengenal pasti 4705 gene daripada* E. aerogenes. *Dua gen sasaran calon gelatinase terpilih ialah NODE_9_length_26866_cov_148.013245_12 mengandungi 1029 bp jujukan dan NODE_24_length_155103_cov_177.082458_62 mengandungi 717 bp jujukan, dengan peratus identiti persamaan tertinggi dengan enzim gelatinase yang terdapat di pangkalan data atas talian Swiss-Prot dan NCBI. Kemudian, pencetus telah direka bentuk bagi setiap gen sasaran berdasarkan rangka bacaan terbuka gen terpilih. Calon gen yang dijangka telah ditunjukkan, 1029 bp dan 717 bp amplikon bagi gen NODE_9_12 dan NODE_24_62. Oleh itu, gen yang mengekod calon gelatinase telah dikenal pasti bagi hidrolisis gelatin babi.*

Kata kunci: Analisis bioinformatik; Enterobacter aerogenes; gelatin babi; gelatinase; hidrolisis

INTRODUCTION

Gelatinase bacteria are prominent due to its ability to hydrolyze gelatin which may affect the quality and safety of gelatin production. These bacteria would affect the viscosity and gelling capacity of gelatin-based products (De Clerck et al. 2004). In addition, contamination of these bacteria in foods and pharmaceutical applications would affect human health as it possesses pathogenic properties. Indeed, gelatinase is a proteolytic enzyme with ability to hydrolyze gelatin into smaller polypeptides, peptides or amino acids. There are certain bacteria that can produce gelatinase including *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Clostridium perfringens* and *Serratia marcescens* with currently no gelatin substrate specificity studied.

Gelatin is a proteinaceous colloid of animal origin which has been used widely in foods, beverage, pharmaceutical (Suhimi & Mohammad 2012) and cosmetic industries (Ismail et al. 2018). Different types of gelatin have different molecular weight distribution which will determine their viscosity, dispersion in water, adhesiveness and gel strength (Gudmundsson & Hafsteinsson 1997). A

bacterial strain was identified as gelatinase producer known as *Enterobacter aerogenes* (Strain EA1) using 16sRNA universal primers and Biolog (Mohd Fadly 2016). This *E. aerogenes* (Strain EA1) was isolated from hot spring water, Melaka, Malaysia and found to produce the highest enzyme activity towards medium containing porcine and fish gelatin (pH7) and slight activity towards bovine gelatin (pH7) when it was grown at 37°C for 24 h. Therefore, there was potential to identify the genes coded for specific gelatinase production from this Gram-negative rod-shaped microorganism of the *Enterobacteriaceae* family.

In order to identify the gelatinase genes responsible for degradation of specific type of gelatin, a preliminary study based on bioinformatics analysis was done to the available raw genomic data. The gelatinase species-specific bacterium to porcine, bovine and fish for instance, offers the possibility of enzymes production in which specific-specific enzymes produce for those species' gelatin. Thus, detection of the specific gelatin using bacterial gelatinase can be further analyzed especially for hydrolyzing species-specific gelatin towards porcine, bovine and fish gelatin to be showed, as gelatin adulteration may occurred in certain products (Shahimi et al. 2018). The data may be useful in other application such as diagnostic using enzymatic approaches for the detection of gelatin species.

MATERIALS AND METHODS

BACTERIA

Bacteria of *E. aerogenes* (Strain EA1) was isolated from hot spring water, Melaka, Malaysia in Jun 2015. The raw data size obtained for the bacterium is 5,009,814 bp using Next Generation Sequencing (NGS) done by Malaysian Genome Institute (MGI), Malaysia. Bioinformatic analysis was done to the raw data before PCR assay of the synthesized cDNA.

BIOINFORMATICS ANALYSIS

Bioinformatics analysis of this bacterium was done using LINUX server system, PuTTY and WinSCP in Malaysian Genome Institute (MGI), consists of raw data pre-processing, genome assembly, gene prediction, gene annotation, gene selection, protein translation and primer design.

In pre-processing, raw data were uncompressed from gunzip file into fastq file, followed by calculation of total reads, trim reads based on quality (Qphred=20), trim reads based on sequence length (minimum 50 bp), count reads for every trimmed fastq file to get info on total high quality reads (QV≥20, Length ≥50bp), screening for phiX reads or other contaminants and count reads for cleaned fastq files for total clean reads. Finally, fastq paired sequence and singletons was selected, shuffled and combined (Mohd Noor 2014).

Velvet Optimiser software was used for data assembly using VelvetOptimiser. Pl script comprise of Velvet h and Velvet g. This assembly produced contigs with N50 which is a length of contigs which cover at least 50% of total genome length and optimum k-mer.

Gene prediction using prodigal script will identify the region of genomic DNA that encoded gene. The output showed protein genes from the whole genome sequence. Gene annotation was studied using Blast Local Alignment Search Tool (BLAST) to find similarities between predicted protein genes and gelatinase protein data available in Swiss-Prot and NCBI online database. Blastp script was used for protein query.

Finally, two gelatinase candidate target genes were selected from the bacterium, based on their similarity percentage, highest hit score and lowest e-value result from gene annotation. Artemis software was used to get back the sequence of each candidate. DNA sequences of each candidate were translated into amino acid sequence using online Expasy tool (expasy.org) and their open reading frame (ORF) were determined. A homologous-protein search of the gelatinase candidate target genes with an existing gelatinase in GenBank was done using Clustal Omega Multiple Sequence Alignment to get the value of identity and similarity between the proteins.

Foward and reverse primers were designed using IDT oligoanalyzer to amplify gelatinase candidate target genes from the bacteria.

cDNA SYNTHESIS

The bacterial culture was revived from 20% (v/v) glycerol stock by inoculating the cultures into nutrient broth which has been added with 10% (w/v) Brain Heart Infusion (BHI). The culture was grown at 37°C, orbital shaking at 80 rpm for 24 h before transferred into new medium (nutrient broth added with 3% (w/v) of porcine gelatin) and then incubated at 37°C, orbital shaking at 80 rpm for 48 h.

RNA extraction was done to 1 mL of well grown bacteria using FavorPrep Blood/Cultured cell Total RNA purification Mini Kit (Favorgen, TW) as described by the manufacturer's instruction. RNA extract concentration and purity were determined at 260 nm and 280 nm using MaestroNano Spectrophotometer (MaestroGen, USA).

Extracted RNA was then converted to complementary DNA (cDNA) using qPCRBIO cDNA Synthesis kit (PCRBIOSYSTEMS, MY) following procedure provided according to manufacturer's protocol.

PCR AMPLIFICATION

PCR amplification was performed in a 25 μL reaction volume containing a mixture of PCRBIO HiFi Polymerase (PCRBIOSYSTEMS, MY), forward and reverse oligonucleotide primers (1st Base, MY) (Table 1), nuclease free water (NFW) and the cDNA template as described by the manufacturer's instruction. All mixture was prepared in a 0.2 mL sterile PCR tubes. Negative and positive controls were included in every cycle. Polymerase chain reaction for each candidate target gene was carried out in Applied Biosystems Veriti Thermal Cycler (Thermo Fisher Scientific, MY) with a

TABLE 1. Forward and reverse primers designed from gelatinase candidate gene of *Enterobacter aerogenes* (Strain EA1)

Primer	Description	Length (bp)	GC%	Tm
Primer pair 1 NODE_9_length_26866_ cov_148.013245_12	Forward primer : 5'- CAC CAT GCC AGA TAC TCG TCC TAG -3'	24	54.2	58.6
	Reverse primer : 5'- TCA TGA TGT GAC TCC AAC ATC CTC C -3'	25	48.0	58.4
Primer pair 2 NODE_24_length_155103_ cov_177.082458_62	Forward primer : 5'- CAC CAT GAT AAC ACT CGA AAT GCG CAA -3'	27	44.4	59.8
	Reverse primer : 5'- TTA AAG GTG CTG TAG AAA AAA GGC GG -3'	26	42.3	57.9

temperature program consisting of initial denaturation at 95°C for 1 min to completely denature the template, followed by 35 cycles of denaturation at 95°C for 15 s, annealing for 15 s (at 51.4°C for the first gelatinase candidate target gene and at 48.9°C for the second gelatinase candidate target gene), polymerization at 72°C for 45 s and final elongation at 72°C for 7 min.

Gel electrophoresis was done using 1% (w/v) agarose gel in 1X TAE buffer (40 mM Tris-acetate, 1 mM EDTA, pH8.0) (1st Base, MY) at 100V for 45 min to view the PCR products. The gel was pre-stained with Maestrosafe™ Nucleic Acid Prestain (Maestrogen, TW) and viewed under gel imager UV trans-illuminator Gel Documentation System (Syngene, UK). GeneRuler™ 1 kb DNA ladder (Thermo scientific, MY) was used as marker.

RESULTS AND DISCUSSION

Full genome sequence analysis is important to redefine bacterial species in this modern era (Stakebrandt et al. 2002). Next generation sequencing (NGS) has got attention in biological research such as gene expression quantification. In this study, total length of genome DNA sequence produced from NGS sequencing process of *E. aerogenes* (Strain EA1) was 5,009,814 base pair (bp), which about the size of circular genome of *E. aerogenes* clinical isolate (5,419,609 bp) sequenced by Diene et al. (2012) and *E. aerogenes* KCTC 2190 (5,280,350 bp) sequenced by Shin et al. (2012). The first complete genome sequencing of *E. aerogenes* species was done on strain KCTC 2190 by 454 GS FLX Titanium pyrosequencing (Roche) (Shin et al. 2012). Output data generated as raw reads from sequencing process have to be processed prior to analysis (Del Fabbro et al. 2013).

The pre-processing gave result on total reads after count reads in every fastq files, total high quality reads after count reads for every trimmed fastq file, total clean reads after screen for phiX reads and count reads for cleaned fastq file, total of paired reads and singleton reads of the bacterial genome for next analysis (Table 2). From the total reads of 87,569,556 bp found in raw data of *E. aerogenes* (Strain EA1), only 68,814,714 bp were classified as total clean reads with 78.58% of high quality reads obtained, which was in the acceptable range (≥70%) to undergo

TABLE 2. Result from pre-process pipeline

Description	Results
Total reads	87,569,556 bp
Total high quality reads	68,814,718 bp
Total clean reads	68,814,714 bp
Total paired reads	61,063,302 bp
Total singleton reads	7,751,412 bp
Percentage of high quality	78.58%

genome assembly. Determination of high quality reads is very crucial as discounting the existence of low quality base calls may add unreliable random sequences to the dataset and affecting any downstream analysis such as generation of false k-mer in genome assembly (Del Fabbro et al. 2013).

E. aerogenes (Strain EA1) were then assembled with N50 contigs size of 124,856 and the most optimum k-mer of 75. De-novo genome assembly used is a process to assemble high quality DNA sequences using VelvetOptimiser.Pl script. Genome assembly analysis will produce N50 contigs and optimum k-mer will be determined. N50 is a length of contigs which cover at least 50% of total genome length. It will assemble genome sequence contigs from the longest to the smallest in length and select 50% base pair (bp) from the whole genome. Less number of contigs with high N50 will produce good sequence quality assemble. Therefore, *E. aerogenes* (Strain EA1) can be concluded as well assembled. However, N50 cannot be used as the only factor in determining good assembly as there are other factors to be considered (Desai et al. 2013; Mohd Faizal 2013). An optimum k-mer should also be determined to get well assembly. K-mer reads between 50 bp and 100 bp total length reads of contigs. High k-mer will produce low quality assemble product while, there will be high risk to misassemble if the k-mer is low because of the short length contigs. Assembly process will also give estimation on genomic size, number of contigs, and GC base content. In this study, *E. aerogenes* (Strain EA1) which has been isolated previously from hot spring water showed high GC base content (55.17%). A study by Musto et al. (2004) showed positive correlation between GC base content and optimum growth temperature factor in

prokaryote, where organism which prefers higher growth temperature usually will have a higher GC base content. Thus, the statement was strongly supported by this study. Table 3 shows results from the assembly process.

TABLE 3. Result of assembly analysis

Description	Results
Number of contigs	120
Total size of contigs	5008879
Longest contig	375400
Shortest contig	149
Number of contigs > 1K nt	81 (67.5%)
Mean contig size	41741
N50 contig length	124856
% GC content	55.17 %

Using Prodigal software, there were 4,705 protein genes predicted from the assembled data of *E. aerogenes* (Strain EA1). Prodigal is a gene-finding program for microbial genomes (Hyatt et al. 2010). In this study, gene prediction analysis using Prodigal software with Prodigal script showed protein sequences from the genomic data. By comparing its genomic size with number of predicted genes, it showed that number of predicted genes was directly proportional to the size of genome. A bigger genome usually will contain more number of genes, as supported by previous sequenced *E. aerogenes* KCTC 2190 which have a larger genomic size (5,280,350 bp) with predicted genes of 4912 (Shin et al. 2012) and *E. aerogenes* clinical isolate which contained 5,419,609 bp circular genome with 5260 predicted coding sequences (Diene et al. 2012). Therefore, 4,705 predicted genes for *E. aerogenes* in this study was fitting its genomic size of 5,009,814 bp.

Gene annotation process involves two important analyses which are identification of protein homologues and domain analysis (Mohd Faizal 2013), which covering the coding genes, pseudogenes, promoters-regulatory regions, untranslated regions and repeats (Reeves et al. 2009). The basic level of annotation was done using BLAST (Basic Local Alignment Search Tool) to find similarities and annotating genomes. BLAST is an algorithm comparing biological sequence information, such as the amino acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. As the target protein in this study was gelatinase enzyme, protein sequence determined from the gene prediction was set as query in blastp against all gelatinase protein sequence located in UniProtKB-Swiss-Prot (http:www.uniprot.org) and NCBI-non redundant (nr) database.

There were 273 hit determined for *E. aerogenes* (Strain EA1) after BLASTP with all gelatinase in the gene banks (NCBI and UniProtKB-Swiss-Prot). The results obtained were based on selected e-value of ≤1e-5 as this value was chosen in most BLASTP analysis and showed significant similarity in homology. To get a more relevant hit as to infer the given gene is present, the lowest percentage of identity chosen was at least 30%. This threshold value has been generally used to indicate a thriving homology modeling (Mohd Faizal 2013; Xiang 2006). 244 putative proteins are coded by the same genes which was NODE_9_12 with identity percentage of 30% to 48%. The other remaining hit proteins have lower percentage of identity below 30%. Another BLAST of NODE_9_12 with NCBI database showed 100% similarity hit with protease prtS of *Enterobacter aerogenes* (GenBank: KGB04922.1). When BLAST was done using non-redundant protein sequence of NCBI as database, seprase from *Erwinia pyrifoliae* hitted NODE_24_62 with the highest similarity of 55%.

Finally, gene selection was done by considering the gene length factor where the longer candidate gene will be preferred to avoid gene data lost (Mohd Faizal 2013). ARTEMIS software as a sequence viewer (Rutherford et al. 2000), was used to retrieve the DNA sequence of the selected gelatinase candidate target genes. Although not all gelatinase candidate target genes were coded as gelatinase, other gene had been selected due to gelatinase activity reported previously from the enzyme (O'Brien & O'Connor 2008). Domain analysis from Conserved Domain Database (CDD) showed that the first gelatinase candidate target gene of *E. aerogenes* (Strain EA1) possessed Zn-dependent metalloprotease domain while the second candidate gene has dienelactone hydrolase as domain. Domain analysis was done to all putative proteins to get more information on the functions, as it is also important in gene nomenclature to be well characterized (Gaudermann et al. 2006; Mohd Faizal 2013). All putative proteins have to be further evaluated to confirm the gelatinase activity. Table 4 shows two gelatinase candidate target genes selected due to highest identity percentage and lowest e-value to be targeted from the bacterium.

Amino acid sequences were translated using Expasy tool to assure the protein functioning. Multiple sequence alignment between gelatinase candidate genes in this study with gelatinase gene (gelE; GenBank: M 37185) from previous study of *Enterococcus faecalis* by Su et al. (1991) and gene 2 (P2) of *Lysinibacillus spahericus* (Rul Aisyah et al. 2016) using Clustal Omega (Clustal O), showed some amino acid sequence homology which indicated the gelatinase gene conserved properties. Figure 1 shows the overall similarities between these enzymes. The first gelatinase candidate target gene of *E. aerogenes* (Strain EA1); NODE_9_length_26866_cov_148.013245_12 contained 1029 base pairs with 342 amino acids showed 26.58 percent of identity with *gelE* and 33.33 percent of identity with gene P2 of *L. spahericus*. On the other hand, the second gelatinase candidate target gene; NODE_24_

TABLE 4. Two gelatinase candidate target genes from BLASTP against gelatinase in NCBI database

Query_Name	%_Identity	E-Value	Acc_Num	Subject
NODE_9_ length_26866_ cov_148.013245_12	30.2	4.00E-26	tr|X6T8U3|X6T8U3_ENTFL	Gelatinase OS=Enterococcus faecalis B318 GN=P011_01203 PE=4 SV=1
NODE_24_ length_155103_ cov_177.082458_62	55.0	1.00E-93	gi|283476997|emb|CAY72889.1|	Seprase [Erwinia pyrifoliae DSM 12163]

(a)

```
NODE_24_62      ------------------------------------------------------------    0
gel_E_su        -------MKGNKILYILGTGIFVG-SSCL-------------------------------    21
NODE_9_12       ------------------------------------------------------------    0
lysini_node_23  MSSFFEWNCQKKLSQILGKGINLKNKNCIKVSKIFMGMGNVIVKKQVISSALALTVIAGG    60


NODE_24_62      ------------------------------------------------------------    0
gel_E_su        ---FSSLFVAAEEQVYS-ESEV-------------------STVLSKLE----KEAISEA    54
NODE_9_12       ------------------------------------------------------------    0
lysini_node_23  FGTFGATTTKAEEQKIQYHQEFKTPAYIGEEWKAPEGLDKKETVFQYLESKKDMFKLAGN   120


NODE_24_62      ------------------------------------------------------------    0
gel_E_su        AAEQYTVVD-RKEDAWGMKHLKLEKQTEGVTVDSDNVIIHLDKNGAVTSVTGNPVDQVV-   112
NODE_9_12       ------------------------------------------------------------    0
lysini_node_23  MDKHFNVVGEEKDAESGTTHVKLVEKHNNIPVYGSDQTVTLDKENNVKAFFGQVIPNLED   180


NODE_24_62      ------------------------------------------------------------    0
gel_E_su        -KIQSVDAIGEEGVKKIVASD-----------NPETKDLVFLAIDKRVNNEGQLFYK-VR   159
NODE_9_12       ------------------------------------------------------------    0
lysini_node_23  KNIPAFAKISAEQAETIAKADIEKEVGKVKNYDGVKKDLYVY------EKDGNYYLAYLV   234


NODE_24_62      ---------------------------------------------------------MITL    4
gel_E_su        VTSSPTGDPVSLVYKVNATDGTIMEKQDLTEHVGS-E---VTLKNSFQV-------TFNV   208
NODE_9_12       ---MPDTRP----------SFVIPPYIL-RRIVDHG-----SAQQQRCAQQTLSHVQTL    40
lysini_node_23  KASISKPAPGYWHYFVDATNGNVIEKYNAVDNITGFGYGVLGGKQSFEIAQDTKTGEFNL   294
                                                                           .:


NODE_24_62      EMRNLAGGEVIHACPAGMAEKPL-P----------------------VIVFYHGFTSSK    40
gel_E_su        PVEKSNTGIALHGTD----NTGVYHAVVDGKNNYSIIQA---PSLATLNQNAIDAYTHGK   261
NODE_9_12       MAHVPGRPAAAHVATPGQLERDIYDA-R-QTQELPGIQVRFEGQPSNGDVAVDEAYDYLG    98
lysini_node_23  FDGKRG--QGIHTFDAENMDENWFNLFS-QILGYTGEEVKSKSKFFE-DKAAVDAHVNAG   350
                             *    :                            ...


NODE_24_62      LVYSYFAVA-----LAEAGFRVIMPDAAEHGARFRGDEEGRMQRFWPILRQNF-------    88
gel_E_su        FVKTYYEDHFQRHSIDDRGMPILSVVDEQHPDAY-------DNAFWDGKAMRYGETSTPT   314
NODE_9_12       ITHDFFWKSYHRDSLDNRGLKLTGSVHY--GHEY-------QNAFWNGQQMVFGDGDGEI   149
lysini_node_23  KVYDYYKKTFNRNSFDDKGAKLISTVHV--GESW-------NNAAWNGVQMMYGDGDGKT   401
                 .  ::         : :  *  :             :          :  *      :


NODE_24_62      -LEFPALRDAIIAEGWLEDERLAVAGASMGGMTALGIMTHHPELKCVACLMGSGYFST--   145
gel_E_su        GKTYASSLDV-------------VG----HEMTHGVTEHTAGLEY---LGQSGALNESY   353
NODE_9_12       FNRFTIAIDV-------------VA----HELSHGVTESEAGLIY---FEQSGALNESL   188
lysini_node_23  FIPLSAGLDV-------------IG----HELTHAVTEHTANLVY---KNEPGALNESL   440
                 *.             .          : .:       *       * :.


NODE_24_62      --L-AQTLFPSPGFDTASLAEWDVTHQLAALA----------------DKPLLLWHGDAD   186
gel_E_su        SDLMGYIISGASNPE--------IGADTQSV-DRKTGIRNLQTPSK---------HGQPE   395
NODE_9_12       SDVFGSLVKQYQRQQTADKADWIIGEGLLAKGIHGQGLRSMSHPGTAYDDPILGKDPQPA   248
lysini_node_23  SDIMGVMVE---------KKSWDLGADIYTPGKPGDALRSLKDPASI-PNPLKPGEGYPD   490
                :  . :         :       :     .


NODE_24_62      DVVPPGDTFRLQQALTQSGLATNLTCHWQKGVRHRITPEALDATT---------------   231
gel_E_su        TMAQYDDRAR----YKGTPYYDQGGVHYNSGIINRIGYTIIQ----------NLGIEKAQ   441
NODE_9_12       HM-------K----DFVHTQQDNGGVHINSGIPNHAFYLAAIALGGFA-------WEKAG   290
lysini_node_23  HY-------N---KRYTGTADNGGVHINSSINNKAAYLVSEGGEHYGVKVTGVGREATE   539
                       .             :   * :..:  ::


NODE_24_62      -AFFLQHL----------------------------------------------------   238
gel_E_su        TIFYSS-LVNYLTPKAQFSDARDAMLAAAKVQYGDEAA--SVVSAAFNSAGIGAKEDIQV   498
NODE_9_12       YAWYDAVCDKSLAKDADFPAFAKLTIHHGEKRFGKET--AEAIRKAWEDVGVTS------   342
lysini_node_23  KIYYRA-LTKYLTANSDFKMMRQAALQSAEDLYGKNSKAVQAVTKAYDAVGVK-------   591
                  ::


NODE_24_62      -----------    238
gel_E_su        NQPSESVLVNE    509
NODE_9_12       -----------    342
lysini_node_23  -----------    591
```

(b)

```
NODE_24_62    MITLEMRNLAGGEVIHACPAGMAEKPLPVIVFYHGFTSSKLVYSYFAVALAEAGFRVIMP 60
CAY72889.1    MIELNTERFAEIECLHAFPSGKRHQALPTVLFYHGYSSSKEVYAYFAVALAQAGYRAVLP 60
              ** *: ..:*  *  :** *:*  .: **.::****::*** **:*******:*:*..::*

NODE_24_62    DAAEHGARFRGDEEGRMQRFWPILRQNFLEFPALRDAIIAEGWLEDERLAVAGASMGGMT 120
CAY72889.1    DADMHGARYDGDDQRRLTRFWDILRTNIDELPQIERALRQHQLVDGARLAVAGASMGGMT 120
              ** ****: **:: *: *** *** *: *:* :. *: . ::. *************

NODE_24_62    ALGIMTHHPELKCVACLMGSGYFSTLAQTLFPSPGFD----TASLAEWDVTHQLAALADK 176
CAY72889.1    ALGALARYPQLHSCACLMGSGYYRQLASTLFPPLVADRAEQEKQLAEYDVSHQLARFANR 180
              *** ::::*:*:. ********: **.****    *    .***:**:**** :*:::

NODE_24_62    PLLLWHGDADDVVPPGDTFRLQQALTQSGLATNLTCHWQKGVRHRITPEALDATTAFFLQ 236
CAY72889.1    PLLIWHGDADEVVPVAESVRLEQALRRSGLDRNLTYLIEKGVGHRITPPALSALTGFFAH 240
              ***:******:*** .::.**:*** :*** ***  :*** ***** **.* *.** :

NODE_24_62    HL----------      238
CAY72889.1    HLPLSAGDGKATR     253
              **
```

FIGURE 1. Sequence alignment, (a) Multiple sequence alignment between both gelatinase candidate target genes of *Enterobacter aerogenes* (Strain EA1), *gelE* of *Enterobacter faecalis,* and gene P2 of *Lysinibacillus sphaericus*. (b) Sequence alignment between between seprase of *Erwinia pyrifoliae* and NODE_24_length_155103_cov_177.082458_62. The highlighted region indicates active sites of the genes. An "*" (asterisk) indicates positions which have a single, fully conserved residue, a "." (period) indicates conservation between groups of weakly similar properties, while a ":" (colon) indicates conservation between groups of strongly similar properties

length_155103_cov_177.082458_62 contained 717 bp with 238 amino acids showed only 17.96 percent of identity with *gelE* of *E. faecalis* and 16.83 percent of identity with gene P2 of *L. spahericus*. Amino acid alignment was proven to show conserved sequences among species including the enzyme's catalytic site as in previous study by Selina Oh et al. (2011). In this study, the highly conserved regions were the positions of the enzymatic active sites. The active sites NA-V-HE-H-Y-E-LR-H found in NODE_9_length_26866_cov_148.013245_12 were similar to the *gelE* of *E. faecalis* (NA-V-HE-H-Y-E-IR-H). Both genes were found to share the same multidomain of Zn-metalloendopeptidase which usually contain three residues HEXXH that bind to zinc atom, where the Glu will act as a nucleophile during catalysis. As the zinc atom been activated by water molecules, a nucleophilic attack will eventually break down the peptide bonds in gelatin (Klimpel et al. 1994).

On the other hand, only some sequence homology of the *gelE* active site was represented in NODE_24_length_155103_cov_177.082458_62, as the known difference in domain classification and active site for seprase was G-X-S-X-G (Figure 1(a)). In this study, the sequence is GASMG where the catalytic domain consists of catalytic serine flanked by glycines to give a catalytic triad. The hydroxyl group of the serine will act as a nucleophile to attack the peptide bond while, the aspartic acid will stabilize the charged tetrahedral intermediates formed in the reaction (O'Brien & O'Connor 2008). A sequence alignment between seprase of *Erwinia pyrifoliae* and NODE_24_length_155103_cov_177.082458_62 showed 55.46 percent of identity with the same sequence of active sites (Figure 1(b)).

Table 1 shows forward and reverse primers designed from the gelatinase candidate target gene with their PCR cycles used (Table 2), respectively. Primers were designed to amplify the gelatinase candidate genes determined from the gene annotation process. Primers were used to target only the selected genes from the whole genome of both bacteria. Primers were designed manually by referring to the target gene as template. The primer sequences were then analyzed using IDT Oligoanalyzer online software to get information on crucial factors in primer design such as the length, GC% content, melting temperature and possibility to form dimer or loops from each primer sequence as these factors may affect the result of PCR amplification. A good primer usually is in the range of 19 bp to 30 bp length, 40% to 60% of GC content and has melting temperature below 60ºC. Therefore, all possible factors should be taken into consideration before polymerase chain reaction (PCR) to ensure the primers can target the amplification site on the gelatinase candidate target genes. Multiple bands observed were due to unspecific binding of the primers on the template which may be reduced by optimization of the PCR condition (Lorenz 2012).

RNA extraction done resulting 29.85 ng/μL of RNA extract was successfully transcribed into cDNA with concentration of 146.28 ng/μL. All gelatinase candidate target genes from the bacterium were successfully amplified from their cDNA, with presence of unspecific bands by estimating the sizes of expected amplicon band on agarose gel after polymerase chain reaction (PCR) assay and gel electrophoresis (Figure 2). The gene candidate of 1029 bp and 717 bp amplicons for Candidate Gene NODE_9_12 and Gene NODE_24_62, respectively, were observed. Hence, previous genome analysis of *E. aerogenes* (Strain EA1) had accurately showed the gelatinase candidate target genes inside the bacterium. The possibility of the gelatinase candidates to be expressed as porcine gelatinase thus, can be such an advantage in future halal gelatin analysis.
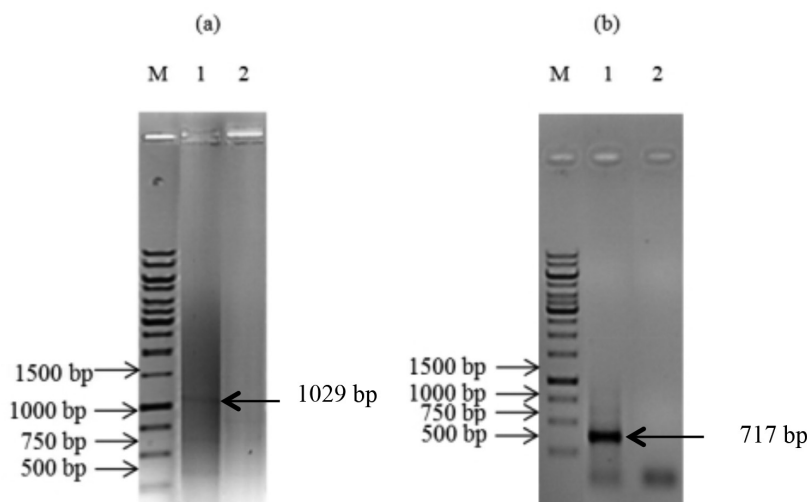
FIGURE 2. Amplicons of gelatinase candidate target genes from *Enterobacter aerogenes* (Strain EA1) (a) PCR products of the first gelatinase candidate target gene. M: 1kb DNA ladder; Lane 1: Candidate gene NODE_9_12; Lane 2: Negative control. (b) PCR products of the second gelatinase candidate target gene. M: 1kb DNA ladder; Lane 1: Candidate gene NODE_24_62; Lane 2: Negative control

CONCLUSION

In the present study, genome analysis of *E. aerogenes* (Strain EA1) resulting two genes encode gelatinase candidate were identified using bioinformatics analysis. PCR amplification using the primers designed based on the target sequence, was capable to amplify the target genes. The data may be useful for enzymatic analysis in detecting porcine gelatin species.

REFERENCES

De Clerck, E., Gevers, D., De Ridder, K. & De Vos, P. 2004. Screening of bacterial contamination during gelatin production by means of denaturing gradient gel electrophoresis, focused on *Bacillus* and related endospore-forming genera. *J. App. Microbiol.* 96: 1333-1341.

Del Fabbro, C., Scalabrin, S., Morgante, M. & Giorgi, F.M. 2013. An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS ONE* 8(12): e85024.

Desai, A., Marwah, V.S., Yadav, A., Jha, V., Dhaygude, K. & Bangar, U. 2013. Identification of optimum sequencing depth especially for De Novo genome assembly of small genomes using next generation sequencing data. *PLoS ONE* 8(4): e60204.

Diene, S.M., Merhej, V., Henry, M., Filali, A.E., Roux, V., Robert, C., Azza, S., Gavory, F., Barbe, V., Scola, B.L., Roult, D. & Rolain, J. 2012. The rhizome of the multidrug-resistant *Enterobacter aerogenes* genome reveals how new 'killer bugs' are created because of a sympatric lifestyle. *Mol. Biol. Evol.* 30(2): 369-383.

Gaudermann, P., Vogl, I., Zientz, E., Silva, F.J., Moya, A., Gross, R. & Dandekar, T. 2006. Analysis of and function predictions for previously conserved hypothetical or putative proteins in *Blochmannia floridanus*. *BMC Microbiology* 6: 1.

Gudmundsson, M. & Hafsteinsson, H. 1997. Gelatin from cods skin as affected by chemical treatments. *J. Food Sci.* 62: 37-39.

Hyatt, D., Chen, G-L., LoCascio, P.F., Land, M.L., Larimer, F.W. & Hauser, L.J. 2010. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinfor.* 11: 119.

Ismail, N.H., Yusop, S.M., Babji, A.S. & Maskat, M.Y. 2018. Kesan penambahan *Kappaphycus alvarezii*, gelatin ikan dan gelatin kaki ayam terhadap ciri-ciri kualiti sosej ayam. *Sains Malaysiana* 47(6): 1157-1165.

Klimpel, K.R., Arora, N. & Leppla, S.H. 1994. Anthrax toxin lethal factor contains a zinc metalloprotease consensus sequence which is required for lethal toxin activity. *Molecular Microbiology* 13(6): 1093-1100.

Lorenz, T. 2012. Polymerase chain reaction: Basic protocol plus troubleshooting and optimization strategies. *J. Vis. Exp.* 63: e3998.

Meyer, R. & Candrian, U. 1996. PCR-based DNA analysis for the identification and characterization of food components. *Lebensmittel-Wissenschaft undTech.* 29(1-2): 1-9.

Mohd Fadly, L. 2016. Pemencilan, pengenalpastian dan aktiviti enzim bagi bakteria gelatinase terhadap pengesanan gelatin bagi tujuan pengesahan halal. Ph.D Tesis, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia (Unpublished).

Mohd Faizal, A.B. 2013. Penentuan saiz dan analisis jujukan genom *Glaciozyma antartica*. MSc. Tesis, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia (Unpublished).

Mohd Noor, M.I. 2014. Next-generation sequencing data analysis workshop. Bioinformatics Training Laboratory, September 2014. *Malaysia Genome Institute*. pp. 17-19.

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valín, F. & Bernardi, G. 2004. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett*. 573(1-3): 73-77.

O'Brien, P. & O'Connor, B.F. 2008. Seprase: An overview of an important matrix serine protease. *Biochim. Biophys. Acta*. 9: 1130-1145.

Reeves, G.A., Talavera, D. & Thornton, J.M. 2009. Genome and proteome annotation: Organization, interpretation and integration. *Journal of the Royal Society Interface* 6(31): 129-147.

Rul Aisyah, M.R., Sahilah, A.M., Safiyyah, S., Rozida, M.K., Mohd. Khan, A., Mohd. Faizal, A.B. & Mohd. Noor, M.I. 2016. Bioinformatics analysis and detection of gelatinase encoded gene in *Lysinibacillussphaericus*. *American Institute of Physics* 1784: 030044.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. & Barrell, B. 2000. Artemis: Sequence visualization and annotation. *Bioinformatics* 16(10): 944-945.

Selina Oh, S.L., Leong, J.M., Abdul Munir, A.M., Nor Muhammad, M. & Farah Diba, A.B. 2011. Cloning and analysis of pyrG gene encoding orotidine 5-monophosphate decarboxylase of *Aspergillus oryzae* Strain S1. *Sains Malaysiana* 40(4): 331-337.

Shahimi, S., Abd Mutalib, S., Nazri, W.S.W., Abdullah, A. & Sani, N.A. 2018. Comparison of DNA profiling between fishes and pork meat using polymerase chain reaction-restriction fragment length polymorphisms (PCR-RFLP) analysis. *Sains Malaysiana* 47(7): 1535-1540.

Shin, S.H., Kim, S., Kim, J.Y., Lee, S., Um, Y., Oh, M-K., Kim, Y-R., Lee, J. & Yang, K-S. 2012. Complete genome sequence of *Enterobacter aerogenes* KCTC 2190. *J. Bacteriol*. 194: 2373-2374.

Stakebrandt, E., Frederiksen, W. & Garrity, G.M. 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int. J. Syst. Evol. Microbio*. 52: 1043-1047.

Su, Y.A., Sulavik, M.C., He, P., Makinen, K.K., Makinen, P.L., Fiedler, S., Wirth, R. & Clewell, D.B. 1991. Nucleotide sequence of the gelatinase gene (gelE) from *Enterococcus faecalis* subsp. Liquefaciens. *Infect. Immun*. 59(1): 415-420.

Suhimi, N.M. & Mohammad, A.W. 2012. Pengoptimuman proses pengeringan semburan gelatin dengan menggunakan kaedah sambutan permukaan. *Sains Malaysiana* 41(8): 983-991.

Xiang, Z. 2006. Advances in homology protein structure modeling. *Current Protein & Peptide Science* 7(3): 217-227.

Safiyyah Shahimi, Sahilah Abd. Mutalib*, Rozida Mohd. Khalid & Maaruf Abdul Ghani
School of Chemical Sciences and Food Technology
Faculty of Science and Technology
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Darul Ehsan
Malaysia

Mohd Fadly Lamri
Politeknik Metro Kuantan
A-5, Jalan Tun Ismail 2, Sri Dagangan 11
25000 Kuantan, Pahang Darul Makmur
Malaysia

Mohd. Faizal Abu Bakar & Mohd. Noor Mat Isa
Malaysia Genome Institute
Ministry of Science Technology and Inovasion
Bangi, 43600 Kajang, Selangor Darul Ehsan
Malaysia

*Corresponding author; email: sahilah@ukm.edu.my