JOURNAL OF PURE AND APPLIED MICROBIOLOGY, August 2014. Vol. **8**(4), p. 2801-2816

In silico Protein Prediction of the 16S ribosomal RNA Gene from Novel Rhizobacterium *Bacillus cereus* UPMLH24

Fitri Abdul Aziz Zakry[1]*, Halimi Mohd Saud[2], Khairuddin B. Abdul Rahim[3] and Osumanu H. Ahmed[1]

[1]Faculty of Agriculture and Food Sciences, Universiti Putra Malaysia Bintulu Campus Sarawak, 97008 Bintulu, Sarawak, Malaysia.
[2]Faculty of Agriculture, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.
[3]Division of Agrotechnology and Biosciences, Malaysian Nuclear Agency, 43000 Kajang, Selangor, Malaysia.

(Received: 18 March 2013; accepted: 20 May 2013)

In this study, computational methods were used to predict information hidden in the small subunit 16S rRNA gene from newly isolated plant growth-promoting rhizobacteria *B. cereus* strain UPMLH24. In the present study, the computational methods employed revealed that the small subunit 16S rRNA gene sequence from *B. cereus* strain UPMLH24 contained a number of small open reading frames that encoded several functional proteins. Computational predictions classified seven phyla of organisms associated with small open reading frames of novel *B. cereus* strain UPMLH24. The data generated from computational predictions in the present study could form the basis for further research to advance new hypotheses in the microbiology of rhizobacteria.

Key words: Bioinformatics, Gene prediction, Open reading frames, Protein, Rhizobacteria, Translation.

In recent years, bioinformatics has emerged as a powerful tool for the rapid investigation of molecular sequences and has taken the center stage in molecular biology research as a tool for transforming the voluminous molecular sequence data into new knowledge and information. Advances in this field have expanded the horizons in biological research. Bioinformatics uses mathematical and statistical concepts to develop computer algorithm in accordance with theories and concepts in biology. With databases of DNA and protein sequences becoming increasingly available in public databases, bioinformatics has evolved into an essential tool in the development of new concepts, theories and experimental approaches in biological research.

* To whom all correspondence should be addressed.
Tel.: 6086-855831; Fax: 6086-855415;
E-mail: zakryfitri@upm.edu.my

The small subunit of the 16S ribosomal RNA gene (16S rRNA) is a highly conserved region found in almost all organisms. In prokaryotes, the 16S rRNA genes occur in at least one copy in the genome. Ribosomal RNAs (16S, 23S, 5S) and ribosomal proteins play very important roles in the formation of ribosomes and in the control of protein translation[1]. The 16S rRNA gene is widely studied as a molecular signature and is commonly adopted as the standard for identification and phylogenetic analysis[2,3] since it contains conserved regions and other more variable sequences that vary according to species or families[3].

In prokaryotic organisms, rRNAs play several important functions in protein synthesis and its control. The most important role of the 16S rRNA in protein synthesis is in the initiation of translation by specific base pairing of its 3' end with the Shine-Dalgarno ribosome sequence and interactions with translation initiation factors[4]. The interaction between tRNAs and the 16S rRNA also

helps to establish final codon/anticodon recognition[5].

Open reading frames (ORFs) are commonly used to identify putative genes in molecular DNA sequences. According to Basrai *et al*.[6], an open reading frame is defined as a segment of DNA capable of encoding a protein beginning with an ATG and ending at a termination or stop codon. The start codon may not always be ATG (coding for methionine), especially in prokaryotes. For example, *Escherichia coli* uses 83% ATG (AUG), 14% GTG (GUG), 3% TTG (UUG) and one or two others, for example ATT (AUU) and CTG (CUG)[7] as start codons. Bioinformatics programs allow the use of alternative start codons for gene prediction. Interesting information and features of gene can be obtained from public bioinformatics databases, for example, those of the National Center for Biotechnology Information (GenBank), European Molecular Biology Laboratory (EMBL) and DNA Data Bank of Japan (DDBJ).

In the early years of DNA sequencing, ORFs were investigated in relation to the complete sequences of the genome. For example, the sequencing of the *Saccharomyces cerevisiae* genome led to the construction of numerous yeast strains and plasmids[8]. However, the generated gene information from the whole genome included only regions of at least 100 contiguous codons, while small open reading frames (sORFs) encoding functional proteins were mostly missed. These missed sORFs were only considered after evidences of small gene expression emerged[9-10]. In addition, sORFs predict small proteins that are involved in several important peptides, such as mating pheromones, energy metabolism-proteins, proteolipids, chaperonins, stress proteins, transporters, transcriptional regulators, nucleases, ribosomal proteins, thioredoxins and metal ion chelators[6].

Whole genome sequence computational analysis usually involves at least 100 contiguous codons and it uses the whole or complete genome sequence[8]. In the present study, from 1 to 99 contiguous codons were screened. Since the 16S rRNA genes are conserved in most organisms they are commonly used to study relatedness among organisms. Hence, in the present study, it is hypothesized that the predicted sORFs from 16S rRNA sequences could be related to the involvement of horizontal gene transfer among organisms in the particular ecosystem. The transferred gene might be involved in cell-cell communications through an array of hormonal signals, for example, hormones produced by eukaryotes and hormone-like compounds produced by bacteria[12]. The cell-to-cell communication is an important mechanism to establish mutual relationship between bacteria and their hosts that subsequently create balanced and suitable ecosystem for both organisms[12].

Bacillus cereus strain UPMLH24 is an indigenous endospore former and microaerophilic bacterium that was first isolated from root surface of pepper plants (*Piper nigrum* L.). It has been established as a plant growth-promoting rhizobacterium after several series of promising findings on the promotion of early plant growth after inoculation[13,14]. Commonly, in fundamental works of PGPR research has generated a substantial number of 16S rRNA sequences used for identification of newly isolated strains and phylogenetic analysis. Those sequences were then submitted to public databases (GenBank, EMBL and DDBJ) for public references making the number of collection of sequences growing further and even faster in parallel with PGPR research progress. Efforts should be made to optimize the use of voluminous 16S rRNA sequences in the databases and its transformation for valuable knowledge and information.

This study was aimed at exploring computational methods to investigate the 16S ribosomal RNA gene sequence of PGPR *Bacillus cereus* UPMLH24 from pepper (*Piper nigrum L.*) for potential protein sequences using bioinformatics analysis.

Methodology

This study made use of selected bioinformatics programs to analyse selected molecular sequence data. Bioinformatics data were interpreted according to theories and concepts of the molecular biology of plant-microbe interactions.

Previously sequenced nucleotides of the 16S ribosomal RNA gene of *Bacillus cereus* strain UPMLH24, that yielded nearly complete sequences of 1418 bases (363 A; 324 C; 434 G; 297 T; G+C content = 53 %), was used in this study to search for gene functions (unpublished data). The sequence has been deposited in GenBank/EMBL/

DDBJ under accession number HQ876004.

The small open reading frames (sORFs) were from 1 to 99 codons[8]. The 1418 bases of *B. cereus* UPMLH24 16S rRNA nucleotide sequence was investigated from three minimal codons lengths, 20, 40 and 60, for the detection of putative sORFs using the Genomics Expression program (http://www.genamics.com/expression/index.htm). No putative ORFs were detected on 16S rRNA nucleotide sequence of *B. cereus* UPMLH24 when minimal codons were lower than 5 and higher than 88; therefore, the present study on the 16S rRNA gene contained only small ORFs or small proteins[6]. The sORFs were identified from methionine (ATG) or alternative start codons and stop codons, or known as termination codons[7].

A number of detected sORFs from 20, 40 and 60 minimal codons were investigated further by extracting each fragment of putative ORF for amino acid sequence. Each amino acid sequence was BLAST (Basic Local Alignment Search Tool) searched (on 5 October 2012) for protein similarity according to PSI-BLAST algorithm and non-redundant protein sequence database of the GenBank. BLAST finds regions of local similarity between sequences by comparing nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. PSI-BLAST allows the user to build a PSSM (position-specific scoring matrix) using the results of the first BLASTp run, making PSI-BLAST algorithm more sensitive compared to the BLASTp algorithm. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families (http://blast.ncbi.nlm.nih.gov/). The PSI-BLAST algorithm aligns protein sequences in the databases with query of putative ORF-amino acid sequence. By using the PSI-BLAST algorithm in the BLASTp program, the program generates search results that classifies between database-protein sequences having sequence(s) with significant alignment, with E-value better than threshold and of sequence(s) with significant alignment where the E-value is lower than threshold. Only the top rank of BLASTp search results was selected on the basis of the highest percentage of identity. Where two or more results gave similar percentage of identity, they were evaluated and ranked based on lower E-value.

The organisms associated with predicted proteins were summarised, classified and listed according to phylum. Phylum classification was according to UNIPROT Taxonomy (http://www.uniprot.org). Each predicted putative protein in the summarised list was analysed, selected and subsequently interpreted for significant predicted proteins with highest identity and their persistency in the overall ORF maps (predicted ORFs available in each of the 20, 40 and 60 minimal codon ORF maps).

The selected sORFs-nucleotide sequences that predicted a functional protein was subjected to MEGABLAST search (highly similar sequences optimisation) using BLASTn program and compared with the whole genome sequence similarity based on the NCBI Genomes (Chromosome) Database and High Throughput Genomic Sequences Database. The BLAST search result was selected from the top ranked listing with the highest similarity. Initially, investigation on the literature was made according to information available in whole genome databases such as NCBI Genome (Chromosome) database and High Throughput Genomics Sequence Databases. Information from the citations and genomic databases were used to trace and acquire more information related to the predicted protein. The literature mining was made using public databases of scientific literature including SciVerse-Scopus (http://www.scopus.com/home.url), PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and Google Scholar (http://scholar.google.com) that were available on the World Wide Web. The mined literature related to predicted putative proteins were reviewed and interpreted. The bioinformatics workflow in the present study is shown in Fig. 1.

RESULTS

Predictive Analysis of 20, 40 and 60-Minimal Codon Small Open Reading Frame (sORF) Maps of the 16S rRNA Gene

A total of 43 sORFs were identified in 20, 40 and 60 minimal codon ORF maps (Table 1; Fig. 2). Forty sORFs-amino acid sequences with sequence lengths in the range of 20 to 89 amino acids were identified after BLAST search and had significant similarities with proteins from the GenBank non-redundant protein database. Of the

Table 1. Top ranked BLAST search results of predicted ORF-amino acid sequence from 20, 40 and 60 minimal codons of 16S rRNA ORF map (*Bacillus cereus* strain UPMLH24)

ORF Map	Putative ORF position on the map	Amino acid sequence extracted from putative ORF [length]	Description	E-value	% Identity
20 minimal codons	1	MRFKMLSGISPGFPE LSQSYGQVTHVLL TRPPLTS [35]	Conserved Hypothetical protein [*Streptomyces* sp. AA4]	7e-05	72
	2	LLRQTFVHCGRFPTA ASRRSLGRVSVPV WPITLSGRLRIVAL VSRYLTN [49]	pG1 protein [*Bacillus amyloliquefaciens*]	9e-26	100
	3	LLPPVGVWAVSQSQ CGRSPSQVGYASL PW [29]	ORF40s [*Pinus koraiensis*]	6e-08	100
	4	MRSRPERVIGHTGTE TRPRLLREAAVGN LPQWTKV [35]	Hypothetical protein FNP_0147 [*Fusobacterium nucleatum sub* sp. polymorphum ATCC 10953]	2e-11	82
	5	LFFPNNRVLRPESLH HSRGVAPSDFRPL RKIPYCCLP [37]	Unknown protein [*Streptococcus suis* 98HAH33]	2e-06	69
	6	MDESLTEQRRVSDE GFRVVKLCC [23]	Hypothetical protein IGI_00001, partial [*Bacillus cereus* HuB2-9]	7e-18	100
	7	LPPTYYRGCWHVVS RGFLVRYRQGASL FN [29]	pG1 protein [*Listeria grayi* DSM 20601]	2e-18	96
	8	LTVPNQKATANYVP AAAVIRRWQALS GIIGRKARAGGFL SLM [42]	Hypothetical protein [*Bacillus cereus* G9241]	1e-19	98
	9	LLPTLSRLSVSYRPE SRLRHWCSSISLRI SPLHMEFHFPLLH SSLPVSNDPPRLS RGLSHQT [62]	pG1 protein [*Lactobacillus jensenii* 269-3]	1e-18	79
	10	LRPYSPGGVLNALTS ALKGGNPLTLSTHRL RRGLPGYLILFAPHAFA PQCQLQTRKSPSPLVFL HISTHFTATHGIPLSSSA LKSPSFQ [89]	Hypothetical protein HMPREF9520_02283 [*Enterococcus faecalis* TX1467]	4e-47	89
	11	LETGRLECRRGKWN SMCSGEMRRDME EHQWRRRLSGL [37]	Hypothetical protein MLEA_001930 [*Mycoplasma leachii* 99/014/6]	2e-07	75
	12	LGDLSAEEESGIPCV AVKCVEIWRNTS GEGDFLVCN [36]	Putative cytoplasmic protein [*Listeria seeligeri* FSL S4-171]	8e-07	82

13	LTTTMHHLSLCSRRR SPISRVFRGCQDL VRFFALLRIKPHA PPLVRAPVNSFEF QPCGRTPQAECL MR [68]	Conserved Hypothetical protein [*Shigella* sp. D9]	5e-04	100
14	MLHRLCGPPSIPLSF SLAAVLPRRSA [26]	No sequences producing significant alignments with E-value better than threshold		
15	LGSTAARLKLKGIDG GPHKRWSMWFNS KQREEPYQVLTSS ENPRDRASPSGAE [53]	Hypothetical protein BACCAP_03831 [*Bacteroides capillosus* ATCC 29799]	4e-18	92
16	LTGARTSGGACGLIR SNAKNLTRS [24]	Ribosomal protein S10 [Medicago truncatula]	3e-06	86
17	LKTLEIGLLLREQSD RWCMVVVSSCRE MLG [30]	No sequences producing significant alignments with E-value better than threshold		
18	LRSLRDLTQHLTTRA DDNHAPPVTLLP KEKPYL [33]	Hypothetical protein HG1285_07402 [*Hydrogenivirga* sp. 128-5-R1-1]	0.001	95
19	LPVTNRRKVGMTSN HHAPYDLGYTRA TMD GTKSCKTARWS [40]	Conserved Hypothetical protein [*Listeria innocua* FSL J1-023]	4e-11	96
20	LQPTIRTENGFMRLA PPRGLAALCTVHCSTC VAQVIRGM MI [41]	Putative lipoprotein [*Clostridium botulinum C str. Eklund*]	3e-06	79
21	MPLMTWATHVLQW TVQRAARPRGGANL IKPFSVRIVGCNSPT [42]	Hypothetical protein GCWU000323_01937 [*Leptotrichia hofstadii* F0254]	2e-04	74
22	LQLAYMKLESLVIAD QHAAVNTFPGLV HTARHTTRVCNT RSRWGNLFGASC SRANGLRACSYE VSGGRVSN	Hypothetical protein STRINF_00003 [*Streptococcus* infantarius subsp infantarius ATCC BAA-TWVTCP [77]	8e-21	98 102]
23	MKLESLVIADQHAA VNTFPGLVHTARHT TRVCNTRSRWGNLF GASCSRANGLRACS YEVSGGRVSNTWVT CP [72]	Hypothetical protein MUY_01125 [*Bacillus licheniformis* WX-02]	5e-17	97
24	LESLVIADQHAAVNT FPGLVHTARHTTRVC NTRSRWGNLFGASC SRANGLRACSYEVS GGRVSNTWVTCP [70]	Hypothetical protein MUY_01125 [*Bacillus licheniformis* WX-02]	4e-16	100
25	LYTPPVTPREFVTPE VGGVTFLEPAAVER MD [31]	No sequences producing significant alignments with E-value better than threshold		

	26	LRLAPKRLPHRLRVL QTLVV [20]	Conserved Hypothetical Protein [*Lactobacillus* jensenii 27-2-CHN]	0.003	86
40 minimal codons	1	LLRQTFVHCGRFPTA ASRRSLGRVSVPVW PITLSGRLRIVALVSR YLTN [49]	pG1 protein [*Bacillus amyloliquefaciens*]	9e-26	100
	2	LTVPNQKATANYVP AAAVIRRWQALSGII GRKARAGGFLSLM [42]	Hypothetical Protein [*Bacillus cereus* G9241]	1e-19	98
	3	LLPTLSRLSVSYRPE SRLRHWCSSISLRISP LHMEFHFPLLHSSLP VSNDPPRLSRGLSHQ T [62]	pG1 protein [*Lactobacillus jensenii* 269-3]	1e-18	79
	4	LRPYSPGGVLNALTS ALKGGNPLTLSTHRL RRGLPGYLILFAPHA FAPQCQLQTRKSPSP LVFLHISTHFTATHGI PLSSSALKSPSFQ [89]	Hypothetical Protein HMPREF9520_02283 [*Enterococcus faecalis* TX1467]	4e-47	89
	5	LTTTMHHLSLCSRRR SPISRVFRGCQDLVR FFALLRIKPHAPPLV RAPVNSFEFQPCG RTPQAECLMR [68]	Conserved Hypothetical Protein [*Shigella* sp. D9]	5e-04	100
	6	LGSTAARLKLKGID GGPHKRWSMWFNS KQREEPYQVLTSSE NPRDRASPSGAE [53]	Hypothetical Protein BACCAP_03831 [*Bacteroides capillosus* ATCC 29799]	4e-18	92
	7	LQPTIRTENGFMRLA PPRGLAALCTVHCS TCVAQVIRGMMI [41]	Putative lipoprotein [*Clostridium botulinum* C str. Eklund]	3e-06	79
	8	MPLMTWATHVLQW TVQRAARPRGGANL IKPFSVRIVGCNSPT [42]	Hypothetical Protein GCWU000323_01937 [*Leptotrichia hofstadii* F0254]	2e-04	74
	9	LQLAYMKLESLVIAD QHAAVNTFPGLVHT ARHTTRVCNTRSRW GNLFGASCSRANGL RACSYEVSGGRVSN TWVTCP [77]	Hypothetical Protein STRINF_00003 [*Streptococcus infantarius sub* sp. *infantarius* ATCC BAA-102]	8e-21	98
	10	MKLESLVIADQHAA VNTFPGLVHTARHT TRVCNTRSRWGNLF GASCSRANGLRA CSYEVSGGRVSNT WVTCP [72]	Hypothetical Protein MUY_01125 [*Bacillus licheniformis* WX-02]	5e-17	97
	11	LESLVIADQHAAVNT FPGLVHTARHTTRVC NTRSRWGNLFGASC SRANGLRACSYEVS GGRVSNTWVTCP [70]	Hypothetical Protein MUY_01125 [*Bacillus licheniformis* WX-02]	4e-16	100

60 minimal codons	1	LLPTLSRLSVSYRPE SRLRHWCSSISLRISP LHMEFHFPLLHSSLP VSNDPPRLSRGLSHQ T [62]	pG1 protein [*Lactobacillus jensenii* 269-3]	1e-18	79
	2	LRPYSPGGVLNALTSAL KGGNPLTLSTHRL RRGLPGYLILFAPHA FAPQCQLQTRKSPSP LVFLHISTHFTATHGI PLSSSALKSPSFQ [89]	Hypothetical Protein HMPREF9520_02283 [*Enterococcus faecalis* TX1467]	4e-47	89
	3	LTTTMHHLSLCSRRR SPISRVFRGCQDLVR FFALLRIKPHAPPLV RAPVNSFEFQPCGRT PQAECLMR [68]	Conserved Hypothetical Protein [*Shigella* sp. D9]	5e-04	100
	4	LQLAYMKLESLVI ADQHAAVNTFPG LVHTARHTTRVCN TRSRW GNLFGASCSRANGL RACSYEVSGGRVSN TWVTCP [77]	Hypothetical Protein STRINF_00003 [*Streptococcus* i*nfantarius sub*sp. *infantarius* ATCC BAA-102]	8e-21	98
	5	MKLESLVIADQHAA VNTFPGLVHTARHT TRVCNTRSRWGNLF GASCSRANGLRACS YEVSGGRVSNTWVT CP [72]	Hypothetical Protein MUY_01125 [*Bacillus licheniformis* WX-02]	5e-17	97
	6	LESLVIADQHAAVN TFPGLVHTARHTTR VCNTRSRWGNLFGA SCSRANGLRACSYE VSGGRVSNTWVTCP [70]	Hypothetical Protein MUY_01125 [*Bacillus licheniformis* WX-02]	4e-16	100

43 sORFs, three sORFs-amino acid sequences (all from the 20 minimal codon map) had no similarity with entries in the non-redundant proteins database. All sORFs in the 40 and 60 minimal codon length maps showed similarities to known proteins.

Table 2 lists the 21 organisms that had relationships with the *Bacillus cereus* proteins, including one sORF where the protein matched one previously reported for *Bacillus cereus*. The organisms were classified according to phylum and other classification categories such as genus/species, predicted protein, position in ORF maps and degree of identity with the newly isolated strain UPMLH24 of *Bacillus cereus*. Computational analysis classified the predicted organisms into seven phyla, namely *Actinobacteria*, *Aquificae*, *Firmicutes*, *Fusobacteria*, *Proteobacteria*, *Tenericutes* and *Viridiplantae*. *Firmicutes* was the most diverse in proteins associated with those of *B. cereus* strain UPMLH24 sORFs (13 different species of organisms; 61.19 %), followed by two species from *Fusobacteria* (9.52 %) and two species from *Viridiplantae* (9.52 %). Only one species (4.76 % of each *phylum*) was identified for the *phyla Actinobacteria*, *Aquificae*, *Proteobacteria* and *Tenericutes*.

Among the predicted proteins that had the most similarities with sORFs of the 16S rRNA gene from *B. cereus* UPMLH24 were pG1 protein, a lipoprotein, a cytoplasmic protein and ribosomal protein S10. The pG1 protein was predicted from two bacterial species, namely *Bacillus amyloliquefaciens* (100 % identity) and *Listeria*

Table 2. Phylum classification of proteins from 21 species of organisms related to predicted open reading frame proteins from 20, 40 and 60 minimal codon maps of the small subunit 16S ribosomal RNA gene of *Bacillus cereus* strain UPMLH24

Phylum	Genus/species	Position in ORF map [No. ORF(s)]	Predicted ORF-protein	% Identity
Actinobacteria	*Streptomyces* sp.	ORF1 of 20MC [1]	Hypothetical protein	72
Aquificae	*Hydrogenivirga* sp.	ORF18 of 20MC [1]	Hypothetical protein	95
Firmicutes	*Bacillus amyloliquefaciens*	ORF2 of 20MC; ORF1 of 40MC [2]	pG1 protein	100
	Bacillus cereus	ORF8 of 20MC; ORF2 of 40MC [2]	Hypothetical protein	98
	Bacillus licheniformis	ORF23 and ORF24 of 20MC; ORF10 and ORF11 of 40MC; ORF5 and ORF6 of 60MC [6]	Hypothetical protein	97-100
	Bacteroides capillosus	ORF15 of 20MC; ORF6 of 40MC [2]	Hypothetical protein	92
	Clostridium botulinum C	ORF20 of 20MC; ORF7 of 40MC [2]	Putative lipoprotein	79
	Enterococcus faecalis	ORF10 of 20MC; ORF4 of 40MC; ORF2 of 60MC [3]	Hypothetical protein	89
	Lactobacillus jensenii	ORF9 of 20MC; ORF3 of 40MC; ORF1 of 60MC [3]	Conserved Hypothetical protein	81
	Listeria grayi	ORF7 of 20MC [1]	pG1 protein	96
	Listeria innocua	ORF19 of 20MC [1]	Conserved Hypothetical protein	96
	Listeria seeligeri	ORF12 of 20MC [1]	Putative cytoplasmic protein	82
	curdlanolyticus	ORF6 of 20MC [1]	Conserved Hypothetical protein	96
	Streptococcus infantarius sub sp. *infantarius*	ORF22 of 20MC; ORF9 of 40MC; ORF4 of 60MC [3]	Hypothetical protein	98
	Streptococcussuis	ORF5 of 20MC [1]	Unknown protein	69
Fusobacteria	*Fusobacterium nucleatum* sub sp. polymorphum	ORF4 of 20MC [1]	Hypothetical protein	82
	Leptotrichia hofstadii	ORF21 of 20MC; ORF8 of 40MC [2]	Hypothetical protein	74
Proteobacteria	*Shigella* sp.	ORF13 of 20MC; ORF5 of 40MC; ORF3 of 60MC [3]	Conserved Hypothetical protein	100
Tenericutes	*Mycoplasma leachii*	ORF11 of 20MC [1]	Putative uncharacterised protein	75
Viridiplantae	*Medicago truncatula*	ORF16 of 20MC [1]	Ribosomal protein S10	86
	Pinus koraiensis	ORF3 of 20MC [1]	ORF40s	100

Note: MC = minimal codons, minimal codons of open reading frame. Analysis and interpretation of gene prediction was made based on 3 ORF maps of 3 different minimal codons (20, 40 and 60 minimal codons of 1418 bases length of 16S ribosomal RNA gene from *Bacillus cereus* strain UPMLH24).

Table 3. Literature mining on selected predicted putative sORF-proteins of small subunit 16S rRNA gene from *Bacillus cereus* strain UPMLH24

Predicted protein (species of organism related to putative sORF-amino acid sequence)	sORF-nucleotide sequence (length)	Results from top BLAST search of sORF-nucleotide sequence	Information from literature mining
pG1 protein (*Bacillus amyloliquefaciens*) Position on map: ORF2-20MC; ORF1-40MC.	CTAGTTGGTGAGGTAA CGGCTCACCAAGGCAA CGATGCGTAGCCGACC TGAGAGGGTGATCGGC CACACTGGGACTGAGA CACGGCCCAGACTCCT ACGGGAGGCAGCAGTA GGGAATCTTCCGCAAT GGACGAAAGTCTGACG GAGCAA (150)	NCBI Genomes (Chromosome) Database: *Bacillus thuringiensis* MC28 chromosome, complete genome E-value = 2e-71 Identity = 100 % High Throughput Genomic Sequences Database: *Bacillus subtilis* subsp. natto BEST195 [AP011541] E-value = 6e-72 Identity = 100 %	The *Bacillus cereus* group includes *B. cereus, B. thuringiensis, B. anthracis*, all having similar biochemical, morphological and genetic relations (Bavykin et al., 2004). Nucleotide sequence of predicted ORF associated with a natto production strain *Bacillus subtilis* natto (Nishito *et al.* 2010). Amino acid sequence of predicted ORF was associated with γ-polyglutamic acid producing strain (C06) of *B. amyloiquefaciencs* (Liu *et al.* 2010).
Hypothetical Protein MUY_01125 (*Bacillus licheniformis* WX-02) Position on map: E-value = 1e-60 Nucleotide sequence of predicted ORF is ORF23 and ORF24 of 20MC; ORF10 and ORF11 of 40MC; ORF5 and ORF6 of 60MC.	CTGGAATCGCTAGTAA TCGCGGATCAGCATGC CGCGGTGAATACGTTC CCGGGCCTTGTACACA CCGCCCGTCACACCACG AGAGTTTGTAACACC CGAAGTCGGTGGGGTA ACCTTTTTGGAGCCAGC CGTGCAGTCGAGCGAA TGGATTAAGAGCTTGCT CTTATGAAGTTAGCGG CGGACGGGTGAGTAAC ACGTGGGTAACCTGCC CATAA (215)	NCBI Genomes (Chromosome) Database: *Bacillus thuringiensis* MC28 chromosome, complete genome Identity = 100 % High Throughput Genomic Sequences Database: *Bacillus subtilis* subsp. natto BEST195 [AP011541] E-value = 7e-58 Identity = 98 %	The *Bacillus cereus* group includes *B. cereus, B. thuringiensis*, *B. anthracis*, all having similar biochemical, morphological and genetic relations (Bavykin *et al.*, 2004). associated with a natto production strain *Bacillus subtilis* natto (Nishito et al. 2010). This strain (WX-02) of *Bacillus licheniformis* is related to γ-polyglutamic acid producing strain which was isolated from saline soil (Wei et al., 2010; Yangtse et al., 2012)

grayi (96 % identity). Putative lipoprotein from *Clostridium botulinum C* (79 % identity), putative cytoplasmic protein from *Listeria seeligeri* (82 % identity) and ribosomal protein S10 from *Medicago truncatula* (86 % identity) were the other proteins displaying similarities to the 16S rRNA of B. cereus UPMLH24. From the total of 30 identified sORFs, six sORFs-proteins were associated with proteins of *Bacillus licheniformis* strain MUY_01125 and were the most frequently encountered, followed by *Enterococcus faecalis* (3 sORFs), *Lactobacillus jensenii* (3 sORFs), *Streptococcus infantarius sub*sp. infantarius (3 sORFs), *Shigella* sp. (3 sORFs), *Bacillus amyloliquefaciens* (2 sORFs), *Bacillus cereus* (2 sORFs), *Bacteroides capillosus* (2 sORFs), *Clostridium botulinum C* (2 sORFs), *Leptotrichia hofstadii* (2 sORFs). One sORF matched proteins of *Streptomyces* sp., *Hydrogenivirga* sp., *Listeria grayi*, *Listeria innocua*, *Listeria seeligeri*, *Paenibacillus curdlanolyticus*, *Streptococcus suis*, *Fusobacterium nucleatum sub*sp. *polymorphum*, *Mycoplasma leachii*, *Medicago truncatula* and *Pinus koraiensis*. Four sORFs having the highest identity (100 %) were predicted as pG1 (*Bacillus amyloliquefaciens*), hypothetical protein (*Bacillus licheniformis*), conserved hypothetical protein (*Shigella* sp.) and the ORF40s protein (*Pinus koraiensis*).

Literature Mining Using Selected Predicted sORFs of Newly Isolated Plant Growth-Promoting Rhizobacterium *Bacillus cereus* UPMLH24

Literatures mining on selected sORF-proteins predicted in the present study are presented in Table 3. The selected sORF-proteins were pG1 protein (*B. amyloliquefaciens*) and hypothetical protein (*B. licheniformis* strain WX-02); these proteins had the highest identity (100 %) and the most frequently predicted in small open reading frame maps of the 16S rRNA gene of *B. cereus* strain UPMLH24 (Table 2). Based on

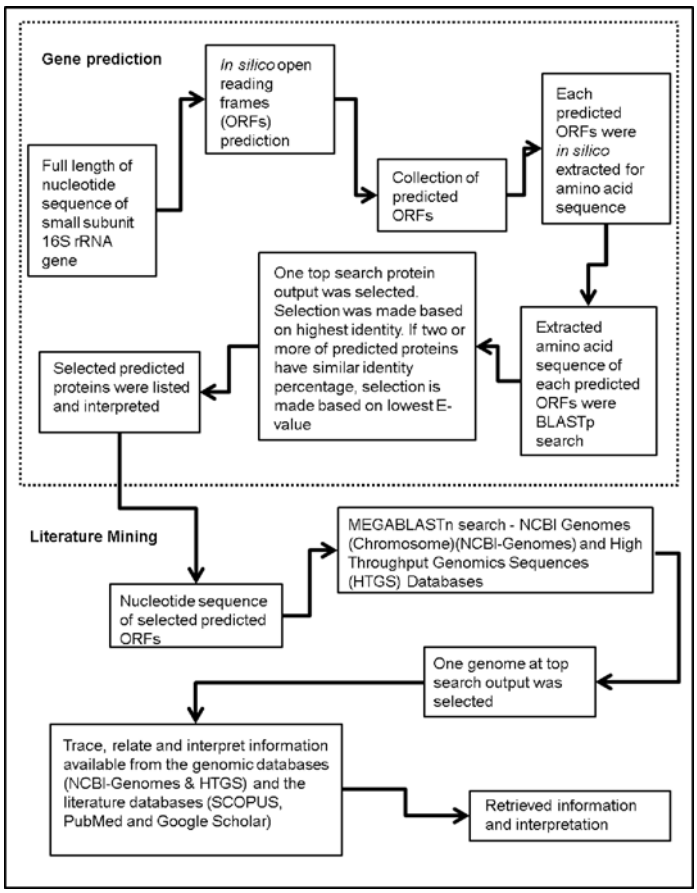

Fig. 1. Bioinformatics workflow

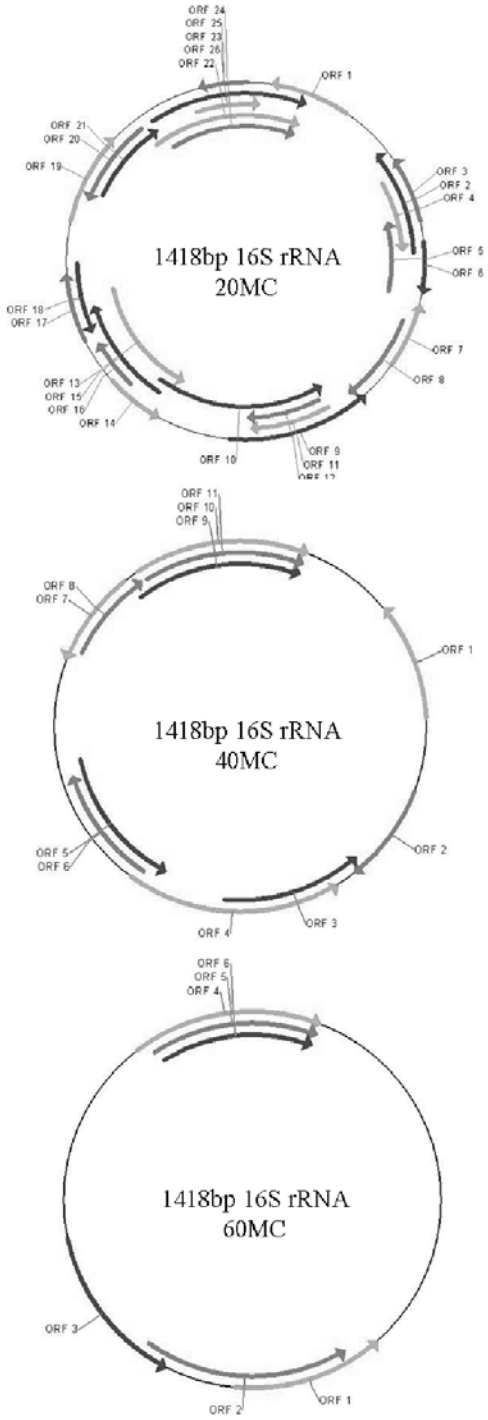

Fig. 2. The 20, 40 and 60 minimal codons (MC) ORF maps of the 1418bp 16S rRNA gene from Bacillus cereus strain UPMLH24. The orientations of the putative ORFs are indicated by the orientations of the arrows.

information in the literature mining activity, the predicted pG1 protein from *B. amyloliquefaciens* may be related to a similar protein from the gamma-polyglutamic acid producing strain of *B. amyloliquifaciens*. BLAST sORF-nucleotide sequence of 16S rRNA gene from *B. cereus* strain indicated that the strain may be related to *B. subtilis sub*sp. natto that is used in natto production.

A hypothetical protein from *B. licheniformis* strain WX-02 was frequently predicted in small open reading frames (sORFs) of 16S rRNA gene of the novel plant growth-promoting rhizobacteria *B. cereus* strain UPMLH24. BLAST sORF-nucleotide sequence indicated that the strain WX-02 is related to halotolerant, gamma-polyglutamic producing bacterium. Based on the NCBI Genomics (Chromosome) Database, the sORFs of this protein that was present in strain of *B. cereus* UPMLH24 was identical with a protein on the *B. thuringiensis* MC28 chromosome.

DISCUSSION

Voluminous and rapidly expanding molecular sequence data in the public databases, viz. GenBank, European Molecular Biology Laboratory and DNA Data Bank of Japan have paved the way to a new era for the development of new knowledge and discovery. However, these free and publicly available data are still not fully utilised to transform the raw molecular sequence data into tools for the advancement of science. In the present study, computational methods have been used to discover possible protein attributes in the small subunit 16S ribosomal RNA genes newly isolated from the plant growth-promoting rhizobacteria *Bacillus cereus* strain UPMLH24.

Computational open reading frames prediction on the 16S rRNA gene from the novel B. cereus strain UPMLH24 produced forty three sORFs, with predictions of translated proteins made for 40 sORFs. These sORFs translated proteins and peptides of 20 to 89 amino acids sequence length. The 16S rRNA gene is a part of small subunit of ribosomal RNA genes with sequence lengths varying between 1404 bp of *Streptomyces coelicolor*[15] and 1549 bp of *B. subtilis*[16]. In the *Escherichia coli*, the small subunit

contains 21 r-proteins denoted S1 to S21 (S for small subunit). Protein lengths range from 70 to 557 amino acids[17]. However, in the large subunit, smaller r-protein can be obtained from a 38 amino acid sequence (L36)[17]. According to Simonen and Palva[18], signal peptides are among the small proteins and they vary between 18 and 35 amino acid residues in length. Hence, some of the sORF proteins predicted in the 16S rRNA gene of UPMLH24, which are mostly found in the 20-minimal codon length map, may be related to signal peptides.

In the present study, computational analysis predicted proteins that are present in seven phyla and 20 species might be associated with novel *B. cereus* strain UPMLH24 on the basis of in silico BLAST protein-protein similarity and comparison. Only one sORF was related to a formerly reported protein of the same species, *Bacillus cereus* strain G9241. The seven phyla that were predicted were *Actinobacteria, Aquificae, Firmicutes, Fusobacteria, Proteobacteria, Tenericutes* and *Viridiplantae*. Among the *phyla,* the Firmicutes were most frequently associated with the predicted sORF proteins. The present computational study is larger than that experimentally reported for *Proteobacteria* by Aguirre-Garrido *et al*.[19], who demonstrated the involvement of three different phyla having three different classes (alpha, beta and gamma). In that study, Firmicutes were among the most abundant taxa besides alpha-proteobacteria and actinobacteria. Those phyla were isolated from three cactus species grown under semi-arid highlands in Central Mexico[19]. Cardoso *et al*.[20] reported that phyla *Bacteroidetes* and *Firmicutes* dominated in the faecal samples of the giant land snail, Achatina fulica. However, both phyla were less abundant in crop fluid samples of *A. fulica*. The abundant and diversity of Firmicutes might be related to the type of samples or level of external or environmental stresses. The Firmicutes might be abundant and mostly populated in the environmental samples or external environment such as soil, plant rhizosphere and faeces[19,20] where bacteria survival and persistency will be affected directly from environmental stresses such as sunlight, heavy metals, chemical fertiliser, pesticide, herbicide and other toxic substances rather than within living organisms except digestive tracts which could be exposed to extreme acidic conditions[21]. This phenomenon was supported by Xiao *et al*.[22] who reported that bacteria isolated from sea anemone of coastal water of the Naozbou island of South China Sea were mostly abundant and diverse with *phylum Firmicutes* at 41%. Chen *et al*.[23] indicated that *Firmicutes* (34%) was among the most abundant and diverse isolates besides *Proteobacteria* (47%) which were isolated from ancient salt deposits from Yipinglang Salt Mine in the Yunan Province of China. The 16S rDNA sequence analysis demonstrated in the previous years, reported that Firmicutes was a dominant phylum associated with the gastrointestinal microbial communities in cattle[24,25], pigs[26] and humans[27], particularly those related to the broad genus Clostridium. Bacteria under phylum Firmicutes is always associated with its capability to form endospore[28], a capsular component which formed for its survivability under hazardous environment[29]. Thus endospore forming species under Firmicutes was more abundant and diverse within competitive and imbalance environment rather than non-endospore forming species.

Of seven phyla traced in the present study, one phylum was associated with plant kingdom, *Viridiplantae*. The 16S rRNA genes from UPMLH24 was predicted with two sORFs that had similarities with two proteins, ribosomal protein S10 (86 % similar) from Medicago truncatula and ORF40s from *Pinus koraiensis* (100 % similar). This might be the first computational evidence indicating that newly isolated strain from *P. nigrum rhizosphere, B. cereus* is plant-associated bacteria or rhizobacteria. The occurrence of 16S rRNA-sORF of B. cereus strain UPMLH24 where similar proteins are present in plants may indicate interspecies communication that occurs between bacteria and plants through bacteria-plant signalling[30]. For example, in Vibrio harveyi, cell-cell communication is carried out by two-component signal transduction proteins, whereby the AI-1 peptide mediates intraspecies communication and the AI-2 peptide mediates interspecies cell-to-cell communication. Moreover, AI-2 is a universal signal that could be involved in interspecies communication[30]. Besides that, indole-3-acetic (IAA) acid is also reported as a reciprocal signalling molecule in bacteria-plant communication. IAA features is the main mechanism of plant growth

promotion by plant growth-promoting rhizobacteria[12,31]. According to Hughes and Sperandio[12], quorum sensing is not restricted to bacterial cell-cell communication, but also involves communication between bacteria and their hosts.

Four functional small proteins were predicted in silico from the open reading frames of the 16S rRNA gene of B. cereus strain UPMLH24, namely pG1 protein (29 to 62 sORF-amino acid sequence length), a lipoprotein (41 sORF-amino acid sequence length), a cytoplasmic protein (36 sORF-amino acid sequence length) and the ribosomal protein S10 (24 sORF-amino acid sequence length). The pG1 protein was the most frequently predicted in sORFs of UPMLH24. From the literature mining activity, pG1 protein was a gamma-polyglutamic acid (PGA) and may be associated with two functional roles viz., natto production[32] or it acted as a biocontrol agent through colonisation on fruit surface on which it then formed a protective layer[33]. Findings in the present study also indicated that highly similar sORFs of the hypothetical protein from *B. licheniformis* strain WX-02 was also related to the poly-d-ã-glutamic acid (PGA) producing bacterial strain isolated from saline soil and designated as a halotolerant strain[34,35]. According to Oh *et al*.[36], PGA was among the cluster of encoded genes that endowed pathogenic bacilli with capsular material, which allowed their escape from the innate host immune response, thereby aiding in pathogenesis. In addition, PGA is a main constituent of the sticky material in natto, a Japanese traditional food product made from soybeans that have been steamed and then fermented by PGA-producing bacteria[37]. However, PGA producing strains can perform different functions, depending on the species and their environment[32,33,38-40]. Thus PGA producing strains have potentially multipurpose applications in various industries, including wastewater treatment, food products, drug delivery, medical adhesives and vaccines. For example, PGA nanoparticles have been used for on-site drug release in cancer chemotherapy and in tissue engineering[41]. Viewed from the perspective of bacterial functionality, the PGA predicted in the present study may be related to endospore formation or the persistency of *B. cereus* in the plant rhizosphere, thereby delivering benefits to their plant host indirectly. Further studies need to be conducted to ascertain the possible roles of PGA in B. cereus and how that affects its immediate environment.

Lipoproteins are proteins containing lipid covalently linked to an N-terminal cysteine residue[42]. Sutcliffe and Russell[42] reviewed the functions of lipoprotein in relation to antibiotic resistance, substrate-binding in transport systems, adhesins, protein secretion, signalling systems, bacterial conjugation, sporulation, and other functions. Lipoproteins are among the cytoplasmic proteins active in many metabolic functions of the organism. Among such cytoplasmic proteins is beta-lactamase, a signal peptide produced by *Bacillus cereus*[18]. Ribosomal protein S10, which was predicted and found located in the 20-minimal codon map of 16S rRNA gene of UPMLH24 (ORF16) produced a short protein sequence (24-amino acid) (Table 2). According to Friedman *et al*.[43], ribosomal protein S10 was involved in anti-termination of transcription. Woolstenhulme and Hill[44] reported that ribosomal protein S10 was one of the several tertiary binding proteins (S2, S3, S10, S14 and S21) that required all primary and secondary proteins to be bound first. Unlike other protein groups, this protein required thermal activation before binding took place[44]. The amino acid length of the predicted sORF-ribosomal protein S10 was shorter (24-amino acids in length) than reported for the *E. coli* ribosomal protein S10 (103 amino acids in length)[17]. The short proteins predicted here could be related to the small size of the 16S rRNA gene; possibly, the full length ribosomal protein S10 sequence was missed in this study. Future studies could aim to obtain the complete sequence of the 5S-28S ribosomal RNA gene that might be about 4500 nucleotides in size[17], hence increasing the possibility of acquiring complete sequence of ribosomal protein S10.

In the present study, it was also found that functional 16S nucleotide fragment encoded for pG1 protein of the newly isolated rhizobacteria *B. cereus* strain UPMLH24 was highly similar (100 %) with sequence found within complete genome of *B. thuringiensis* MC28 chromosome. According to Bavykin *et al*.[45], constituents of the *Bacillus cereus* group, which included *B. cereus, B. thuringiensis* and *B. anthracis* were related in biochemical, morphological and genetic aspects. *B. cereus* can be distinguished from *B.*

thuringiensis by screening for the presence of the cry gene or its crystal protein that was associated with the latter[46].

Computational predictions in the current investigation indicated that many of the predicted proteins belonged to species that were potentially pathogenic to animals and humans. Amongst them were the animal and human pathogens *Clostridium botulinum C, Enterococcus faecalis, Fusobacterium nucleatum subsp. polymorphum, Leptotrichia hofstadii, Mycoplasma leachii, Shigella* sp. and *Streptococcus suis*[47-52]. This observation was in agreement with Berg *et al*.[53] who reviewed the occurrence of potentially human pathogenic species in the rhizosphere of plants. According to Berg *et al*.[53], various bacterial genera contained root-associated strains that could initiate bivalent interactions with both plant and human hosts. The occurrence of opportunistic and pathogenic bacterial strains predicted computationally, indicated that each pathogenic strain could have multiple resistance against diverse antibiotics in the rhizosphere. In addition, microbial competition and enhanced horizontal gene transfer among pathogenic rhizobacterial strains would also contribute to the high levels of natural resistance[53,54] and the evolution of environmental microbiota[55].

CONCLUSION

In conclusion, computational methods employed in the present study have provided useful information to help unravel the relationship between bacteria community and their host plant. Data generated from the computer program in the present study can be useful for developing new hypotheses for future research. Computational methods also indicated that the small subunit 16S rRNA gene sequence was able to provide useful sequence data from which new information can be obtained. Such data also predicted species from seven phyla that may be associated with the novel *Bacillus cereus* strain UPMLH24 from the *Piper nigrum* rhizosphere through protein-protein identity. Four putative functional proteins that may be associated with the *B. cereus* strain UPMLH24 were the gamma-polyglutamic acid protein, lipoprotein, cytoplasmic protein and ribosomal protein S10. The predicted gamma-polyglutamic acid protein was the most frequently identified. However, further experimentation needs to be conducted in the future to characterise structural features, chemical properties and biological functionalities of the predicted proteins. These computational methods provide possible clues for understanding plant root microbiome and also help to understand the origin of pathogenic species of root-associated bacteria.

ACKNOWLEDGEMENTS

Authors gratefully acknowledge the Universiti Putra Malaysia, Malaysian Pepper Board and Ministry of Higher Education Malaysia for research and financial support.

REFERENCES

1. Martwn, J.F., Barreiro, C., González-Lavado, E., Barriuso, M. Ribosomal RNA and ribosomal proteins in corynebacteria. *J. Biotechnol.*, 2003; **104**: 41-53.
2. Wertz, J.E., Goldstone, C., Gordon, D.M., Riley, M.A. A molecular phylogeny of enteric bacteria and implications for a bacterial species concept. *J. Evol. Biol.*, 2003; **16**: 1236-1248.
3. Clarridge III, J.E. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious disease. *Clin. Microbiol. Rev.*, 2004; **17**: 840-862.
4. Xue, S., Barna, M. Specialized ribosomes: a new frontier in gene regulation and organismal biology. *Nat. Rev. Mol. Cell Biol.*, 2012; **13**: 355-369.
5. Moazed, D., Noller, H.F. Binding of tRNA to the ribosomal A and P sites protects two distinct sets of nucleotides in 16S rRNA. *J. Mol. Biol.*, 1990; **211**: 135-145.
6. Basrai, M.A., Hieter, P., Boeke, J.D. Small open reading frames: beautiful needles in the haystack. *Genome Res.*, 1997; **7**: 768-771.
7. Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y. The complete genome sequence of *Escherichia coli* K-12. *Science*, 1997; **277**: 1453-1462.
8. Sopko, R., Andrews, B. Small open reading frames: not so small anymore. *Genome Res.*, 2006; **16**: 314-315.
9. Kumar, A., Harrison, P.M., Cheung, K.H., Lan,

N., Echols, N., Bertone, P., Miller, P., Gerstein, M.B., Snyder, M. An integrated approach for finding overlooked genes in yeast. *Nat. Biotechnol.*, 2002; **20**: 58-63.

10. Oshiro, G., Wodicka, L.M., Washburn, M.P., Yates III, J.R., Lockhart, D.J., Winzeler, E.A. Parallel identification of new genes in *Saccharomyces cerevisiae*. *Genome Res.*, 2002; **12**: 1210-1220.
11. Kessler, M.M., Zeng, Q., Hogan, S., Cook, R., Morales, A.J., Cottarel, G. Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome. *Genome Res.*, 2003; **13**: 264-271.
12. Hughes, D.T., Sperandio, V. Inter-kingdom signaling: communication between bacteria and their hosts. *Nat. Rev. Microbiol.*, 2008; **6**: 111-120.
13. Zakry, F.A.A., Halimi, M.S., Abdul-Rahim, K.B., Osumanu, H.A., Wong, S.K., Franklin, R.K., Stephen, L.C.T., Make, J. Isolation and plant growth-promoting properties of rhizobacterial diazotrophs from pepper vine *(Piper nigrum* L.). *Mal. Appl. Biol. J.*, 2010; **39**: 41-45.
14. Zakry, F.A.A., Halimi, M.S., Abdul-Rahim, K.B., Ahmed, O.H. Variable responses on early development of shallot (*Allium ascalonicum*) and mustard (*Brassica juncea*) plants to *Bacillus cereus* inoculation. *Mal. J. Microbiol.*, 2012; **8**: 47-50.
15. Mahasneh, I.A., Al-Thani, R.F., Abuannadi, M.A., Qandil, M.A. A complete sequence of the 16S rRNA gene of a novel *Streptomyces coelicolor* (AB588124)(QU66c-2002) isolated from the soil of Qatar. *Biotechnol.*, 2011; **10**: 167-174.
16. Green, C.J., Stewart, G.C., Hollis, M.A., Vold, B.S., Bott, K.F. Nucleotide sequence of the *Bacillus subtilis* ribosomal RNA operon, rrnB. *Gene*, 1985; **37**: 261-266.
17. Stelzl, U., Connell, S., Nierhaus, K.H., Wittmann-Liebold, B. Ribosomal proteins: role in ribosomal functions. In: Encyclopedia of Life Sciences, Nature Publishing Group, 2001.
18. Simonen, M., Palva, I. Protein secretion in *Bacillus* species. *Microbiol. Rev.*, 1993; **57**: 109-137.
19. Aguirre-Garrido, J.F., Montiel-Lugo, D., Hernández-Rodrwguez, C., Torres-Cortes, G., Millán, V., Toro, N., Martwnez-Abarca, F., Ramwrez-Saad, H.C. Bacterial community structure in the rhizosphere of three cactus species from semi-arid highlands in Central Mexico. *Antonie van Leeuwenhoek*, 2012; **101**: 891-904.
20. Cardoso, A.M., Cavalcante, J.J., Vieira, R.P., Lima, J.L., Grieco, M.A.B., Clementino, M.M., Martins, O.B. Gut bacterial communities in the giant land snail *Achatina fulica* and their modification by sugarcane-based diet. PLoS ONE, 2012; **7**: e33440.
21. Robbe, C., Capon, C., Maes, E., Rousset, M., Zweibaum, A., Zanetta, J.P., Michalski, J.C. Evidence of regio-specific glycosylation in human intestinal mucins presence of an acidic gradient along the intestinal tract. *J. Biol. Chem.*, 2003; **278**: 46337-46348.
22. Xiao, H., Chen, Y., Liu, Z., Huang, K., Li, W., Cui, X., Yi, L. Phylogenetic diversity of cultivable bacteria associated with a sea anemone from coast of the Naozhou island in Zhanjiang, *China. Acta Microbiologica Sinica*, 2009; **49**: 246.
23. Chen, Y.G., Li, H.M., Li, Q.Y., Chen, W., Cui, X.L. Phylogenetic diversity of culturable bacteria in the ancient salt deposits of the Yipinglang Salt Mine, PR China. *Acta Microbiologica Sinica*, 2007; **47**: 571.
24. Whitford, M.F., Forster, R.J., Beard, C.E., Gong, J., Teather, R.M. Phylogenetic analysis of rumen bacteria by comparative sequence analysis of cloned 16S rRNA genes. *Anaerobe*, 1998; **4**: 153–163.
25. Tajima, K., Aminov, R.I., Nagamine, T., Ogata, K., Nakamura, M., Matsui, H., Benno, Y. Rumen bacterial diversity as determined by sequence analysis of 16S rDNA libraries. *FEMS Microbiol. Ecol.*, 1999; **29**: 159–169.
26. Leser, T.D., Amenuvor, J.Z., Jensen, T.K., Lindecrona, R.H., Boye, M., Moller, K. Culture-independent analysis of gut bacteria: the pig gastrointestinal tract microbiota revisited. *Appl. Environ. Microbiol.*, 2002; **68**: 673–690.
27. Suau, A., Bonnet, R., Sutren, M., Godon, J.J., Gibson, G.R., Collins, M.D., Dore, J. Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl. Environ. Microbiol.* 1999; **65**: 4799–4807.
28. Vollmer, W. Bacterial outer membrane evolution via sporulation. *Nat. Chem. Biol.*, 2011; **8**: 14-18.
29. Errington, J. Regulation of endospore formation in *Bacillus subtilis*. *Nat. Rev. Microbiol.*, 2003; **1**: 117-126.
30. Federle, M.J., Bassler, B.L. Interspecies communication in bacteria. *J. Clin. Invest.*, 2003; **112**: 1291-1299.
31. Lambrecht, M., Okon, Y., Broek, A.V., Vanderleyden, J. Indole-3-acetic acid: a reciprocal signaling molecule in bacteria-plant interactions. *Trends Microbiol.*, 2000; **8**: 298-

300.
32. Nishito, Y., Osana, Y., Hachiya, T., Popendorf, K., Toyoda, A., Fujiyama, A., Itaya M., Sakakibara, Y. Whole genome assembly of a natto production strain *B. subtilis* natto from very short read data. *BMC Genomics*, 2010; **11**: 243.
33. Liu, J., He, D., Li, X.Z., Gao, S., Wu, H., Liu, W., Gao, X., Zhou, T. ã-Polyglutamic acid (γ-PGA) produced by *Bacillus amyloliquefaciens* C06 promoting its colonization on fruit surface. *Intl. J. Food Microbiol.*, 2010; **142**: 190-197.
34. Wei, X., Ji, Z., Chen, S. Isolation of halotolerant *Bacillus licheniformis* WX-02 and regulatory effects of sodium chloride on yield and molecular sizes of poly-γ-glutamic acid. *Appl. Biochem. Biotechnol.*, 2010; **160**: 1332-1340.
35. Yangtse, W., Zhou, Y., Lei, Y., Qiu, Y., Wei, X., Ji, Z., Qi, G., Yong, Y., Chen, L., Chen, S. Genome sequence of *Bacillus licheniformis* WX-02. *J. Bacteriol.*, 2012; **194**: 3561-3562.
36. Oh, S.Y., Budzik, J.M., Garufi, G., Schneewind, O. Two capsular polysaccharides enable *Bacillus cereus* G9241 to cause anthrax-like disease. *Mol. Microbiol.*, 2011; **80**: 455-470.
37. Urushibata, Y., Tokuyama, S., Tahara, Y. Characterization of the *Bacillus subtilis* ywsC gene, involved in c-polyglutamic aicd production. *J. Bacteriol.*, 2002; **184**: 337-343.
38. Candela, T., Mock, M., Fouet, A. CapE, a 47-amino-acid peptide, is necessary for *Bacillus anthracis* polyglutamate capsule synthesis. *J. Bacteriol.*, 2005; **187**: 7765-7772.
39. Candela, T., Fouet, A. Poly-γ-glutamate in bacteria. *Mol. Microbiol.*, 2006; **60**: 1091-1098.
40. Ho, G.-H., Ho, T.-I., Hsieh, K.-H., Su, Y.-C., Lin, P.-Y., Yang, J., Yang, K.-H., Yang, S.-C. γ-polyglutamic acid produced by *Bacillus subtilis* (natto): structural characteristics, chemical properties and biological functionalities. *J. Chinese Society*, 2006; **53**: 1363-1384.
41. Buescher, J.M., Margaritis, A. Microbial biosynthesis of polyglutamic acid biopolymer and applications in the biopharmaceutical, biomedical and food industries. *Crit. Rev. Biotechnol.*, 2007; **27**: 1-19.
42. Sutcliffe, I.C., Russell, R.R.B. Lipoproteins of Gram-positive bacteria. *J. Bacteriol.*, 1995; **177**: 1123-1128.
43. Friedman, D.I., Schauer, A.T., Baumann, M.R., Baron, L.S., Adhya, S.L. Evidence that ribosomal protein S10 participates in control of transcription termination. *Proc. Natl. Acad. Sci. USA*, 1981; **78**: 1115-1118.
44. Woolstenhulme, C.J., Hill, W.E. The genesis of ribosome structure: how a protein generates RNA structure in real time. *J. Mol. Biol.*, 2009; **392**, 645-656.
45. Bavykin, S.G., Lysov, Y.P., Zakhariev, V., Kelly, J.J., Jackman, J., Stahl, D.A., Cherni, A. Use of 16S rRNA, 23S rRNA, and gyrB gene sequence analysis to determine phylogenetic relationships of *Bacillus cereus* group microorganisms. *J. Clin. Microbiol.*, 2004; **42**: 3711-3730.
46. Chen, M.L., Tsen, H.Y. Discrimination of *Bacillus cereus* and *Bacillus thuringiensis* with 16S rRNA and gyrB gene based PCR primers and sequencing of their annealing sites. *J. Appl. Microbiol.*, 2002; **92**: 912-919.
47. Eribe, E.R.K., Olsen, I. Leptotrichia species in human infections. *Anaerobe*, 2008; **14**: 131-137.
48. Werner, G., Gfrörer, S., Fleige, C., Witte, W., Klare, I. Tigecycline-resistant *Enterococcus faecalis* strain isolated from a German intensive care unit patient. *J. Antimicrob. Chemoth.*, 2008; **61**: 1182-1183.
49. Baums, C.G., Valentin-Weigand, P. Surface-associated and secreted factors of *Streptococcus suis* in epidemiology, pathogenesis and vaccine development. *Anim. Health Res. Rev.*, 2009; **10**: 65-83.
50. Manso-Silvan, L., Vilei, E.M., Sachse, K., Djordjevic, S.P., Thiaucourt, F., Frey, J. Mycoplasma leachii sp. nov. as a new species designation for *Mycoplasma* sp. bovine group 7 of Leach, and reclassification of *Mycoplasma mycoides* subsp. *mycoides* LC as a serovar of *Mycoplasma mycoides* subsp. *capri. Intl. J. Syst. Evol. Microbiol.*, 2009; **59**: 1353-1358.
51. Steeves, C.H., Potrykus, J., Barnett, D.A., Bearne, S.L. Oxidative stress response in the opportunistic oral pathogen *Fusobacterium nucleatum*. *Proteomics*, 2011; **11**: 2027-2037.
52. Grondin, C., Imbert, P., Ficko, C., Mérens, A., Dutasta, F., Bigaillon, C., Rapp, C. Shigella flexneri bacteremia in two immune competent adult travelers. *J. Travel Med.*, 2012; **19**: 258-260.
53. Berg, G., Eberl, L., Hartmann, A. The rhizosphere as a reservoir for opportunistic human pathogenic bacteria. *Environ. Microbiol.*, 2005; **71**: 4203–4213.
54. Berg, G., Zachow, C., Cardinale, M., Müller, H. Ecology and human pathogenicity of plant-associated bacteria. *Regulation of Biological Control Agents*, 2011: 175-189.
55. Martinez, J.L. The role of natural environments in the evolution of resistance traits in pathogenic bacteria. *Proc. Royal Soc. B: Biol. Sci.*, 2009: **276**, 2521-2530.