IOSR Journal of Biotechnology and Biochemistry (IOSR-JBB)
ISSN: 2455-264X, Volume 7, Issue 1 (Jan. – Feb. 2021), PP 46-50
www.iosrjournals.org

Characterization of a Hypothetical Protein from *Bacillus cereus*: an *In Silico* Approach

Md. Fazley Rabbi[1]*, Aradhan Sarkar[1]
[1]*Department of Biotechnology and Genetic Engineering, Noakhali Science and Technology University, Noakhali-3814, Bangladesh.*
**Corresponding author*

Abstract
Background: *Bacillus cereus is a Gram positive, rod-shaped bacterium commonly found in soil and foods. The bacterial strain G1-1 contains many hypothetical proteins yet to be annotated. Annotation of these proteins might reveal new insights about this pathogenic organism; hence a hypothetical protein WP_000822694.1 was selected in this study for comprehensive characterization through in silico approach.*
Materials and Methods: *Various bioinformatics resources were used in this study for characterization of the target protein. Physicochemical properties were estimated using ProtParam tool. CELLO and PSLpred server was used for subcellular localization prediction. Functional analysis was performed using NCBI-CDD and Pfam server. Secondary and tertiary structure (3D) of the protein was determined by SOPMA and MODELLER server respectively. Moreover, quality of the 3D structure was validated by several structure assessment tools. Finally, active site of the protein was analyzed by CASTp server.*
Results: *The 172 amino acid containing protein was found as a stable protein. The protein was predicted to be a toxin CdiA protein of RNAse superfamily used in interbacterial competition. 3D structure of the protein was successfully determined which passed all the quality assessment tools. Active site analysis revealed several key interacting residues.*
Conclusion: *The protein was found to be a toxin protein used for growth inhibition of neighboring bacterial cells. Further experimentations are needed to validate our findings.*
Keywords: *Characterization, Hypothetical protein, Bacillus cereus, Functional analysis*

Date of Submission: 28-01-2021 Date of Acceptance: 11-02-2021

I. Introduction

Bacillus cereus is a Gram-positive, rod-shaped, facultative anaerobic, spore forming bacterium commonly found in soil, on vegetables, and in many raw and processed foods [1]. Food poisoning by this bacterium may occur when foods are prepared and kept without adequate refrigeration for several hours before serving [2,3]. Other than food poisoning, *B. cereus* induces local and systemic infections including septicemia, pneumonia, endocarditis, meningititis and encephalitis, especially in immunosuppressed individuals [4-8]. There are two main types of intestinal illnesses caused by *B. cereus*: one is diarrheal and another one leads more to nausea/vomiting. The pathogenicity of *B. cereus*, whether intestinal or nonintestinal, is intimately associated with the production of tissue-destructive exoenzymes [9].

The bacterial genome contains many hypothetical proteins (HPs) whose functions are still unexplored. A hypothetical protein is a protein whose existence has been predicted but in vivo function has not been determined [10]. Characterization of HPs of a bacterial genome may give insights about new domains and motifs, pathways and protein networks etc [11,12]. 1,150 genomes of *Bacillus cereus* are available in NCBI database (http://www.ncbi.nlm.nih.gov/) [13]. Among them, strain G1-1 was isolated from ocean sediment on October 3, 2018 and submitted by Qingdao University, China. The total length of the genome is 5.86 Mb with a total protein count of 5,725 and GC content of 35.19%. Among 5,725 proteins, 890 proteins are hypothetical. Annotation of these proteins is important in order to understand its pathogenicity and discover potential pharmacological targets. We selected the hypothetical protein WP_000822694.1 of *B. cereus* G1-1 strain for computational characterization in this study.

II. Materials and Methods

Selection of hypothetical protein: We browsed NCBI database and searched for hypothetical proteins of *B. cereus* G1-1 strain. Among the hypothetical proteins, a 172 amino acid containing protein WP_000822694.1 was selected and its primary sequence in FASTA format was retrieved for further analysis.

Determination of physicochemical properties: Various physicochemical properties including molecular weight, aliphatic index (AI), isoelectric point (pI), molecular formulae, extinction coefficients, GRAVY (grand average of hydropathy) etc. of the target protein were determined using ProtParam (http://web.expasy.org/protparam/) [14] tool of ExPASy.
Prediction of subcellular localization: Subcellular location of the hypothetical protein was determined using CELLO (http://cello.life.nctu.edu.tw/) [15] and PSLpred [16] tool.
Function determination of the hypothetical protein: Function of the hypothetical protein was determined based on domain and motif analysis. The tools used for this purpose included NCBI CDD [17] and Pfam [18]. Protein sequence in FASTA format was given as input to identify conserved domain and motif.
Secondary structure determination: Secondary structure of the protein was determined using SOPMA[19] server. Determination of the secondary structure can give us insight about the three dimensional structure of a protein.
Three dimensional structure determination: Homology modeling method was used to determine three dimensional (3D) structure of the hypothetical protein. We used MODELLER through HHpred tools [20] of the Max Planck Institute for Developmental Biology to develop 3D structure. Among the template proteins, 5E3E_F (PDB ID: 5E3E) was selected to initiate modeling which is an X-ray diffraction model of a toxin protein of *Yersinia kristensenii*.
Quality assessment of the 3D structure: The quality of the 3D structure was assessed by several quality assessment tools including PROCHECK [21], Verify3D [22] through SAVES (https://saves.mbi.ucla.edu/) server and QMEAN [23] program. The Z value for the template and target protein was determined by ProSA-web server [24].
Active site determination: To identify the active site of the protein, computed atlas of surface topography of proteins (CASTp) (http://sts.bioe.uic.edu/castp/) server [25] was used. The PDB file of the developed 3D structure was uploaded as input. The best active site and interacting amino acids of the target protein were determined.

III. Results

Physicochemical properties: Many physical and chemical properties of the hypothetical protein WP_000822694.1 were determined by ProtParam tool. The target protein contains 172 amino acid residues and has molecular weight of 19184.03. Instability index (II) of the protein was found to be 30.18 indicating it as a stable protein. Detailed characteristics are summarized in Table 1.

Table 1: Physicochemical properties of the hypothetical protein estimated by ProtParam tool

Properties	Value
No. of amino acids	172
Molecular weight	19184.03
Theoretical pI	9.44
Negatively charged residues (Asp + Glu)	23
Positively charged residues (Arg + Lys)	31
Extinction coefficients (M-1 cm-1)	9970
Estimated half-life (in vitro)	30 hours
Instability index (II)	30.18
Aliphatic index	80.47
Grand average of hydropathicity (GRAVY)	-0.653

Subcellular localization: Cello server predicted the protein to be a membrane and extracellular protein with reliability score of 2.470 and 2.102 respectively whereas PSLpred server predicted the protein as extracellular with reliability index of 1.

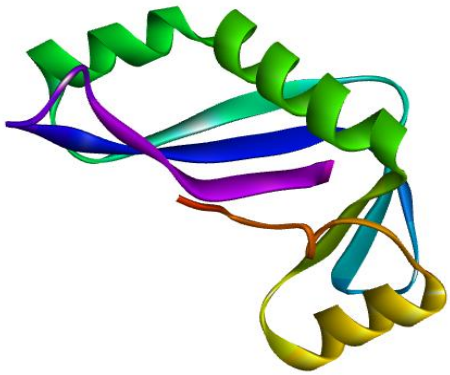

Figure 1: Three dimensional (3D) structure of the target protein.

Functional analysis: The domain of the target protein was predicted using NCBI CDD and Pfam server. NCBI CDD predicted the target protein to contain bacterial CdiA-CT RNAse domain (C terminal toxin domain of contact dependent growth inhibition, Cdi protein of RNAse superfamily) at 58-169 amino acid residues with an E-value (expected value) of 7.27e-39. Pfam server predicted the same domain at same position with an E-value (expected value) of 1.9e-29. The result was consistent in both platforms indicating the protein to be a toxin CdiA protein of RNAse superfamily.

Table 2: Ramachandran plot statistics of the target protein

Ramachandran Plot Statistics	Number of a.a residues	Percentage (%)
Residues in most favored regions [A, B, L]	100	94.3%
Residues in additional allowed regions [a, b, l, p]	4	3.8%
Residues in generously allowed regions [~a, ~b, ~l, ~p]	1	0.9%
Residues in disallowed regions	1	0.9%
Number of non-glycine and non-proline residues	106	100%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	5	
Number of proline residues	2	
Total number of residues	115	

Structure determination: According to SOPMA prediction, alpha helix predominates in secondary structure (44.19%) followed by random coil (37.21%). Other types of secondary structure included extended strand (12.79%) and beta turn (5.81%). 3D structure was determined by MODELLER using the template protein 5E3E_F (PDB ID: 5E3E). The developed structure is illustrated in Figure 1.

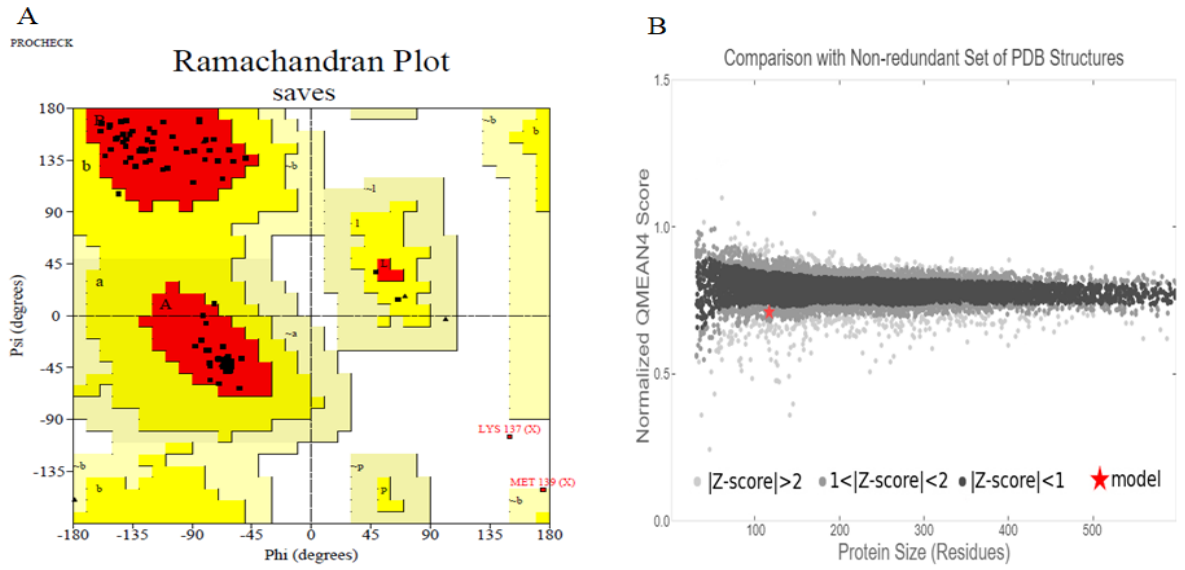

Figure 2: Quality assessment of the developed 3D structure. (A) Ramachandran plot of the target protein. (B) Graphical output of QMEAN result.

Quality assessment result: According to PROCHECK result, 94.3% amino acid residues reside within the most favored region in 'Ramachandran plot' indicating a good quality model (Table 2 and Figure 2A). The 3D structure successfully passed the Verify 3D server where 81.74% of the residues have averaged 3D-1D score >= 0.2. QMEAN4 value determined through QMEAN program was -1.39 which is also considered as good (Figure 2B). According to PROSA server, the Z score of the target and template protein was -5.98 and -5.07 respectively (Figure 3). This implies possible homology between the template and the target protein.

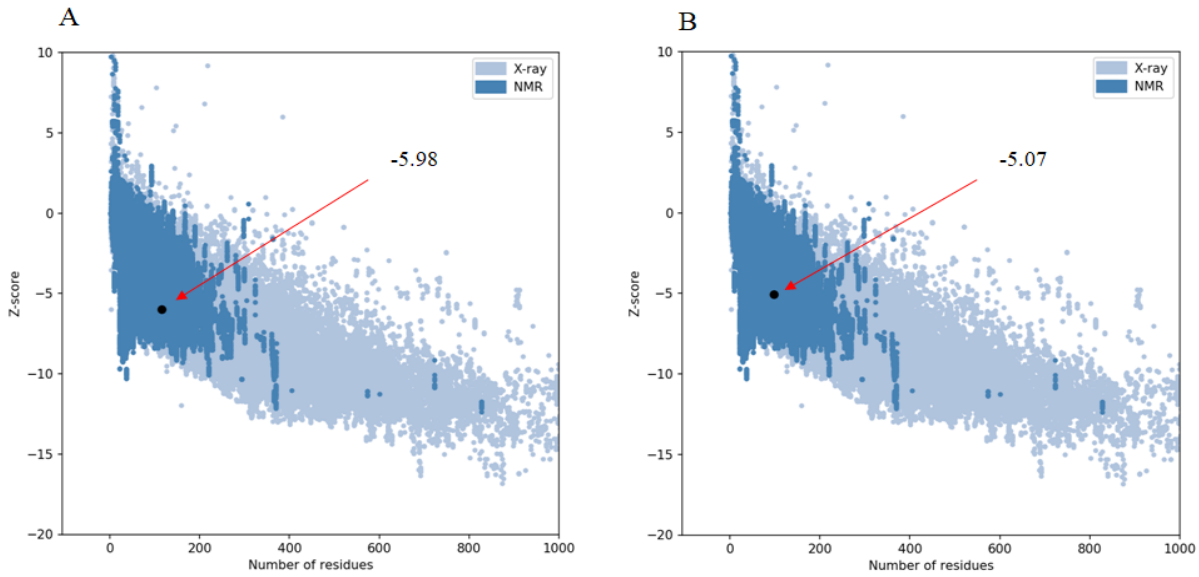

Figure 3: Z scores of the proteins determined by ProSA server (A: target protein, B: template protein).

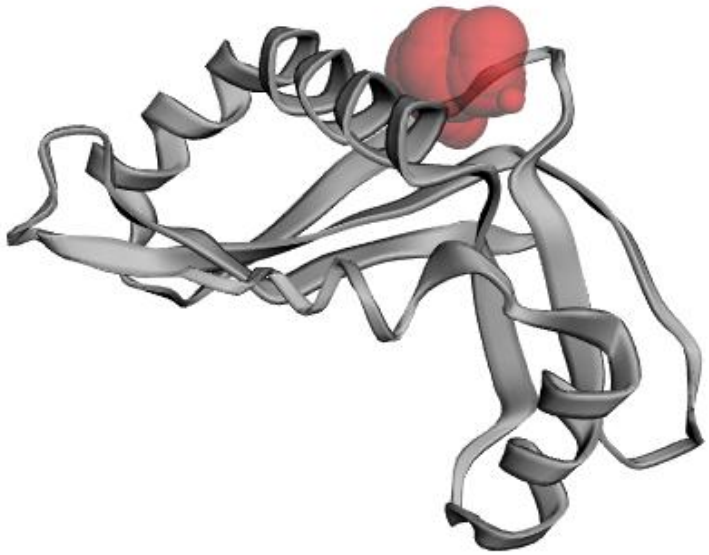

Figure 4: Active site of the protein predicted by CASTp server (red sphere indicates best active site).

Active site analysis: The active site of the protein was analyzed using CASTp server (Figure 4). Amino acids of the best active site was found to be Thr[92], Lys[95], Ala[96], Asp[99], Thr[124], Thr[125], His[126], Phe[128], Pro[129], Val[130], Val[146], Thr[148] with solvent accessible (SA) area of 103.495 and volume of 52.097.

IV. Discussion

Rapid, low-cost sequencing technologies have generated vast amount of data which requires annotation. Various bioinformatics resources have been used to annotate the proteins from various pathogenic microorganisms. In this study, a hypothetical protein WP_000822694.1 from *Bacillus cereus* G1-1 strain was selected for comprehensive characterization. Physicochemical properties and subcellular localization of the protein was determined (Table 1). Functional analysis revealed the protein to be a toxin CdiA protein of RNAse A superfamily. The protein is used by many bacteria as a mechanism in interbacterial competition. The bacteria utilize this protein to inhibit growth of neighboring bacterial cells by contact dependent inhibition mechanism [26]. Secondary and tertiary structure of the protein was determined by SOPMA and MODELLER tool respectively. Alpha helix was found to be the prevalent type in secondary structure. 3D structure of the target protein (Figure 1) was determined based on homology modeling method. Interestingly, the 3D structure passed all the quality assessment tools. Finally, the active site and interacting amino acid residues of the protein was determined.

Further research on this important protein might reveal new information about bacterial pathogenicity and interbacterial interaction.

V. Conclusion

The hypothetical protein was successfully characterized from both structural and functional aspects. The protein was found to be a key protein of *Bacillus cereus* used in interbacterial competition.

References

[1]. Nguyen AT, Tallent SM. Screening food for Bacillus cereus toxins using whole genome sequencing. Food Microbiol. 2019;78:164-170.
[2]. Asaeda G, Caicedow G, Swanson C. Fried rice syndrome. Jems. 2005;30(12):30-32.
[3]. Glasset B HS, Guiller L, Cadel-Six S, Vignaud ML, Grout J, et al. Large-scale survey of Bacillus cereus-induced food-borne outbreaks: epidemiologic and genetic characterization. EuroSurveillance. 2016;21(48).
[4]. Veysseyre F, Fourcade C, Lavigne JP, Sotto A. Bacillus cereus infection: 57 case patients and a literature review. Med Mal Infect. 2015;45(11-12):436-440.
[5]. Frankard J, Li R, Taccone F, Struelens MJ, Jacobs F, Kentos A. Bacillus cereus pneumonia in a patient with acute lymphoblastic leukemia. Eur J Clin Microbiol Infect Dis. 2004;23(9):725-728.
[6]. Gaur AH, Patrick CC, McCullers JA, et al. Bacillus cereus bacteremia and meningitis in immunocompromised children. Clin Infect Dis. 2001;32(10):1456-1462.
[7]. Arnaout MK, Tamburro RF, Bodner SM, et al. Bacillus cereus causing fulminant sepsis and hemolysis in two patients with acute leukemia. J Pediatr Hematol Oncol. 1999;21(5):431-435.
[8]. Ramarao N, Belotti L, Deboscker S, et al. Two unrelated episodes of Bacillus cereus bacteremia in a neonatal intensive care unit. Am J Infect Control. 2014;42(6):694-695.
[9]. Granum PE. Bacillus cereus and its toxins. J Appl Bacteriol Symp. 1994;Suppl.76:615-665.
[10]. Lubec G, Afjehi-Sadat L, Yang JW, John JP. Searching for hypothetical proteins: theory and practice based upon original data and literature. Prog Neurobiol. 2005;77(1-2):90-127.
[11]. Nimrod G, Schushan M, Steinberg DM, Ben-Tal N. Detection of functionally important regions in "hypothetical proteins" of known structure. Structure. 2008;16(12):1755-1763.
[12]. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999;96(8):4285-4288.
[13]. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Res. 2002;30(1):17-20.
[14]. Gasteiger E. HC, Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A. Protein Identification and Analysis Tools on the ExPASy Server. (In) John M Walker (ed): The Proteomics Protocols Handbook, Humana Press. 2005:pp. 571-607.
[15]. Yu CS, Chen YC, Lu CH, Hwang JK. Prediction of protein subcellular localization. Proteins. 2006;64(3):643-651.
[16]. Bhasin M, Garg A, Raghava GP. PSLpred: prediction of subcellular localization of bacterial proteins. Bioinformatics. 2005;21(10):2522-2524.
[17]. Lu S, Wang J, Chitsaz F, et al. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 2020;48(D1):D265-d268.
[18]. El-Gebali S, Mistry J, Bateman A, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47(D1):D427-d432.
[19]. Combet C, Blanchet C, Geourjon C, Deléage G. NPS@: network protein sequence analysis. Trends Biochem Sci. 2000;25(3):147-150.
[20]. Zimmermann L, Stephens A, Nam SZ, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol. 2018;430(15):2237-2243.
[21]. Laskowski R A MMW, Moss D S, Thornton J M. PROCHECK - a program to check the stereochemical quality of protein structures. J App Cryst. 1993;26: 283-291.
[22]. Eisenberg D, Lüthy R, Bowie JU. VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol. 1997;277:396-404.
[23]. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics. 2011;27(3):343-350.
[24]. Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 2007;35(Web Server issue):W407-410.
[25]. Tian W, Chen C, Lei X, Zhao J, Liang J. CASTp 3.0: computed atlas of surface topography of proteins. Nucleic Acids Res. 2018;46(W1):W363-w367.
[26]. Batot G, Michalska K, Ekberg G, et al. The CDI toxin of Yersinia kristensenii is a novel bacterial member of the RNase A superfamily. Nucleic Acids Res. 2017;45(9):5013-5025.

Md. Fazley Rabbi, et. al. "Characterization of a Hypothetical Protein from *Bacillus cereus*: an *In Silico* Approach." *IOSR Journal of Biotechnology and Biochemistry (IOSR-JBB)*, 7(1), (2021): pp. 46-50.