BIOINFORMATICS BASED INDICATION OF *BACILLUS LENTUS* NCTC4824EXPRESSINGSIRTUIN-LIKE PROTEIN

Beenish Rafique, Muhammad Sufian* and Raja TahirMahmood

Department of Biotechnology, Mirpur University of Science and Technology (MUST), Mirpur-10250, AJK, Pakistan

*Correspondence :msufian@must.edu.pk

ABSTRACT

Bacillus species as part of probiotics of the so-called "live drugs" and food additives are approved for human and animal consumption. *Bacillus* genus members are commonly employed in industry for large-scale enzyme synthesis.*Bacillus lentus* is aerobic, Gram positive bacteria. *B. lentus* has two industrially significant strains: NCTC 4824 and NBRC 16444. It can be used in a variety of biotechnology sectors, including detergent and food manufacturing. It is utilized in industry to make different enzymes, like proteases and peptidases. Household detergents are the furthermost well-known use of these alkaline proteases. In modern biology and medicine, bioinformatics is critical for data management.Hypothetical proteins of bacteria present in NCBI Protein and UniProt databases are identified and characterized by bioinformatics. In biochemical and physiological pathways, hypothetical proteinsmay play a significant role.In our study, NCBI Protein database was used to download the complete genome of *B. lentus* NCTC4824 in FASTA format. After retrieving hypothetical proteins from the NCTC4824 genome,NCBI Protein BLAST was performed for sequence similarity search against locally-designed BIPs dataset. The obtained best hithaving highest percent identity, and query coverage, and e-value was found to NAD-dependent protein deacetylase of another specie of *Bacillus*. This enzyme has a significant importance in biotechnology.Sirtuin Sir2 (deacetylase enzyme for NAD+ dependent acetate consumption) influences cell ageing, chromosomal stability, energy metabolism in response to dietary cues, gene silencing, and stress response.The findings have elevated the importance of an industrially significant bacterial strain i.e. *B. lentus* NCTC4824.

Keywords: *Bacillus lentus*, Sirtuin, hypothetical proteins, gene silencing

INTRODUCTION

Species of *Bacillus* as part of probiotics of the so-called "live drugs" and food additives approved for human and animal consumption include only certain types of bacilli *B. subtilis, B. licheniformis, B. coagulans, B. toyoi (cereus), B. clausii (lentus),* and *B. polyfermentans.* The bacteria *B. lentus*are part of the non-pathogenic transit microflora of the intestines of humans and animals, but have their own characteristic difference. *B. lentus* are known to be alkaline tolerant and are known producers of highly alkaline protease. Different strains of *B. lentus* are ATCC 10840, NG121 and NCTC 4824 (Sharma *et al*., 2006; Xu and Jean, 2003).

B. lentus is aerobic, gram positive bacteria that produce endospores. It is found in soil. Considered non-pathogenic to human. *B. lentus* grow on solid media (e.g. nutrient agar) at 37°C, within 24-48 hours. *B. lentus* also grow on liquid media (e.g. nutrient broth) at 37°C.The complete genome assembly of *B. lentus* NCTC4824 is sequenced. The accession number of bacterium is LS483476.1. Total sequence length is 4,384,366 base pairs (bp.) and total un-gapped length is 4,384,366 bp. Total number of chromosomes and plasmids are 1.

Bacillus genus members are commonly employed in industry for large-scale enzyme synthesis. *B. lentus*is utilised in industry to make a variety of enzymes, including proteases and peptidases (Jorgensen *et al*., 2000).One of these industrial organisms is the alkalophilic*B. lentus*, which produces BLAP(*B. Lentus* Alkaline Peptidases), a commercial peptidase preparation. Nowadays, one of the furthermost well-known applications of BLAP is in household detergent (Jorgensen *et al*., 2000).It was observed that BLAP is inhibited by its products in a competitive manner, as a result, it was used to hydrolyze casein in an enzyme membrane reactor (EMR) (Eisele *et al*., 2013).Proteases are found in many industrial processes and products.The most major bacterial supplier of proteases is *Bacillus*, providing large quantities of proteolytic enzymes that are neutral and alkaline with unique features for example, excellent stability in the presence of severe temperatures, pH, organic solvents, detergents, and oxidising chemicals. As a result, numerous ways for producing *Bacillus* proteases at a low cost have been devised, including optimising fermenter conditions. Protease enzymes are employed in many different applications, such as detergent, food processing, and skin dehairing. From various *Bacillus* species, many commercial proteases have been described, purified, and produced (Contesini *et al*., 2018).

The use of computation and analysis tools to record and understand biological data is known as bioinformatics.Bioinformatics is necessary for data management in modern biology and

medicine.Software programs for computers,the bioinformatics toolkit, like BLAST and Ensembl, which require internet connectivity. Bioinformatics is becoming more widely used for a variety of other crucial functions, such as analysinggene expression and variation, analysing and predicting structure and function of genes and proteins, as well as gene regulator networks prediction and detection, in addition to genome sequence data processing (Bayat, 2002).

Sequence analysis is a wide field with numerous sub-domains. Sequence alignment can disclose crucial information about the structural and functional areas within a sequence. It is used to investigate the evolutionary course of sequences by finding orthologs and homologs of the sequences. Statistical approaches are used to generate sequence profiles and more precisely identify distantly related sequences. The advancement of sequencing technologies has ushered in the next-generation era, which has allowed for personalised medicine and the finding of haplotypes and quasi-species (Speed, 2003).

The field of molecular biology known as functional genomics explores the roles and interactions of genes (and proteins).The field of functional genomics focuses on the dynamic aspects of genomic data, such astranscription of genes, and translation, gene control, and significant interactions between proteins, rather than static aspects, such as the sequence or the structures of DNA. Characteristics of the genome that are connected to its function, such as polymorphism and mutation (for example, single nucleotide polymorphism (SNP) analysis), in addition to measuring molecular activity, all of this is a part of functional genomics. Transcriptomics (gene expression), proteomics (protein production), and metabolomics(metabolite production) are some of the "-omics" that are included in this category (Mlecnik *et al*., 2018).

Hypothetic proteins (HPs) are proteins that are anticipated to be expressed using an open reading frame but without any experimental evidence of translation. Only about 2% of the genetic material in the genome proteins are coded, the others are either non-coding or have not yet been detected.(Ijaq *et al*., 2019). Because these proteins demonstrate weak link to known annotated proteins, they are referred to Putative Conserved Proteins (PCPs) or Hypothetical Proteins (HPs) (Shahbaaz *et al*., 2013). Despite the fact that they lack functional characterization, they serve a critical role in comprehending biochemical and physiological mechanisms, such as discovering novel structures and functions, markers and pharmaceutical

targets, as well as early identification and advantages for proteomic and genomic studies (Ijaq *et al*., 2019).

Proteins play a role in every biological process that takes place in living organisms. Protein interacts with a wide range of substances. Enzyme, a protein that catalyses chemical reactions, is the most well-known example of protein function. (Ismaya and Wangsa 2011). Microbial enzymes have attracted interest for their broad use in industries and medicine due to their stability, catalytic activity, and ease of manufacturing and optimization.

Enzymes use in a variety of industries (such as food, agriculture, chemicals, and medicine) is rapidly increasing due to their shortened processing time, minimal energy input, cost effectiveness, non-toxic, and environmentally benign features. Chemicals contained in industrial and household trash can be degraded or converted by microbial enzymes (phenolic compounds, nitriles, and amines, etc) (Singh *et al*, 2016).

MATERIALS AND METHODS

Working System

All computational experiments were carried out using Google Chrome version 87.0.4278.0 (Official Build) dev (64-bit) on a 64-bit processor Intel(R) Core(TM) m3-7Y30 CPU @ 1.00GHz 1.61GHz and 8.00 GB RAM on a Windows 10, version 2004 (OS Build 19041.508) operating system.

Literature Mining

In order to carry out my research work, firstly I choose bacterial specie after comprehensive study through research articles, a bacterium named *B. lentus* that is rod shaped and gram-positive. It has two strains NCTC 4824 and NBRC 16444. I selected it because of its interesting features like it is amongst the biggest known bacteria. Currently, it is used as industrial organism to produce a variety of proteins like proteases and peptidases. It has a wide range of uses in biotech sectors, including detergent and food processing. I carried out my research work with *B. lentus* NCTC 4824 strain.

Retrieval of Genomic Datasets

The complete proteome of *B. lentus* NCTC 4824 under the accession number LS483476.1was retrieved from the NCBI Protein Database (https://www.ncbi.nlm.nih.gov/protein/) (Benson *et*

al., 2016). Hypothetical proteinswere extracted from complete proteome in FASTA format using FaBox tool (https://users-birc.au.dk/~palle/php/fabox/index.php) (Villesen, 2007).

Dataset of Biotechnological Important Proteins

Dataset of biotechnological important proteins (BIPs) was made through an extensive study of literature of industrially significant proteins. The protein sequences were retrieved in FASTA format from NCBI Protein database(Benson *et al*., 2016). Proteins datasetwas compiled along with references, industrial uses, and proteomic sequences.

Sequence Similarity Search

For sequence similarity search perform NCBI Protein BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) of two datasets in which Query Sequence is Hypothetical proteins of *B. lentus* and Subject Sequence is BIPs dataset of 100 proteins. Search for the best hits using three values: percent identity (>70%), e-value (near to 0) and query coverage (near to 100%). Also the query hypothetical protein should be of large size as to avoid partial submissions in NCBI Protein database.

Validation Experiments

The obtained best hit from BIPs was searched in the NR database using NCBI Protein BLAST in which Query Sequence is accession number of best hitagainst NCBI NR database. This is done to check the consensus of naming and similarity with other proteins.

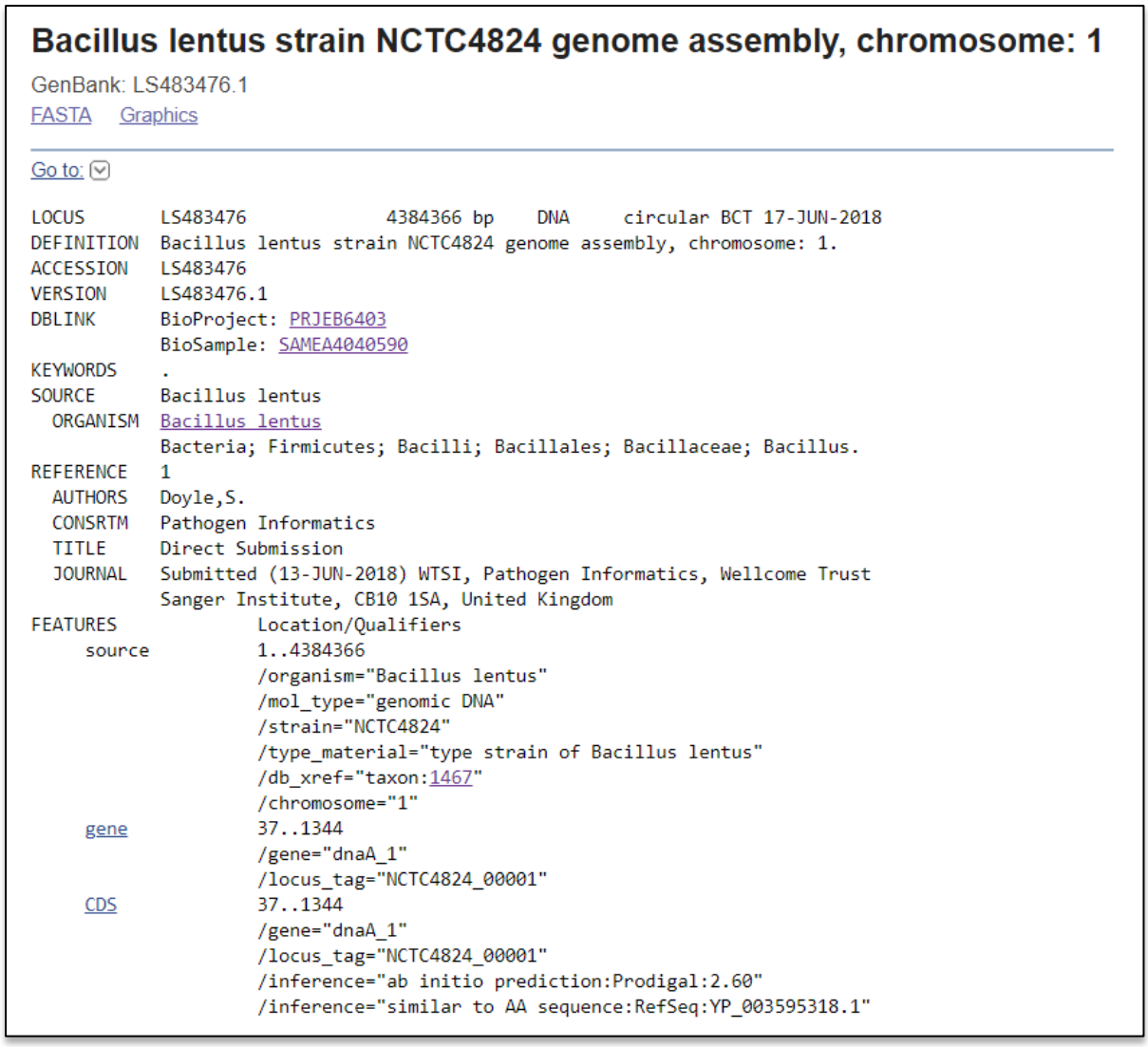
RESULTS AND DISCUSSION

Literature Mining

The genome size of*B. lentus*NCTC 4824 strain is 4,383,366 bp. It was retrieved from NCBI Genome database (https://www.ncbi.nlm.nih.gov/genome) under accession number LS483476.1 (Fig.1) (Kitts *et al*. 2016).

Retrieval of Genomic Dataset

The complete proteome of *B. lentus* NCTC4824 contains 4,072 proteins out of which 483 were hypothetical proteins, which were extracted from complete proteome in FASTA format using FaBox tool (Villesen, 2007).



Bacillus lentus strain NCTC4824 genome assembly, chromosome: 1

GenBank: LS483476.1

FASTA Graphics

Go to:

```
LOCUS       LS483476             4384366 bp    DNA     circular BCT 17-JUN-2018
DEFINITION  Bacillus lentus strain NCTC4824 genome assembly, chromosome: 1.
ACCESSION   LS483476
VERSION     LS483476.1
DBLINK      BioProject: PRJEB6403
            BioSample: SAMEA4040590
KEYWORDS    .
SOURCE      Bacillus lentus
  ORGANISM  Bacillus lentus
            Bacteria; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus.
REFERENCE   1
  AUTHORS   Doyle,S.
  CONSRTM   Pathogen Informatics
  TITLE     Direct Submission
  JOURNAL   Submitted (13-JUN-2018) WTSI, Pathogen Informatics, Wellcome Trust
            Sanger Institute, CB10 1SA, United Kingdom
FEATURES             Location/Qualifiers
     source          1..4384366
                     /organism="Bacillus lentus"
                     /mol_type="genomic DNA"
                     /strain="NCTC4824"
                     /type_material="type strain of Bacillus lentus"
                     /db_xref="taxon:1467"
                     /chromosome="1"
     gene            37..1344
                     /gene="dnaA_1"
                     /locus_tag="NCTC4824_00001"
     CDS             37..1344
                     /gene="dnaA_1"
                     /locus_tag="NCTC4824_00001"
                     /inference="ab initio prediction:Prodigal:2.60"
                     /inference="similar to AA sequence:RefSeq:YP_003595318.1"
```

Figure 1: Complete Genome of *B. lentus* NCTC4824

Sequence Similarity Search

For sequence similarity search performed NCBI Protein BLAST of two datasets in which Query Sequence is Hypothetical proteins of *B. lentus*NCTC4824and Subject Sequence is BIPs dataset of 100 proteins. The obtained best hit is NAD-dependent protein deacetylase of *B.onubensis*. The NAD-dependent protein deacetylase of *B. onubensis* have percent identity is 85.37%, e-value is 0.0 and query coverage is 100% (Fig. 2). Results with good score were checked for pairwise alignments using dot representation for identities (Fig. 3).

Hypothetical Protein

- Query Accession No: LS483476.1
- Query Name: Hypothetical Protein

- Query Organism: *B. lentus* strain NCTC4824

>gi|1054047867|ref|WP_066139086.1| hypothetical protein [Bacillus lentus]
MLTQQYQNIIGTILEKIEEADAIVVGGAAGMSAAGGYNWYLDDENFREHFNVFAEKYGIDSIFGGFYYKF
RTEEERWAYLATLINFVTEVPIGQPYKDLYEIIKDKNYYILTTNQDTQFLQVFPEEKVSAIQGNWTYLQC
SGPCHDGIYPYAEQAKELCAHIDGTKIPSDMVPKCPECGGPMELWVRSFVFLEGEKYRDEHHKYRTFLLE
NQNKKILFLELGVGQMTPMFIKEPFWNMTYTWPDAYYITINPKDALLPQELKNKGLAVHEDIATVLSNVL
TEQQKERKQGAV

Best Known Hit in BLASTp Results

- Subject Accession No: WP_099353704.1
- Subject GI: 1272174496
- Subject Name: NAD-dependent protein deacetylase
- Subject Organism: *B.onubensis*
- Research Article(s) for Biotech Significance (PubMed IDs): PMID: 23226010

>WP_099353704.1 NAD-dependent protein deacetylase [Bacillus onubensis]
MLPHQYQNNIDTILQKIKEADAIVVGGASGMSAAAGYNWYRDDENFRKYFNAFAVEYGIDSIFGGFYYKF
RTEEERWAYLATLINFVADVPIGQPYKDLYEILKDKNYYILTTNQDTQFLQVFPEEKVSAIQGNWTYLQC
SGPCHDGIYPYAEQAKELCNHIEGTKIPSSMVPTCPECGGPMELWVRSFVFLEGSKYHDEHRKYREFLLE
NHNKKVLFLELGVGRMTPMFIQEPFWNMTYSWPDAYYITINPKDALLPQQLKDKNKGIAIHEDIAPVLRD
VLTKQQTEGEKGAV

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
hypothetical protein [Bacillus lentus]	610	610	100%	0.0	100.00%	WP_066139086.1
NAD-dependent protein deacetylase [Bacillus onubensis]	531	531	100%	0.0	85.37%	WP_099353704.1
NAD-dependent protein deacetylase [Lysinibacillus sp. OL1]	507	507	99%	5e-180	81.23%	WP_131520440.1
hypothetical protein [Bacillus sp. MRMR6]	496	496	100%	1e-175	84.69%	WP_075688611.1

Figure 2: BLAST results of BIPs

NAD-dependent protein deacetylase [Bacillus onubensis]
Sequence ID: **WP_099353704.1** Length: 294 Number of Matches: 1
Range 1: 1 to 294

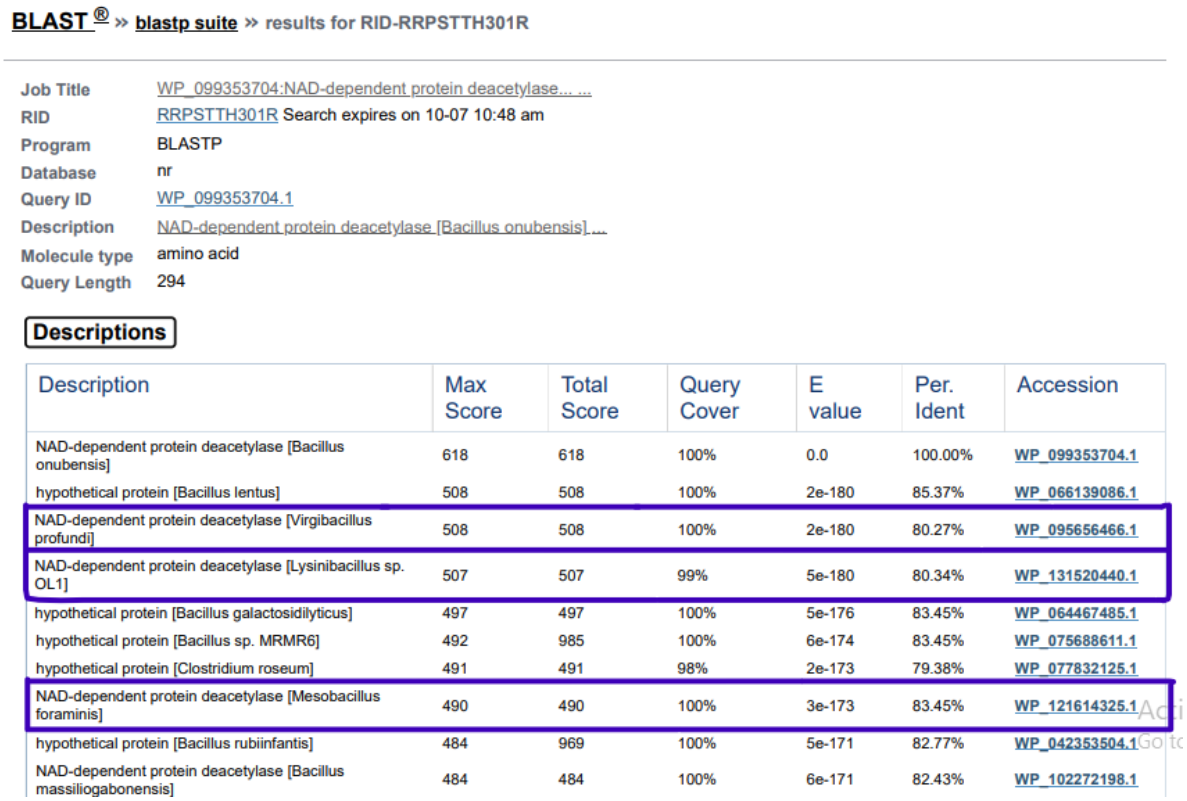
Score	Expect	Method	Identities	Positives	Gaps	Frame
531 bits(1369)	0.0()	Compositional matrix adjust.	251/294(85%)	271/294(92%)	2/294(0%)	

```
Query  1    MLTQQYQNTIGTILFKIFEADAIVVGGAAGMSAAGGYNWYIDDENFRFHFNVFAFKYGID  60
Sbjct  1    ..PH....N.D...Q..K..........S.....A.....R......KY..A..VE....  60

Query  61   SIFGGFYYKFRTEEERWAYLATLINFVTFVPIGQPYKDLYEITKDKNYYILTTNQDTQFL  120
Sbjct  61   ............................AD.............L..................  120

Query  121  QVFPEEKVSAIQGNWTYLQCSGPCHDGIYPYAEQAKELCAHIDGTKIPSDMVPKCPECGG  180
Sbjct  121  ........................................N..E......S...T......  180

Query  181  PMELWVRSFVFLEGEKYRDEHHKYRTFLLENQNKKILFLELGVGQMTPMFIKEPFWNMTY  240
Sbjct  181  ..............S..H...R...E.....H...V........R......Q........  240

Query  241  TWPDAYYITINPKDALLPQEL--KNKGLAVHEDIATVLSNVLTEQQKERKQGAV  292
Sbjct  241  S..................Q.KD....I.I.....P..RD...K..T.GEK...  294
```

Figure 3: BLAST result in pairwise alignment format

Validation Experiments

The NAD-dependent protein deacetylase was searched in NR database using NCBI Protein BLAST.It was observed that NAD-dependent protein deacetylase [*Virgibacillusprofundi*], NAD-dependent protein deacetylase [*Lysinibacillus sp. OL1*] and NAD-dependent protein deacetylase [*Mesobacillusforaminis*] has the close homologs with same name (Fig. 4). So, the hypothetical protein WP_066139086.1 of *B. Lentus* is predicted to be having the function of NAD-dependent protein deacetylase. This prediction came from having the sequence similarity of 85%from *B. Onubensis.*



BLAST® » blastp suite » results for RID-RRPSTTH301R

Job Title	WP_099353704:NAD-dependent protein deacetylase... ...
RID	RRPSTTH301R Search expires on 10-07 10:48 am
Program	BLASTP
Database	nr
Query ID	WP_099353704.1
Description	NAD-dependent protein deacetylase [Bacillus onubensis] ...
Molecule type	amino acid
Query Length	294

Descriptions

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
NAD-dependent protein deacetylase [Bacillus onubensis]	618	618	100%	0.0	100.00%	WP_099353704.1
hypothetical protein [Bacillus lentus]	508	508	100%	2e-180	85.37%	WP_066139086.1
NAD-dependent protein deacetylase [Virgibacillus profundi]	508	508	100%	2e-180	80.27%	WP_095656466.1
NAD-dependent protein deacetylase [Lysinibacillus sp. OL1]	507	507	99%	5e-180	80.34%	WP_131520440.1
hypothetical protein [Bacillus galactosidilyticus]	497	497	100%	5e-176	83.45%	WP_064467485.1
hypothetical protein [Bacillus sp. MRMR6]	492	985	100%	6e-174	83.45%	WP_075688611.1
hypothetical protein [Clostridium roseum]	491	491	98%	2e-173	79.38%	WP_077832125.1
NAD-dependent protein deacetylase [Mesobacillus foraminis]	490	490	100%	3e-173	83.45%	WP_121614325.1
hypothetical protein [Bacillus rubiinfantis]	484	969	100%	5e-171	82.77%	WP_042353504.1
NAD-dependent protein deacetylase [Bacillus massiliogabonensis]	484	484	100%	6e-171	82.43%	WP_102272198.1

Figure 4: BLAST results of NAD-dependent protein deacetylase

Significance of NAD-dependent protein deacetylase

A transcriptional unit with two genes including Sirtuin Sir2 (deacetylase enzyme for NAD+-dependent acetate consumption) and a universal stress protein genes. The proposed transcriptional units for stress-sensitive inorganic sulphate absorption and acetate utilization could explain biological mechanisms that restrict *Bacillus* species survival by sulphate and acetate. Given the relevance of sirtuins in mammalian physiology, more research on the USP-Sir2 transcriptional unit of *B. megaterium* should help explain mammalian acetate metabolism in glucose-limiting circumstances like caloric restriction.

B. megaterium soil inhibiting bacteria has six USP genes, one of which is on a plasmid. Another *Bacillus* species that lives in the soil, *B. subtilis*, has two USP genes. A transcriptional unit consisting of sirtuin (Sir2) gene and a USP gene discovered solely in *B. megaterium* was of particular interest. Deacetylase enzyme Sir2removes acetyl groups from lysine amino acids in proteins, found in bacteria and higher eukaryotes,using the nicotinamide adenine dinucleotide ion (NAD+). Sirtuin (SrtN), a NAD+-dependent deacetylase, and AcuC, a non-NAD+-dependent deacetylase, are necessary in *B. subtilis* to keep the enzyme acetyl coenzyme A (Ac-

CoA) synthetase (AcsA) active (i.e. deacetylated), so that the cell can grow at low acetate concentrations.

Sir2 proteins influence cell ageing, chromosomal stability, and energy consumption in response to nutritional cues, gene silencing, and stress response. The sirtuin gene family has been studied intensively as a potential therapeutic target for age related illnesses, obesity, cardiovascular disease, and cancer(Williams *et al*., 2012).

CONCLUSIONS

The hypothetical protein WP_066139086.1 of *B. lentus* is predicted to be having the function of NAD-dependent protein deacetylase which is also referred as Sirtuins. This prediction came from having the amino acid sequence identity of 85% and similarity of 92% from *B. onubensis.*Only 23 out of 294 amino acids are found to be different. Further investigation is needed to inquire this uniqueness of WP_066139086.1.

REFERENCES

Bayat A (2002). Science, medicine, and the future: Bioinformatics. BMJ: Brit Med J. 324(7344): 1018.

Benson D A, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman D J, Ostell J and Sayers E W (2016). GenBank. Nucl acids Res. 45(D1): D37-D42.

Contesini FJ, Melo RR D and Sato HH (2018). An overview of Bacillus proteases: from production to application. Crit Rev Biotechnol. 38(3):321-334.

Eisele T, Stressler T, Kranz B and Fischer L (2013). Bioactive peptides generated in an enzyme membrane reactor using Bacillus lentus alkaline peptidase. Europ Food Res Technol. 236(3): 483-490.

Ijaq J, Malik G, Kumar A, Das P S, Meena N, Bethi N and Suravajhala P (2019). A model to predict the function of hypothetical proteins through a nine-point classification scoring schema. BMC bioinformat. 20(1): 14.

Ismaya I, Wangsa W (2011). The Role of Protein Biochemistry in Biotechnology. INOVASI Online. 19: 11-18.

Jørgensen P L, Tangney M, Pedersen P E, Hastrup S, Diderichsen B and Jørgensen ST (2000). Cloning and sequencing of an alkaline protease gene from Bacillus lentus and amplification of the gene on the B. lentus chromosome by an improved technique. Appl environ microbial. 66(2): 825-827.

Kitts P A, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V and Kimchi A (2016). Assembly: a resource for assembled genomes at NCBI. Nucl acids Re. 44(D1): D73-D80.

Mlecnik B, Galon J and Bindea G (2018). Comprehensive functional analysis of large lists of genes and proteins. J proteomic. 171: 2–10.

Shahbaaz M, Imtaiyaz H M, and Ahmad F (2013). Functional annotation of conserved hypothetical proteins from Haemophilusinfluenzae Rd KW20. PloS oneline 8(12): e84263.

Sharma N, Kapoor G and Neopaney B (2006). Characterization of a new bacteriocin produced from a novel isolated strain of Bacillus lentus NG121. Anton Van Leeuwen. 89(3): 337-343.

Shi M, Gao J and Zhang M Q (2017). Web3DMol: interactive protein structure visualization based on WebGL. Nucl Acids Res. 45(W1): W523–W527.

Singh R, Kumar M, Mittal A and Mehta P K (2016). Microbial enzymes: industrial progress in 21st century. Biotech, 6(2): 174.

Speed T P (2003). Biological sequence analysis. *arXiv* preprint math/0304372.

Villesen P (2007). FaBox: an online toolbox for fasta sequences. Mol ecology Notes, 7(6): 965-968.

Williams BS, Isokpehi R D, Mbah AN, Hollman A L, Bernard C O, Simmons SS and Garner B L (2012). Functional annotation analytics of Bacillus genomes reveals stress responsive acetate utilization and sulfate uptake in the biotechnologically relevant Bacillus megaterium. Bioinformatic Biol Insights, 6: BBI-S7977.

Xu D and Coˆte J C (2003). Phylogenetic relationships between Bacillus species and related genera inferred from comparison of 3′ end 16S rDNA and 5′ end 16S–23S ITS nucleotide sequences. Int J Systemat Evolut Microbiol. 53(3): 695-704.