Research paper

Pan-Genome Analysis of *Bacillus licheniformis* to Find Patterns in Industrially Important Genes

Syed Muhammad Iqbal Azimuddin [1,2,3*], Dureshahwar Waseem [1]

[1]Center for Genome Research, Department of Biosciences,

Mohammad Ali Jinnah University, Karachi-75400, Pakistan.

[2]Xinjiang Institute of Ecology and Geography,

818 South Beijing Road, Urumqi 830011, Xinjiang, China.

[3]University of Chinese Academy of Sciences, Beijing 100864, China.

[*]Corresponding Author: Dr. Syed Muhammad Iqbal Azimuddin

(iqbal.azimuddin@jinnah.edu)

ABSTRACT

Bacillus licheniformis found in the soil and feathers of terrestrial birds. In an aerobic environment *B. licheniformis* makes bacteriocins and under anaerobic conditions produces lichenin. To produce commercially important heterologous proteins, Bacillus has been identified as a safe genus including many strains from *B. licheniformis*. Pan-genome analysis of the bacteria from the perspective of targeted industrial important genes/proteins was investigated in this study. In the present study, *B. licheniformis* genomic data of 177 strains and their industrially significant protein sequences were downloaded from Refseq and UniProtKB databases respectively. Bioinformatics analysis (i.e. standalone BLAST) was performed with the help of different Python scripts and tools on all strains to find patterns (genes presence/absence frequency polymorphism) for the targeted industrially important genes. A heat map was constructed to show how the targeted genes were related to each other, and how the strains were related to each other based on trends in the presence and absence of genes. The current pan-genome analysis on 39 *B. licheniformis* genomes predicted that bacteria have an open pan-genome consisting of 165,775 gene count, and a core genome gene count consisting of 124,882-120,026 genes. The pan-genome cluster count was comprised of 7233-8151 genes and a core genome cluster count was comprised of 3012- 3075 genes. According to clusters of orthologous genes (COG) categories. Metabolism and non-metabolism genes were separated, the biggest part of the core genome contains genes with metabolic functions (45.29%), while (11%) genes were involved in non-metabolic function and found in the housekeeping process. This core genome showed strong conservation in metabolic genes. We also found that the ribosomal proteins in the bacterial genome fall in the COG category (J), encodes for Translation, ribosomal structure & biogenesis. To investigate the evolutionary relatedness of the bacterial genome a phylogenetic tree shows 14 separate groups. Groups A and B, and groups D and E were closely related to each other, while groups D and E were distantly related to groups A and B. According to the heat map, groups A and B contain a maximum number of genes except for the Beta-mannosidase genes. This phylogenetic study provides how species evolve due to genetic changes.

KEYWORDS: *Bacillus licheniformis*; Pan-Genome; Standalone blast; clusters of orthologous genes; serine protease; peptidase; mannosidase.

INTRODUCTION

The family Bacillaceae includes a variety of Gram-positive, endospore-forming bacteria together referred to as the genus Bacillus. It is made up of several species from a wide variety of ecological niches, including the intestinal tract of humans, animals, various fermented foods, vegetables, milk, and more. A common saprophytic organism in the environment, *Bacillus licheniformis* is a Gram-positive, endospore-forming bacteria. This species is closely related to *Bacillus subtilis*, an organism that, in terms of depth of study, is only surpassed by *Escherichia coli* [1]. *B. licheniformis* is a facultative anaerobe, in contrast to most other bacilli, which are primarily aerobic, which may enable it to flourish in more ecological niches. However, because the species often survive in the soil as endospores, the importance of this trait to ecological denitrification may be minimal. Some isolates of *B. licheniformis* are capable of denitrification [2]. Numerous species in this genus including *B. clausii, B. cereus, B. subtilis, and B. licheniformis,* are regarded as safe and have strong probiotic reputations [3]. *B. licheniformis* and its extracellular compounds have several industrial and agricultural uses. Several proteases, amylase, penicillinase, cellulases, fumarate hydratase, arginase, asparaginase, β-mannanase, and many pectinolytic enzymes are among the industrial enzymes that have been produced using this species for decades [4]. The detergent industry and the process of dehairing and bating leather both use proteases from *B. licheniformis* [5]. The significant benefit of involving microorganisms in the development of α-amylases is the low-cost manufacturing capacity and organisms are not difficult to control to get enzymes of needed qualities. Amylases have wide applications, particularly in food Industries. It is essential for commercial applications because of its major function in starch hydrolysis processes used in food, paper, fermenting. and fabric production [6]. Lipases from *B. licheniformis* have the most diverse variety of uses. A variety of sources, including microbial, vegetable, and animal sources, can produce lipase. The development of commercially available microbial lipases in the detergents sector is the main innovation that allowed harsh chlorine bleach to be replaced with lipase and lowered sewage and industrial contamination of freshwater [7]. *B. licheniformis* strains are utilized to make a variety of specialty compounds, including acetic acid, inosine, inosinic acid, and poly—glutamic acid, as well as peptide antibiotics like bacitracin and proticin.

Since the eight *S. agalactiae* strains were used as the first examples of pan-genomes in 2005 [8], this analysis has been applied to improve our study of bacterial genomic architecture. A pan-genome introduced an extensive collection of genes from several strains of a species. Pan-genome analyses are frequently used to examine variations in gene function and the evolution of species. Horizontal gene transfers result in a large variation in the genes that each strain possesses [8]. Pan-genome analyses identify a collection of "core genes" that are found in all strains, "accessory genes" that are found in two or more strains, and "unique genes" that are unique to a particular strain [9]. The pan-genome can be divided into open and close pan-genomes, which can be used to describe specific species and are typically influenced by factors like population size and HGT [10]. Up till now, several popular tools have been built to analyze the entire genome, which includes PanOCT [11], Roary [12], Panseq, BPGA, PGAP, and PanViz, and R tools like FindMyFriends and PanVizGenerator can be used to analyze the pan-genome data.

To conduct a pan-genome analysis for the current study, we have curated 177 strains of *B. licheniformis* that are present in the NCBI database. Pan-genome analysis of *B. licheniformis* was performed to acquire a better understanding of the functional variations that affect the strain's dynamic evolutionary processes.

MATERIALS AND METHODS

Selection of species and strains

The organism *B. licheniformis* has been selected through a literature survey and 177 different strains of *B. licheniformis* were downloaded from RefSeq Database shown in Figure 1

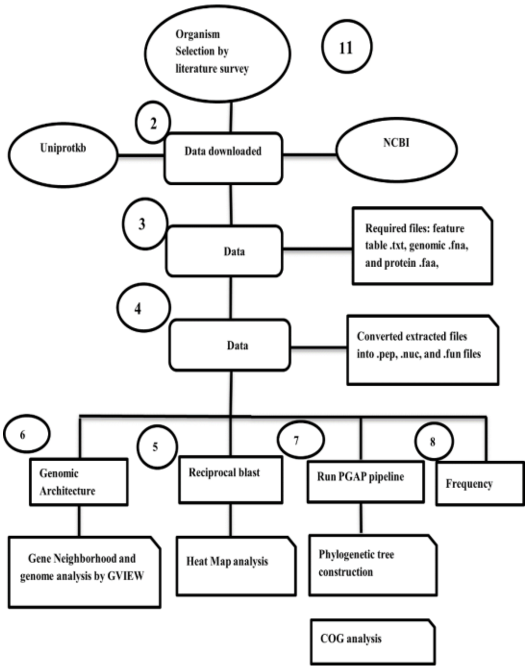

Organism Selection by literature survey
11
2
Uniprotkb
Data downloaded
NCBI
3
Data
Required files: feature table .txt, genomic .fna, and protein .faa,
4
Data
Converted extracted files into .pep, .nuc, and .fun files
6
5
7
8
Genomic Architecture
Reciprocal blast
Run PGAP pipeline
Frequency
Gene Neighborhood and genome analysis by GVIEW
Heat Map analysis
Phylogenetic tree construction
COG analysis

Figure 1: The workflow of the study.

Analysis through Reciprocal Blast

Standalone blast that enables BLAST searches to be carried out on local systems against databases downloaded from NCBI or built locally. These programs operate through command windows like DOS and take input through command line switches that are text-based. No graphical user interface is present [13]. To perform the reciprocal blast by running Blastp a query protein was downloaded from UniProtKB and merged all 177 strains into a single FASTA file, then a protein database was created with all the downloaded query sequences, a python script BLASTP was run for protein sequences, and 16 output files of query proteins was generated, and the result was analyzed as shown Figure 2 [14].

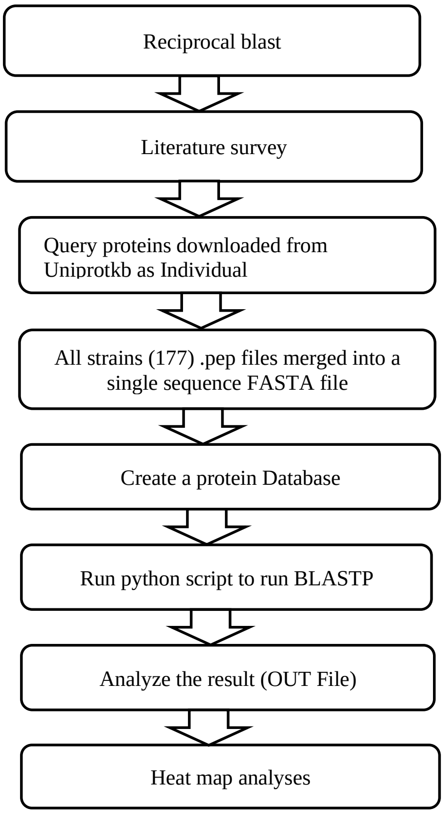

Reciprocal blast
Literature survey
Query proteins downloaded from Uniprotkb as Individual
All strains (177) .pep files merged into a single sequence FASTA file
Create a protein Database
Run python script to run BLASTP
Analyze the result (OUT File)
Heat map analyses

Figure 2: shows the flow chart of the standalone blast.

Phylogenetic and Heat Map Analysis

The UPGMA approach utilizes a sequential clustering algorithm and nearby topological connections are constructed and arranged after reducing the similarity and a dendrogram was generated in a step-by-step way [15]. The dendrogram was displayed using an interactive tree of life (iTOL, V6) in rectangular mode. iTOL is an open-source and freely available online tool that shows, manipulates, and annotates trees [16]. The heat map was visualized using the Python script which contains values displaying different colors for each value to be plotted. In a chart, the darker colors typically indicate higher values than the lighter ones [17].

Core-Genome and Pan-Genome Analysis

PGAP used BLASTP and BLASTN to examine the sequence similarity to analyze the results. The Identity score and E_value in BLAST were 50% and 1e-10 approximately. The outcomes were grouped by the MCL algorithm [18]. The core genome and pan-genome of 39 bacterial strains were deduced through PGAP which was further separated as shown Figure 3.

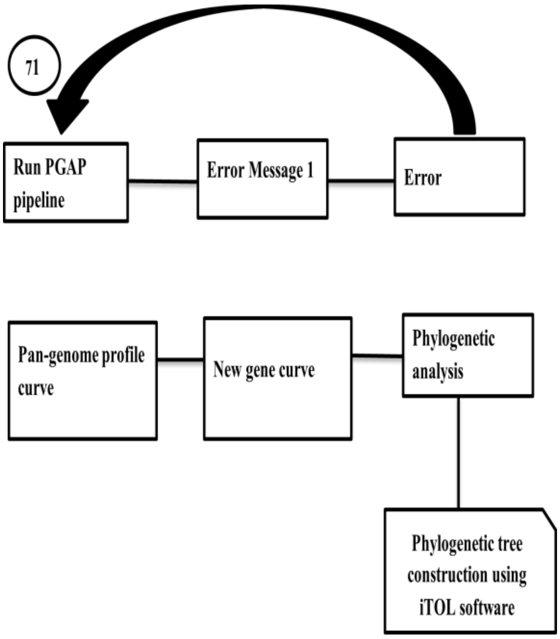
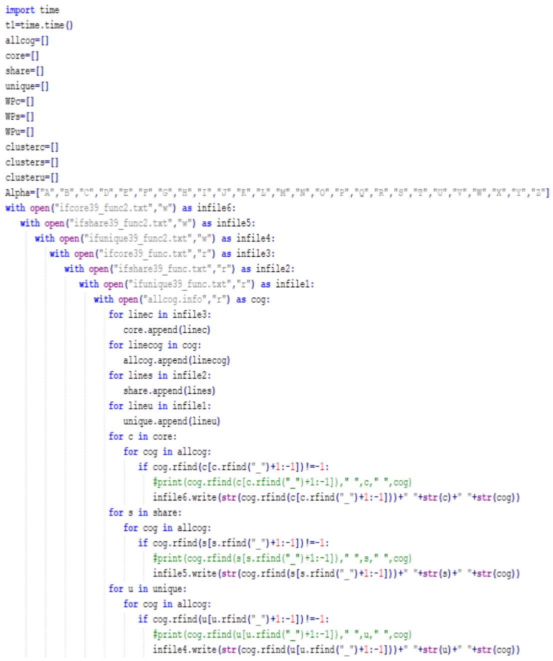


Figure 3: Shows the steps of bacterial pan-genome analysis on 39 strains.

Clusters of Orthologous Group (COG)

Clusters of orthologous group COG is a searchable database for bacterial genomics. The purpose of using the COG database is to get inside the evolutionary categorization of protein families (https://www.ncbi.nlm.nih.gov/research/cog) [19]. Figure 4 represents a COG python script of protein generated by analyzing the query sequence of entire genomes.



```
import time
t1=time.time()
allcog=[]
core=[]
share=[]
unique=[]
WPc=[]
WPs=[]
WPu=[]
clusterc=[]
clusters=[]
clusteru=[]
Alpha=["A","B","C","D","E","F","G","H","I","J","K","L","M","N","O","P","Q","R","S","T","U","V","W","X","Y","Z"]
with open("ifcore39_func2.txt","w") as infile6:
  with open("ifshare39_func2.txt","w") as infile5:
    with open("ifunique39_func2.txt","w") as infile4:
      with open("ifcore39_func.txt","r") as infile3:
        with open("ifshare39_func.txt","r") as infile2:
          with open("ifunique39_func.txt","r") as infile1:
            with open("allcog.info","r") as cog:
              for linec in infile3:
                core.append(linec)
              for linecog in cog:
                allcog.append(linecog)
              for lines in infile2:
                share.append(lines)
              for lineu in infile1:
                unique.append(lineu)
              for c in core:
                for cog in allcog:
                  if cog.rfind(c[c.rfind("_")+1:-1])!=-1:
                    #print(cog.rfind(c[c.rfind("_")+1:-1])," ",c," ",cog)
                    infile6.write(str(cog.rfind(c[c.rfind("_")+1:-1]))+" "+str(c)+" "+str(cog))
              for s in share:
                for cog in allcog:
                  if cog.rfind(s[s.rfind("_")+1:-1])!=-1:
                    #print(cog.rfind(s[s.rfind("_")+1:-1])," ",s," ",cog)
                    infile5.write(str(cog.rfind(s[s.rfind("_")+1:-1]))+" "+str(s)+" "+str(cog))
              for u in unique:
                for cog in allcog:
                  if cog.rfind(u[u.rfind("_")+1:-1])!=-1:
                    #print(cog.rfind(u[u.rfind("_")+1:-1])," ",u," ",cog)
                    infile4.write(str(cog.rfind(u[u.rfind("_")+1:-1]))+" "+str(u)+" "+str(cog))
```

Figure 4: Python script to analyze core, shared, and unique genome COG categories and their representative percentages.

Ethical approval

The study protocol was approved by Kohat University of Science and Technology, Kohat ethics review board. From each woman written consent was obtained after explaining the objectives of the study prior to her enrolment

Data analysis

SPSS (Statistical Package for Social Scientists) statistical software version 18 was used for data analysis. P value was calculated and 0.05 was considered significant at the 95% CI (Confidence Interval) level.

Frequency Element:

A python script was used to count the frequency of genes found in bacterial strains. The flowchart of frequency element at 177 strains, by running the python script three output files were generated, sorted frequency count, gene frequency, and frequency element. These files were used to create a gene frequency graph to identify genes frequency in all 177 strains in Figure 5.

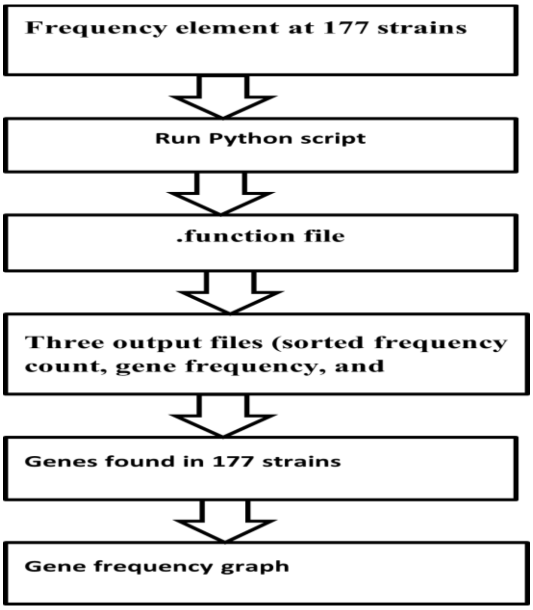

Frequency element at 177 strains
Run Python script
.function file
Three output files (sorted frequency count, gene frequency, and
Genes found in 177 strains
Gene frequency graph

Figure 5: Indicated the flowchart of frequency elements.

Antibiotic Resistance

Bacteria acquire resistance to antibiotics and antibiotic-resistant bacteria continue to develop, proliferate, and infect the host [20]. Clustal omega was used to construct a phylogenetic tree of 35 antibiotic-resistance genes sequences from *Bacillus species*, [21].

RESULTS AND DISCUSSION

Genome Size

Before the era of genome sequencing, the general structure and organization of bacterial genomes were well understood. It was believed that bacterial genome sizes vary by at least an order of magnitude and that even within a single bacterial species, there might be significant diversity in genome size [22]. Bacterial genomes are packed with genes and have a small amount of non-coding DNA, an increase in genome size results in an increase in the number of genes [23]. It has been proposed that a variety of distinct variables operate as the main regulators of genome size in bacteria and archaea.

Table 1 represents the gene count of all the top 60 and lowest 60 strains of *B. licheniformis* and it was carried out with statistical analysis (mean and standard deviation) based on the largest and smallest genome size. The gene count of the largest genome size was 4891 with a mean (average) value of 4448.1 and a standard deviation of 112.4. The gene count of the smallest genome size was 3408 with a mean (average) value of 4091.3 and a standard deviation of 99.3. Low standard deviation indicates the smallest genome was more stable and has greater consistency than the largest genome [24].

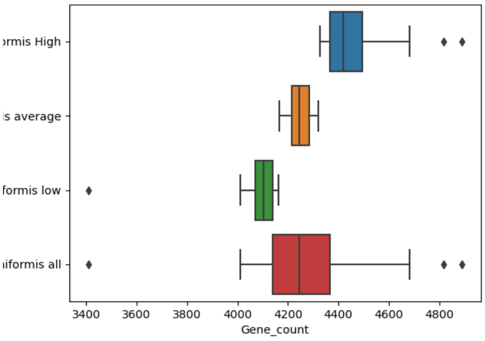


Figure 6: Represents the boxplot of all 177 *B. licheniformis* strains.

Figure 6 represents the boxplot that was used to compare all *B. licheniformis* strains with respect to their gene count. Box plots commonly known as Whisker's plots are a graphical technique that helps define the top and lower limits above which any data lying would be regarded as outliers. They are typically represented as quartiles as well as inter-quartile ranges. The strains including GCF_001896285 (YNP2-TSU), GCF_007830975 (NCTC 1024), and GCF_001925025 (B4089) were outliers because these data values occurred more than the upper limit or lesser than the lower limit. *B. licheniformis* strain YNP2-TSU draft genome sequence was discovered in the hydrothermal-vegetative microbiomes of Yellowstone National Park. Through automated annotation, these assembled sequence contigs projected 4,230 coding genes, 66 tRNAs, as well as 10 rRNAs. New cellulolytic bacteria may have an impact on the development of second-generation biofuels. These unclassified cellulolytic thermophiles can be found in abundance in the heated springs and hydrothermal characteristics of Yellowstone National Park [25].

Strain NCTC 1024 contains 4,301 coding genes, 50 tRNAs, and 1 rRNAs. Strain B4089 is involved in sporulation and spore germination protein including SpoVAB, SpoVAA, SpoVABEA, and SpoVAF. Dipicolinic acid (DPA) is released and partial core hydration occurs after spore germination, which is frequently started by germinant receptor (GR) reacting to nutrient germinants [26]. *B. licheniformis* spoVA operon, which has four genes in the sequence spoVAA, -B, -Ea, and -F, only expresses itself during sporulation in the forming spore These three SpoVA proteins including , SpoVAA, -B, and -Eb, were required for the proper production of *B.licheniformis* spore formation [27]. These proteins are necessary for the spore's latent state and the ingestion of dipicolinic acid (DPA). Since DPA-less *B. licheniformis* spores spontaneously germinate after being released from sporangium also during spore purification, DPA is necessary for spore stability [28]. Based on predictions from original sequences and, in some cases, the localization of proteins showed in growing bacteria, SpoVAA, -B, and -F likely to be integral membrane proteins [29]. While A subunits of a spore's nutrition germinant receptors have a large amount of sequence identity with SpoVAF (GRs) [30]. SpoVAA, -B, and -Ea show no visible similarity to recognized proteins. Bacterial strain B4089 contains 4,138 coding genes, 74 tRNAs, as well as 21 rRNAs.

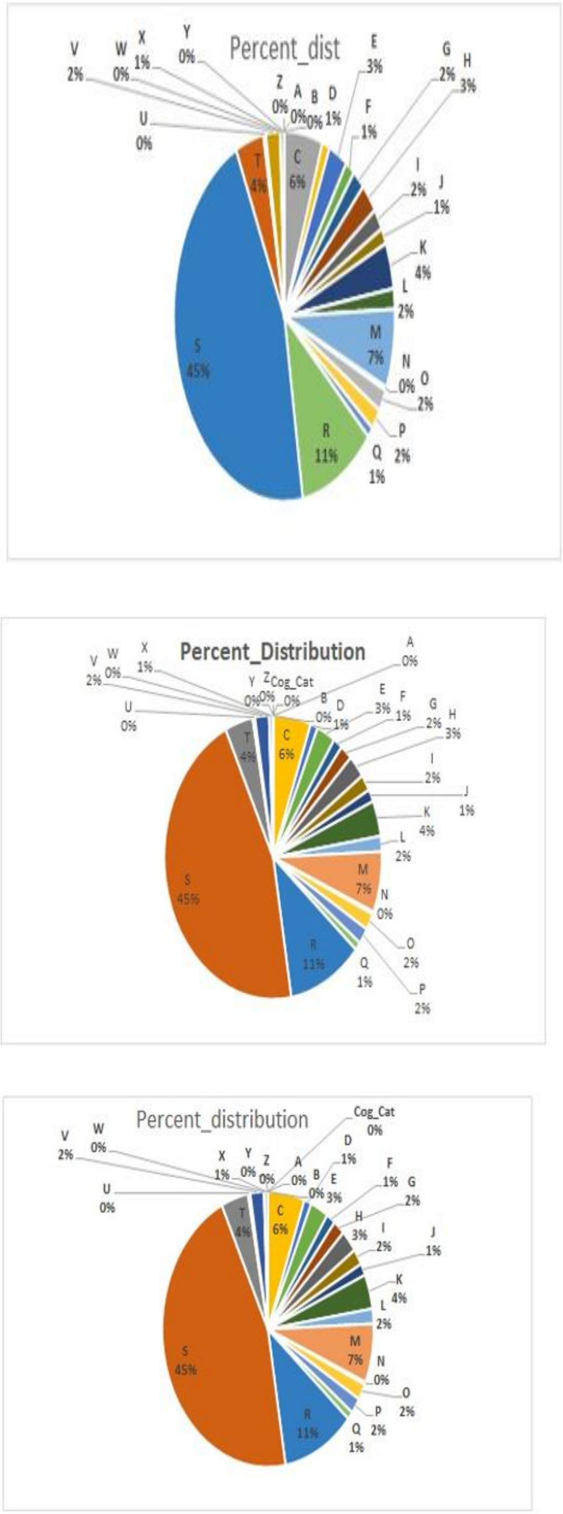


Figure 7: Indicated the Pie chart of percent distribution in comparison of genome sized based on the COG categories. Function unknown (S) is the largest COG category contains 45% genes count in these three genomes.

Figure 7 shows the pie charts of YNP2-TSU, NCTC 1024, and B4089 strains indicating that no visible difference was found in comparing the genomes as per COG comparison based on percent distribution.

Function unknown (S) is the largest COG category and contains 45% genes distribution including Spore maturation protein SpmB, conserved protein AF1501, DUF2226 family, membrane protein YckC, RDD family, metal-binding protein MJ1132, UPF0058 family, and so on. General function prediction only (R) contains 11% genes including Acetyltransferase, isoleucine patch superfamily, Tetratricopeptide (TPR) repeat, alkaline phosphatase, TctA family transporter, and Spore maturation protein SpmA, GTPase-interacting Yip1 domain, and Alpha/beta superfamily hydrolase. Cell wall/membrane/envelope biogenesis (M) contains 7% of genes including Nucleoside-diphosphate-sugar epimerase, UDP-glucose 4-epimerase, UDP-glucose 6-dehydrogenase, and Lipoprotein-anchoring transpeptidase ErfK/SrfK. These COG categories together involved in the largest part of *B. licheniformis* genome.

In prokaryotic chromosome, genes are not randomly dispersed; rather, they are located in functional neighborhoods, and frequently cluster in certain genomic areas [31].

Table 2 indicates industrially important proteins based on Clusters of the Orthologous group of proteins (COGs) and their neighborhood in (YNP2-TSU and B4089 genome size) with their genes number. A group of COG (G, E, D, M, O, and J) with their WP_Nos ensures that such pairs were highly conserved during evolution, their upstream and downstream genes in both (YNP2-TSU and B4089) genome remain consistent on either strands + and -. While a COG category (T, C, and K) was partially conserved, their neighborhood was flipped on upstream and downstream in YNP2-TSU and B4089 genomes.

Figure 6 Represents the boxplot of all 177 *B. licheniformis* strains. Box-plot of the distribution of size of the genomes (Mb) of all *B. licheniformis* strains. The horizontal black line at the center of the box plot represents the median. The bottom and top of the box represent the first and third quartiles. The external edges of the whiskers represent the inner 10th and 90th percentiles [32]. The red boxplot indicates all bacterial strains and contains an outlier at right upstream of strain YNP2-TSU and left upstream and compared them with the average gene count.

Standalone Blast:

Basic Local Alignment Search Tool (BLAST) [33] was used to find homologues of query sequences. Table 3 refers the industrially important proteins (query sequences) along with their WP_Nos were used in different Industries. Bacillus species are appealing industrial organisms for several purposes, such as their high growth rates, which result in short fermentation process time, their ability to produce enzymes in the extracellular environment, and the Food & Drug Administration's GRAS commonly recognized them as safe.

For species like *B. subtilis and B. licheniformis*. lipase, pectate-lyase, glycosyl hydrolase, proteases, peptidase, beta-mannosidase, and amylases are commonly available feed enzymes that are primarily used to enrich the feed of animals and poultry, textiles, detergents, chemicals, and biofuels are the markets where they operate [34]. They pervade every area of our everyday lives, and as a result, the marketplaces are vast.

Amylases are employed in the starch processing industries for the hydrolysis of polysaccharides such as starch into simple sugar constituents. Currently, these enzymes make up around 30% of the enzymes produced worldwide. Starch-converting enzymes are utilized in various of different industrial processes outside starch hydrolysis, including baking and as anti-salting agents or in detergents for porcelain and laundry [35].

Numerous industries, such as paper and pulp, fabric, laundry, biofuel production, the food and feed industries, brewing, and agriculture, have demonstrated the potential use of microbial cellulases [36]. Asparaginase is an enzyme that inhibits and limits the proliferation of malignant cells, making it a popular cancer treatment [37]. Arginase inhibitor of arginase/ornithine pathway used as a therapy for cardiovascular, CNS, and cancers involving high arginase expression and inhibitor has medicinal importance.

Table 3 shows categorized strains presence/absence trends of 16 industrially important proteins. If all the industrially important proteins were present in each set of strain such as SMIA-2 (GCF_009965255) then the strain is considered as highly rich strain. *Bacillus species* SMIA-2 is a crucial Brazilian strain for the synthesis of thermostable enzymes with industrial applications, including amylases, proteases, and cellulases, using a variety of industrial fermentation substrates, including sugarcane bagasse, corn steep liquor, as well as food waste [38]. If eight industrially important proteins were present in a given strain LMG 19409 (GCF_007832495) then it is moderately rich strain. This strain LMG 19409, was found to synthesize exopolysaccharide and lead to the rope rotting of Normandy cider (France) [39]. If only three proteins were present in each strain 127185/2 (GCF_001939535) then it is lowly rich strain.

Heat Map

Heat map (Figure 8) shows top 25 strains based on 16 industrially important proteins. Heat map shows how the absence/presence of selected genes established the connectivity among the strains and genes under study. A heat map was based on two parameters (i) how the strains were related to each other (ii) how the genes were related to each other based on presence/absence trends. Green color indicates genes were present and black color indicates genes were absent in a particular GCF-strains. Maximum industrially important genes were present in two strains LMG 20170 (GCF_007994255) and SMIA-2 (GCF_009965255) within the criteria of 100% identity match with query protein. The left side of the dendrogram shows asparaginase and fumarate hydratase were present in many strains within the criteria of 99% and 98% identity as indicated in table 3. In previous study, they created a protein-protein interaction (PPI) network of *B. licheniformis* strain WX-02 using the interolog method and a domain-based approach. This network has 2,448 nodes and 2,864 edges. A PPI enrichment analysis was performed by showing the PPI network as a heat map depends on several COG categories, The COG heat map was further examined, and the results revealed that protein interaction in *B. licheniformis* WX-02 mostly occurred in same functional categories [40].

Pan genome & core-genome features of *B. licheniformis:*

We construct the pan-genome of *B. licheniformis* by using the data set of 39 different strains. The pan-genome is comprised of the core genome (Genes are shared by all strains of a species), the shared genome (Genes are shared by two or more strains), and the unique genome (genes are specific to one strain only [41].

Table 7 represents the number of genes present in core, shared and unique genome, hence indicate the change in trend by changing the clustering criteria for example (50,60,70,80, and 90 identity percent. This indicated that the pan-genome gene count was 165,775, core genome gene count was (124882-120026), shared genome gene count was (38961-43412), and unique genome gene count was (1932-2337) in 39 *B. licheniformis* strains.

Identity-percent 50, 60, 70, 80, and 90 percent means the chances that two Sequences are similar. It can provide a valuable measure for evolutionary relationship [4].

Table 8 indicates that the pan-genome cluster count was 7233-8151 genes. Based on the dataset, the core, shared and unique genome cluster count along with their gene cluster percentage was 3012- 3075 core cluster genes (41.64 - 37.72%), 2296-2747 shared cluster genes (31.74%-33.71%), and 1925- 2329 unique cluster genes (26.61%-28.57%) respectively, Percent identity from (50%, 60%, 70%,80%, and 90%) was the clustering criteria adopted. As 50% criteria was the loose criteria in which the cluster count decreases while 90% criteria were the strongest criteria in which the cluster count increases.

Previously, a comparative analysis of 17 *E. coli* genomes was reported [43]. These 17 genomes under investigation have 5,020 genes. The conserved core genome count was 2,344 genes mainly encode core metabolic functions.

Clusters of Orthologous Groups (COG) database was used to perform functional annotation of genes within the pan genome, which revealed a distinct allocation of functional categories between the three pan genome sets. When metabolism and non-metabolism genes were separated, the biggest part of the core genome contains genes with metabolic functions (45.29%), for cellular processing and signaling (37.9%), and for information storage and signaling (10.73%). This contains genes includes in metabolism involving glycolysis, gluconeogenesis, pentose phosphate pathway and TCA cycle. A lower portion of metabolic genes was found in the shared genome contains (12.91%, respectively). The pan-genome size increases with the addition of new genes and predicting further addition of new genes by the evolution in the species.

Pan-genome profile evaluation and statistics:

Figure 9 Graph of New gene curve. shows the new gene distribution curve of the B.licheniformis pan-genome in which the new genes were supplemented by adding genome.

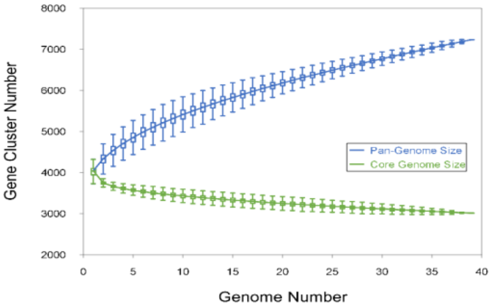


Figure 2: Pan-genome and core genome of *B. licheniformis*

The pan-genome curve increased from 4000 – to 7233 genes and the core genome curve was at a minimum of 4000-3012 genes (95% confidence interval) and therefore was almost stable as new genomes are added. The bacteria under study have open pan-genome other studies indicated those that colonized different environments and had a wide range of genetic exchange pathways, including *Streptococci, Meningococci, H. pylori, Salmonella, and E. coli* [8].

The Core Genome's Size and Content Provide Insight into Essential *B. licheniformis* Capabilities.

The average *B. licheniformis* genome codes 4,251 genes; hence, the average core-genome size varies from (72-75%), shared genome size (26.18%, 23.50%) and unique genome size (1.41%, 1.17%) as shown in table 6. The portion of core genome is extremely large as compared to other organisms like *C. difficile* (core shows the 25% of genome average) and *E. coli* (core shows 40% of genome average). Core genes related to a single or another gene family are actively maintained over a similar collection of genomes at the nucleotide sequence level [44].

Among 3012-3075 core genes 11% were found in the housekeeping process of COG categories (K and J). These genes were involved in non-metabolic functions like transcription (4%), translation, ribosomal structure, & biogenesis (7%). The COG method relies on whole gene (protein) sequences, which enables very straightforward and reliable identification of probable Orthologs and paralogs between all proteins produced in each genome. The COG technique, like other Orthologs discovery approaches uses sequence similarity search against specified proteomes to find pairwise best hits. These COGs carry a mean of (22%) of all genes in all organisms, on average, they make up almost half (46.2%) of the core genes in the clade, which contains more than 30 organisms [45].

Genes included in metabolic function were more conserved like carbohydrate transport and metabolism (G, 6%), Amino acid transport and metabolism (E, 8%), energy production and conversion (C, 5%). Genes that were involved in the cellular process and signaling such as Cell cycle control, cell division, chromosome partitioning (D, 5%), Posttranslational modification, protein turnover, chaperones (O, 6%), Signal transduction mechanism (T, 9%), and Cell wall/membrane/Envelope biogenesis (M, 10%). On the other hand, the accessory genome was enriched in involved in general function prediction only (R, 13%), and both core and shared genome had a high proportion of genes involved in metabolism including carbohydrate transport and metabolism (G, 4%), and Amino acid transport and metabolism (E, 4%). Amino acid transport and metabolism (E) were the largest category in core genome of most of the species [41]. While in the core of *B. licheniformis*, Carbohydrate transport and metabolism (G) was the most abundant category.

Frequency Element:

The Bar graph indicates gene frequency in all 177 strains. Out of 124,882 genes, approximately 28916 genes were found only in 1 strain while this number was reduced to 4860 and were found in 2 strains as indicated in figure 10. The left bar indicated genes were involved in horizontal gene transfer and showed resistance to antibiotic biosynthesis monooxygenase. Members of the monooxygenase family involved in antibiotic biosynthesis catalyze O_2-dependent oxidations as well as oxygenations without the use of any inorganic or metallic cofactors. Numerous tiny, structurally simple, naturally solvent- and temperature-stable enzymes that catalyze both cofactor-dependent and O2-dependent oxidations or monooxygenations are members of the antibiotic biosynthesis monooxygenase family [46]. The bulk of housekeeping genes are found in the core genome, which also contains "accessory" genomic fragments. Previous studies showed that the core genome represents 77.1% of its total genes shared by all strains of *B. subtilis* and is highly conserved [47].

Phylogenetic trees show 14 separate groups. Group A and B, and group D and E were closely related to each other, while group D and E were distantly related to group A and B Figure 15. The finding complies the geographical origin of the strains [48].

Table 1: Genome size variation and their gene counts in different strains.

largest genome (top 60)	**Gene count**
largest genome size	4891
Mean	4448.1
standard deviation	112.4
Smallest genome (lowest 60)	**Gene count**
smallest genome size	3408
Mean	4091.3
standard deviation	99.3

Table 2: Represents the 16 industrially important query proteins based on their Identity percent, e-value, COG categories and their percentage found in core, shared and unique genome based on clustering criteria (0.5, 0.6, 0.7, 0.8 and 0.9).

Protein	**WP-No**	**Identity**	**e-value**	**cog cat**	**COG %**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Alpha-amylase	WP_025807921	100%	0	G	6.47	Core	Core	Core	Core	Core
Arginase	WP_145871597	100%	0	E	8.09	core	Core	Core	Core	Core
Cellulase	WP_016885470	100%	0	G	6.47	Core	Core	Core	Core	Core
fumarate hydratase	WP_202604381	100%	0	C	5.28	Core	Core	Core	Core	Core
serine protease a	WP_003178689	99%	0	D	5.24	Core	Core	Core	Core	Core
serine protease b	WP_185955468	100%	0	O	6.03	Core	Core	Core	Core	Core
Asparaginase	WP_025808111	99%	0	J E	3.62, 8.09	Core	Core	Core	Core	Core
Lipase	WP_003183220	60%	3.00E-80	O	6.03	Core	Core	Core	Core	Core
pectate lyase	WP_003185613	100%	2.00E-163	G	4.42	share	share	share	share	share
glutamate methyl esterase	WP_003182242	100%	0	T	9.2	core	core	core	core	core
Transpeptidase	WP_003178129	100%	0	D M	5.24,10.12	core	core	core	core	core
Peptidase	WP_025808265	99%	0	D	5.24	core	core	core	core	core
BlaI/MecI/CopY family transcriptional	WP_003178582	99%	7.00E-90	K	7.11	core	core	core	core	core
glycosylhydrolase	WP_025807837	93%	0	E R	4.07,12.96	share	share	share	share	share
beta-mannosidase	WP_003179605	75%	0	G	6.47	core	core	core	core	core
pectate trisaccaride lyase	WP_202631970	100%	0	G	4.42	shared	shared	shared	shared	shared

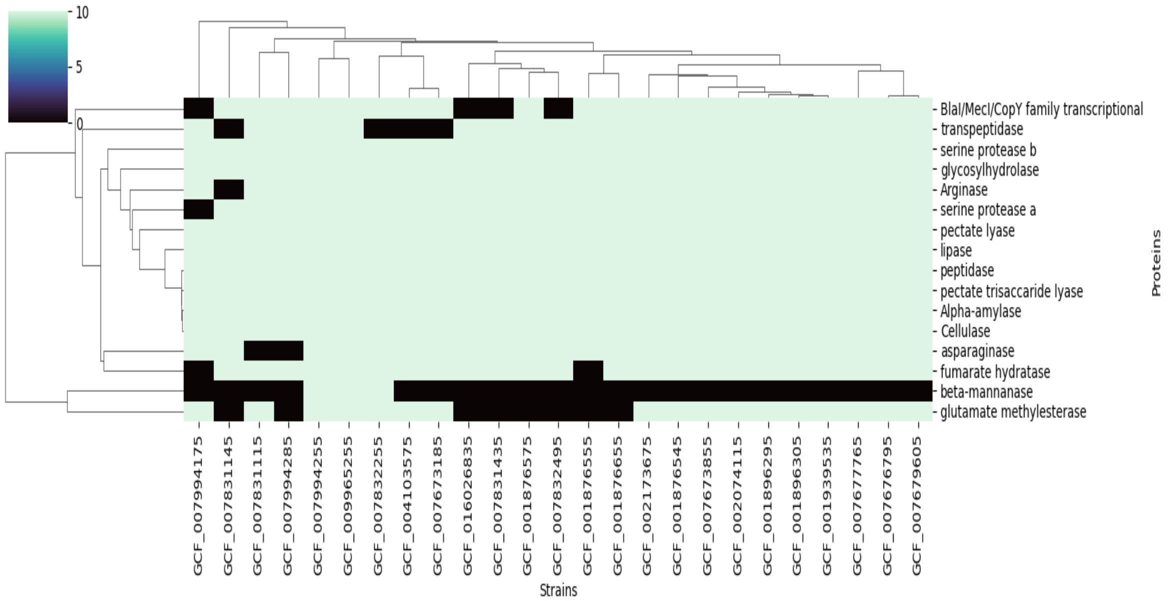


Figure 3: Heatmap of Strains (n=25) exhibits the presence of majority query proteins through blast hit.

Table 3: Represents pan-genome count in core. Shared and unique genome.

Genome	**50%**	**60%**	**70%**	**80%**	**90%**
Core genome count	124882	123328	122089	121256	120026
Share genome count	38961	40448	41613	42358	43412
Unique genome count	1932	1999	2073	2161	2337
Core genome percentage	75.33%	74.39%	73.64%	73.14%	72.40%
Share genome percentage	23.50%	24.39%	32.29%	25.55%	26.18%
Unique genome percentage	1.17%	1.20%	27.08%	1.31%	1.41%
Total WP count is	165775	165775	165775	165775	165775

Table 4: Represents the total cluster size with respect to their percent criteria in core, shared and unique genomes.

Cluster	**50%**	**60%**	**70%**	**80%**	**90%**
core cluster count	3012	3087	3098	3093	3075
share cluster count	2296	2379	2463	2561	2747
unique cluster count	1925	1992	2066	2153	2329
Total cluster count	7233	7458	7627	7807	8151

core cluster percentage	41.64%	41.40%	40.61%	39.61%	37.72%
share cluster percentage	31.74%	31.89%	32.29%	32.81%	33.71%
unique cluster percentage	26.61%	26.71%	27.08%	27.57%	28.57%

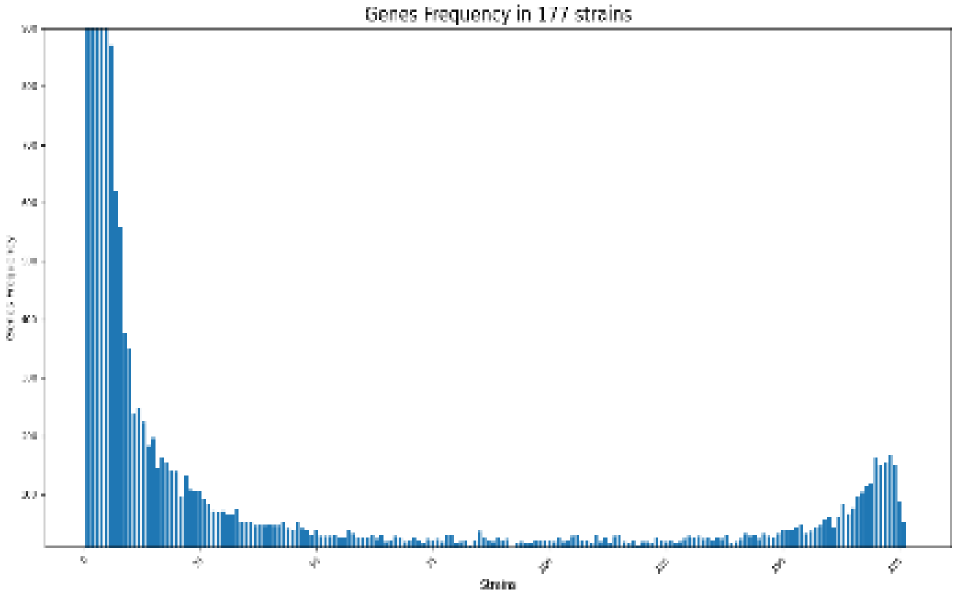


Figure 4: The bar graph represents the frequency of genes found in putative horizontal gene transfer on left side and putative core genome genes on right side of the graph.

REFERENCES

1. Mc Calla, T. M. (1978). Introduction to Soil Microbiology: By Martin Alexander. John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10016. 1977. 467 p. $18.95, Wiley Online Library.

2. Rey, M. W., et al. (2004). "Complete genome sequence of the industrial bacterium Bacillus licheniformis and comparisons with closely related Bacillusspecies." Genome biology 5(10): 1-12.

3. Cutting, S. M. (2011). "Bacillus probiotics." Food microbiology 28(2): 214-220.

4. Erickson, R. (1976). "Industrial applications of the bacilli: a review and prospectus." Microbiology: 406-419.

5. Eveleigh, D. E. (1981). "The microbiological production of industrial

chemicals." Scientific American 245(3): 154-179.

6. Abdel-Fattah, Y. R., et al. (2013). "Production, purification, and characterization of thermostable α-amylase produced by Bacillus licheniformis isolate AI20." Journal of Chemistry 2013.

7. E Ibrahim, N. and K. Ma (2017). "Industrial applications of thermostable enzymes from extremophilic microorganisms." Current Biochemical Engineering 4(2): 75-98.

8.Medini, D., et al. (2005). "The microbial pan-genome." Current opinion in genetics & development 15(6): 589-594.

9. Errington, J. (1993). "Bacillus subtilis sporulation: regulation of gene expression and control of morphogenesis." Microbiological reviews 57(1): 1-33.

10. Tettelin, H., et al. (2008). "Comparative genomics: the bacterial pan-genome." Current Opinion in Microbiology 11(5): 472-477.

11. Fouts, D. E., et al. (2012). "PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species." Nucleic acids research 40(22): e172-e172.

12.Page, A. J., et al. (2015). "Roary: rapid large-scale prokaryote pan genome analysis." Bioinformatics 31(22): 3691-3693.

13. Tao, T. (2010). "Standalone BLAST setup for windows PC." BLASTVR Help [Internet], National Center for Biotechnology Information (US), Bethesda, MD.

14. Cock, P. J., et al. (2009). "Biopython: freely available Python tools for computational molecular biology and bioinformatics." Bioinformatics 25(11): 1422-1423.

15. Garcia-Vallvé, S. and P. Puigbo (2009). "DendroUPGMA: a dendrogram construction utility." Universitat Rovira i Virgili.

16. Letunic, I. and P. Bork (2021). "Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation." Nucleic acids research 49(W1): W293-W296.

17.Netek, R., et al. (2018). "Implementation of heat maps in geographical information system–exploratory study on traffic accident data." Open Geosciences 10(1): 367-384.

18. Enright, A. J., et al. (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic acids research 30(7): 1575-1584.

19. Galperin, M. Y., et al. (2019). "Microbial genome analysis: the COG approach." Briefings in Bioinformatics 20(4): 1063-1070.

20. Rajer, F. and L. Sandegren (2022). "The Role of Antibiotic Resistance Genes in the Fitness Cost of Multiresistance Plasmids." Mbio 13(1): e03552-03521.

21. Sievers, F. and D. G. Higgins (2021). The clustal omega multiple alignment package. Multiple sequence alignment, Springer: 3-16.

22.Herdman, M. (1985). "The evolution of bacterial genomes." The evolution of genome size: 37-68.

23.Mira, A., et al. (2001). "Deletional bias and the evolution of bacterial genomes." Trends in Genetics 17(10): 589-596.

24. Tintle, N. L., et al. (2012). "Evaluating the consistency of gene sets used in the analysis of bacterial gene expression data." BMC bioinformatics 13(1): 1-12.

25.O'Hair, J. A., et al. (2017). "Draft genome sequence of Bacillus licheniformis strain YNP1-TSU isolated from Whiterock Springs in Yellowstone National Park." Genome Announcements 5(9): e01496-01416.

26. Setlow, P. (2014). "Germination of spores of Bacillus species: what we know and do not know." J Bacteriol 196(7): 1297-1305.

27. Setlow, P. and E. Johnson (2012). Spores and their significance.(eds Doyle, M., & Beuchat, L.) Food microbiology: fundamentals and frontiers. 45–79, ASM Press, Washington, DC.

28.Paidhungat, M., et al. (2000). "Characterization of spores of Bacillus subtilis which lack dipicolinic acid." J Bacteriol 182(19): 5505-5512.

29.Setlow, P. (2003). "Spore germination." Current opinion in microbiology 6(6): 550-556.

30. Vernikos, G., et al. (2015). "Ten years of pan-genome analyses." Current Opinion in Microbiology 23: 148-154.

31. Rogozin, I. B., et al. (2002). "Connected gene neighborhoods in prokaryotic genomes." Nucleic acids research 30(10): 2212-2223.

32. Touchon, M., et al. (2016). "Genetic and life-history traits associated with the distribution of prophages in bacteria." The ISME journal 10(11): 2744-2754.

33.Madden, T. (2003). "The BLAST sequence analysis tool." The NCBI handbook.

34. Danilova, I. and M. Sharipova (2020). "The practical potential of bacilli and their enzymes for industrial production." Front Microbiol 11: 1782.

35. Tiwari, S., et al. (2015). "Amylases: an overview with special reference to alpha amylase." J Global Biosci 4(1): 1886-1901.

36. Kuhad, R. C., et al. (2011). "Microbial cellulases and their industrial applications." Enzyme research 2011.

37. Jeyaraj, S. K., et al. (2020). "A study on production and evaluation of L-asparaginase obtained from Bacillus subtilis." Test Eng Manag 82: 4413-4416.

38. Bernardo, S. P. C., et al. (2020). "Draft genome sequence of the thermophilic bacterium Bacillus licheniformis SMIA-2, an antimicrobial-and thermostable enzyme-producing isolate from Brazilian soil." Microbiology Resource Announcements 9(17): e00106-00120.

39.Grande, M., et al. (2006). "Inhibition of Bacillus licheniformis LMG 19409 from ropy cider by enterocin AS-48." Journal of Applied Microbiology 101(2):422-428.

40. Han, Y.-C., et al. (2016). "Prediction and characterization of protein-protein interaction network in Bacillus licheniformis WX-02." Scientific reports 6(1): 1-11.

41.Park, S.-C., et al. (2019). "Large-scale genomics reveals the genetic characteristics of seven species and importance of phylogenetic distance for estimating pan-genome size." Front Microbiol 10: 834.

42.Pearson, W. R. (2014). BLAST and FASTA similarity searching for multiple sequence alignment. Multiple sequence alignment methods, Springer: 75-101.

43.Rasko, D. A., et al. (2008). "The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates." J Bacteriol 190(20): 6881-6893.

44. Charlebois, R. L. and W. F. Doolittle (2004). "Computing prokaryotic gene ubiquity: rescuing the core from extinction." Genome research 14(12): 2469-2477.

45. Segata, N. and C. Huttenhower (2011). "Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies." PLoS One 6(9): e24704.

46. Brito, P. H., et al. (2018). "Genetic competence drives genome diversity in Bacillus subtilis." Genome biology and evolution 10(1): 108-124.

47. Machovina, M. M., et al. (2016). "Monooxygenase substrates mimic flavin

to catalyze cofactorless oxygenations." Journal of Biological Chemistry 291(34): 17816-17828.

48. De Clerck, E. and P. De Vos (2004). "Genotypic diversity among Bacillus licheniformis strains from various sources." FEMS Microbiology Letters 231(1): 91-98.