Functional Annotation Of Hypothetical Proteins From The Bacillus Paralicheniformis Strain Bac84 Reveals Proteins with Biotechnological Potentials – an In Silico Approach

Md Atikur Rahman (✉ md.atikur.rahman@uni-jena.de)
Friedrich Schiller University Jena
Uzma Habiba Heme
Friedrich Schiller University Jena
Md. Anowar Khasru Parvez
Jahangirnagar University

Article

Keywords: Bacillus paralicheniformis strain Bac84, functional annotation, hypothetical proteins, 5 in silico, biotechnological potentials, extreme environments

Posted Date: July 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1862992/v1

TITLE:

Functional Annotation Of Hypothetical Proteins From The *Bacillus Paralicheniformis Strain Bac84* Reveals Proteins with Biotechnological Potentials – an In Silico Approach

Keywords: *Bacillus paralicheniformis strain Bac84*, functional annotation, hypothetical proteins, in silico, biotechnological potentials, extreme environments.

ABSTRACT

A significant number of proteins in the genome of the *Bacillus paralicheniformis strain Bac84* are annotated as functionally uncharacterized hypothetical proteins. Investigating these proteins' functions may help us to find novel targets for biotechnological applications. Therefore, the purpose of our research was to functionally annotate the hypothetical proteins from its genome. We employed a structured in-silico approach incorporating numerous bioinformatics tools and databases for functional annotation and characterization. Sequences of 414 hypothetical proteins were evaluated and we were able to successfully attribute a function to 37 hypothetical proteins. Moreover, we performed receiver operating characteristic analysis to assess the performance of various tools. Eight proteins were predicted with biotechnological potentials such as coenzyme A biosynthesis, phenylalanine biosynthesis, antibiotic biosynthesis, and others. Evaluation of the performance of the tools showed an accuracy of 98% which represented the rationality of the tools used. This work shows that this annotation strategy will make the functional characterization of unknown proteins easier and can find the target for further investigation.

INTRODUCTION

Bacillus paralicheniformis is a newly discovered species in the *Bacillus* genus (Dunlap et al., 2015). It is phylogenetically closely related to *B. licheniformis* (Dunlap et al., 2015; Du et al., 2019). In the biotechnology sector, B. licheniformis has already been employed to produce biochemicals, enzymes, antibiotics, and other things (Rey et al., 2004; Dunlap et al., 2015). Several current investigations have indicated that *B. paralicheniformis* species have a strong potential for the biosynthesis of antimicrobial compounds (Dhakal et al., 2013; Othoum et al., 2018). One of the strains can also inhibit plant pathogenic microbes (Wang et al., 2017). In this way, *B. paralicheniformis* may be of biotechnological relevance but still, it has remained largely unexplored.

B. paralicheniformis is a gram-positive, facultatively anaerobic, rod-shaped, motile, and endospore-forming *Bacillus* species (Dunlap et al., 2015). The *B. paralicheniformis* strains are found in a variety of habitats, including soil, freshwater, marine, and niches associated with food (Dunlap et al., 2015; Wang et al., 2017; Othoum et al., 2018). This strain is adapted to survive in extreme conditions such as high osmolarity which provides it with metabolic capabilities similar to industrial strains (Othoum et al., 2018). The *B. paralicheniformis strain Bac84* was isolated from the Red Sea which is an ecosystem of harsh, extremely saline, and high temperature (Othoum et al., 2018). Hence, this strain may be a potential microbial cell factory to produce both thermo-tolerant and osmotolerant enzymes that may be more suitable for use in industry as well as able to survive frequent exposure to these extreme conditions (Nielsen et al., 2017).

The genome of *B. paralicheniformis strain Bac84* has been fully sequenced and published (Othoum et al., 2018). According to the National Center for Biotechnology Information database - NCBI repository, it encodes 4,237 proteins (CP023665.1). However, 414 coding sequences have been anticipated to encode for proteins without any expressional and functional data. These sequences have been assigned as "hypothetical". These hypothetical proteins (HPs) have constituted a considerable portion of the genome. Functional annotation is necessary for these HPs to find the possible roles in the cell which can lead to an understanding of new structures, and functions in this bacterium. Several studies have revealed the expression of HPs (Jagannadham et al., 2011; Jagannadham and Chowdhury, 2012; Ijaq et al., 2020). Homology-based gene annotation has been assigned previously to predict the unknown functions of numerous HPs in several organisms (Doerks et al., 2004; Hawkins and Kihara, 2007; Shahbaaz et al., 2013; Vickers, 2017). Additionally, numerous bioinformatics tools are available to determine the functions of the HPs such as Pfam, InterPro, CATH, SUPERFAMILY, SMART, CDD-BLAST SCANPROSITE, and many more (Gough et al., 2001; Geer et al., 2002; Liu and Karmarkar, 2008; Punta et al., 2012; Shahbaaz et al., 2013; Ijaq et al., 2015). Moreover, the STRING database is also an essential way of protein-protein interaction (PPI) determination to understand the protein functions in a biological network (Jeong et al., 2016; Szklarczyk et al., 2021). Hence, the PPI study of these HPs can lead to inferences about their biological functions (Snider et al., 2015). Furthermore, the tertiary structure modeling through homology searches utilizing the SWISS-MODEL server is important to find the function of unknown proteins (Waterhouse et al., 2018).

In this study, we aimed to determine the functional roles of the HPs from the *B. paralicheniformis strain Bac84*. We utilized an annotation-based workflow to determine the functions of the HPs for the identification of new biotechnologically important proteins as well as novel proteins contributing to the survival of this bacterium in extreme environments. We successfully identified potential target proteins in the *B. paralicheniformis strain Bac84*. It may eventually be possible to develop new biotechnological applications based on further experimental validation of these identified proteins.

RESULTS AND DISCUSSION

Analysis of The Hypothetical Proteins from the *B. Paralicheniformis Strain Bac84* Genome

DNA sequencing technologies are advancing, and high throughput sequencing technologies have allowed a significant number of bacterial genome sequencing. Sequence homology techniques are commonly used for the annotation of genes (Stormo, 2009). Nevertheless, these homology techniques alone are not always able to predict functions accurately and lead to false annotations (Schnoes et al., 2009). Hence, multiple bioinformatic tools must be employed to assign functional annotations of HPs. In this study, we applied a number of effective tools and databases to do the annotation of HPs from the *B. paralicheniformis strain Bac84*.

We first identified the domains of the HPs which are structural, functional, and evolutionary parts of a protein, therefore providing the functional role of a protein (Rao et al., 2014). We extensively analyzed all the 414 HPs sequences using Pfam, InterPro, CATH, SUPERFAMILY, SMART, SCANPROSITE, and CDD-BLAST (Supplementary Table S3). The results were evaluated aiming to assign functions to HPs and it revealed 37 HPs which demonstrated similar functions from three or more programs listed in **Table 1**. In this way, functional annotations were assigned with strong confidence.

Table 1: Hypothetical proteins functionally annotated from the B. paralicheniformis strain Bac84.

No.	**HP ID**	**Inferred function**
1	WP_158700706.1	Metal-dependent hydrolase
2	WP_230368348.1	Catalytic core DNA breaking-rejoining enzymes
3	WP_095290960.1	RNA polymerase sporulation sigma factor SigK
4	WP_026579962.1	YhzD-like protein
5	WP_224146215.1	Response regulator aspartate phosphatase
6	WP_095291534.1	The YqzH-like protein family
7	WP_003179940.1	The YgaB-like protein family
8	WP_020449960.1	Inner membrane protein YiaA-like
9	WP_105981192.1	YqaH-like protein
10	WP_020453622.1	Bacteriophage A118-like, holin
11	WP_006638778.1	Metal-responsive transcriptional regulator
12	WP_003180123.1	Sigma-M inhibitor protein YhdK
13	WP_025810847.1	Streptogramin lyase
14	WP_020450411.1	RlpA-like domain superfamily
15	WP_105980832.1	Phenylalanyl-tRNA synthetase
16	WP_009328837.1	Flavin-phosphopantothenoylcysteine decarboxylase/Flavin prenyltransferase
17	WP_003180732.1	Pathogenicity locus - Putative mitomycin resistance proteins
18	WP_199792123.1	YetA-like protein
19	WP_020451108.1	ESAT-6-like superfamily
20	WP_020451191.1	YkyB-like protein
21	WP_026579751.1	Transcription regulator DksA-related
22	WP_105980957.1	Nudix_Hydrolase super family
23	WP_023857538.1	YhzD-like protein
24	WP_020451915.1	Heat Shock protein (Hsp20 proteins)
25	WP_020452052.1	HesB-like domain superfamily
26	WP_026579290.1	YqfQ-like protein
27	WP_020452371.1	RmlC-like cupin superfamily
28	WP_234026546.1	Chromosome segregation protein SMC
29	WP_023855527.1	Response regulator aspartate phosphatase
30	WP_105981186.1	Putative phage metallopeptidase
31	WP_105981199.1	Alpha/Beta hydrolase fold
32	WP_003185659.1	Swarming motility protein SwrA
33	WP_023857076.1	Acyl-CoA N-acyltransferase
34	WP_023856950.1	BslA (Biofilm surface layer A)
35	WP_026580354.1	Immunity protein WapI-like / YxiJ super family
36	WP_023856884.1	Six-hairpin glycosidase superfamily
37	WP_020453535.1	Prephenate dehydratase

to the HPs. For the rest HPs (n = 377), domains were recognized from less than three mentioned bioinformatic tools which are needed further assessments.

Further, the GO terms were determined using the ARGOT$^{2.5}$ server (Lavezzo et al., 2016) that provides results based on the confidence scores. 133 HPs have GO term predictions among the 414 targets and the distribution among the GO categories was depicted in **Figure 1**. The rest of the HPs with no GO terms can be found in Supplementary Table S5. Among the three categories, the largest cluster was cellular components followed by molecular functions and biological processes. We found seven different GO terminologies in the cellular component category including 45 having membrane function (**Figure 1B**). Although studying membrane proteins is difficult, it is well known that many membrane proteins play important roles in gram-positive bacteria's physiology (Lee et al., 1992; Desvaux et al., 2006). The membrane proteins come first in the interaction among cells and the environmental stresses (Walian et al., 2012). These membrane HPs need to be analyzed as these may have considerable roles in the survival mechanism of the *B. paralicheniformis strain Bac84* in extreme environments. For biological processes, twenty-five different GO terminologies were identified, mostly associated with transcription and DNA-related processes (**Figure 1C**). Transcriptional regulation is a crucial process for a living organism. The cell can respond to intracellular and external signals such as environmental cues or nutritional insufficiency through this transcription-controlling process. According to the GO annotation, the molecular function category showed twenty-one GO terminologies; mostly indicated to several enzymatic functions, and the others related to protein binding (**Figure 1D**). Here, the DNA and protein interactions are involved in many biological processes (Karthik et al., 2014). Additionally, the proteins with enzymatic functions have potential biotechnological applications (Gurung et al., 2013; Cabrera and Blamey, 2018).

Additionally, 15 HPs carried homologous sequences with described functions were found in BlastP analysis whereas the remaining HPs were matched to uncharacterized family proteins and/or hypothetical proteins (Supplementary Table S6). All the 15 HPs that matched with functional proteins in the BlastP analysis were functionally similar to the anticipated functions.

Furthermore, the DEG database was utilized to predict fundamental genes (Supplementary Table S7). This database adapts both the in vitro and in vivo experiments to detect fundamental genes which are essential for cellular machinery (Luo et al., 2021). Though different challenging lab experiments were used to detect the essential genes such as RNA interference, gene knockouts, and transposon mutagenesis (Wei et al., 2013), this DEG database offers an alternative for predicting essential genes. In our analysis, we did not find any essential genes among the targeted 37 HPs.

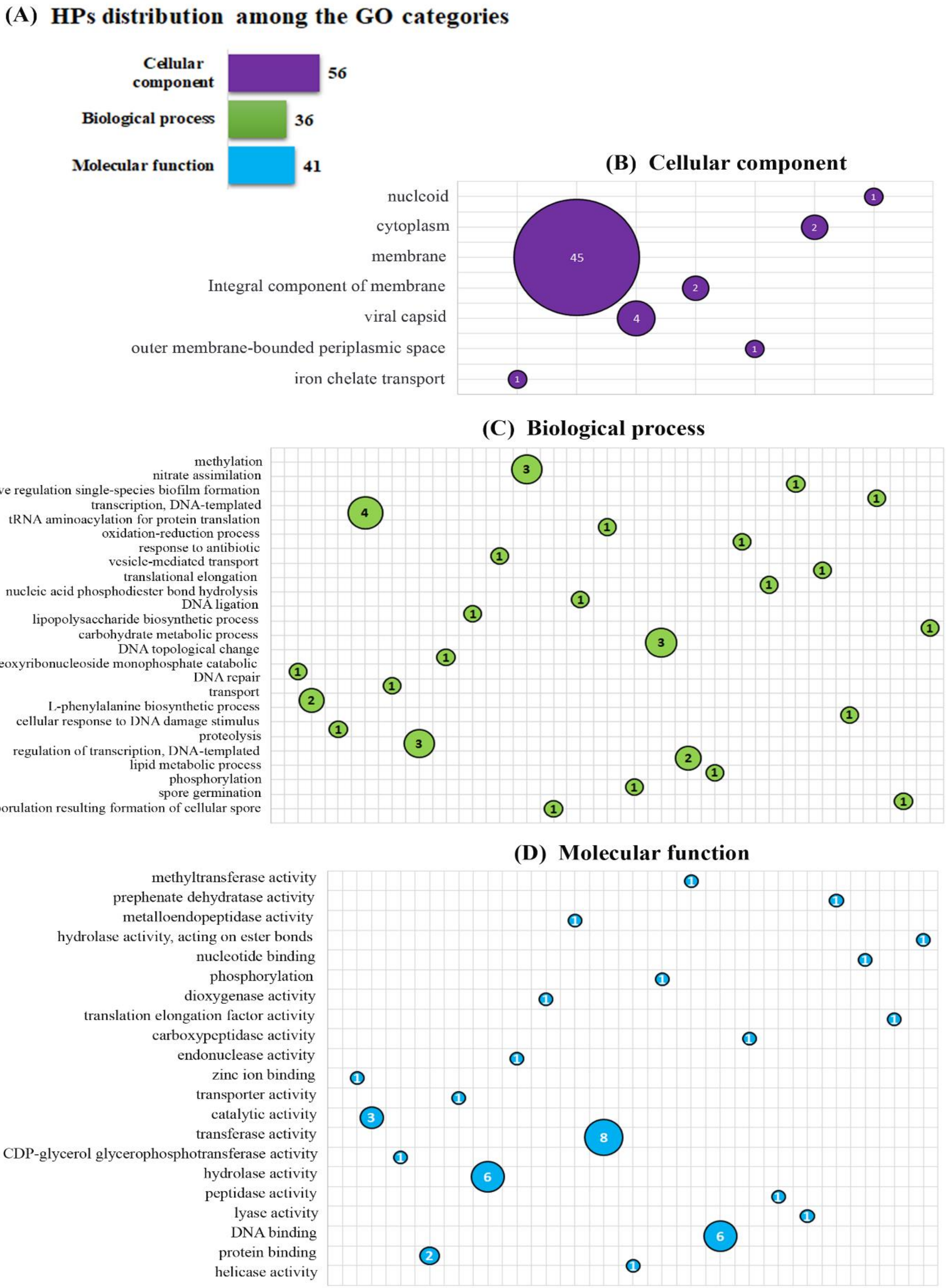

Figure 1: The gene ontology of all the 414 HPs. (A) The distribution of the HPs among the three gene ontology categories. (B) Graph of the cellular components. (C) Graph of the biological processes. (D) Graph of the molecular functions.

Physicochemical Characterization and Subcellular Localization

To evaluate the physicochemical characteristics and their cellular distribution the sequences of the screened 37 HPs were used (Supplementary Table S8). Most of the studied proteins had molecular weight (MW) values over 10000 Da. Proteins with a lower MW (< 10000 Da) need special modifications for analysis in the SDS-PAGE system (Hashimoto et al., 1983). Hence, the first few HPs with lower MW require special attention to perform further lab experiments. The pH value of a protein at which it carries no net electrical charge is known as isoelectric point pI. For our selected HPs, it ranged from 4.4 to 10.48 and 11 proteins have acidic nature (pI < 7), whereas others were found to be basic. Along with the MW, the pI also helps in the laboratory analysis of proteins (da Costa et al., 2018).

The aliphatic index (AI) is used to evaluate the protein thermostability and our HPs were in the range of 55.19-145.1. The range of temperatures at which a protein will be stable increases with increasing AI values (Ikai, 1980). Protein WP_003180123.1, associated with growth and survival after salt stress showed the highest value of 145.1. The instability index (II) was applied to get the idea regarding in vitro protein stability. 15 HPs were considered to be unstable, and 22 HPs were stable. The cut-off values >40 and <40 were used to categorize stable and unstable proteins, respectively (Guruprasad et al., 1990). The GRAVY indicates the interactive nature of a protein with water (Jaspard et al., 2012). Among these 37 HPs, only four (WP_158700706.1; WP_003180123.1; WP_023857538.1 and WP_020453535.1) showed positive values which indicates that these might be hydrophobic.

Moreover, the cellular localization of proteins is vital for their biological functions in a specific environment (Yu et al., 2006; Naqvi et al., 2015). Among the 37 HPs, most of the proteins were determined as cytoplasmic. The cytoplasmic proteins are in the regulation of several functional processes including biosynthesis, regulatory activities, and transport which may help environmental bacteria to compete with the neighboring organisms in the same ecological niche (Nakashima and Nishikawa, 1994). Additionally, we only found 4 proteins to have signal peptides that are critically related to protein secretion (Owji et al., 2018).

Protein-Protein Interactions

To determine the interaction partners of the HPs, we performed a protein-protein interaction analysis (Gazi et al., 2020). In this study, protein WP_095290960.1, RNA polymerase sporulation sigma factor SigK showed a very strong interaction (score 0.930) with the sporulation stage IV protein A (spoIVA) which is involved in sporulation (Roels et al., 1992). WP_006638778.1 interacted with EndoA – a putative RNase (score 0.988) functional as endoribonuclease (Pellegrini et al., 2005). WP_009328837.1 was found to interact with the yacB (score 0.987) which catalyzes the phosphorylation of pantothenate (Brand and Strauss, 2005). The protein WP_023855527.1 showed interaction with the Raca protein which is required for the formation of axial filaments (Schumacher et al., 2016). All these findings along with the other predictions (S9 table and S2 figure) strengthened our functional predictions.

Tertiary Structure Predictions

X-ray crystallography has become a robust approach to determining novel protein structures (Chance et al., 2002). The functional annotation methods in combination with the protein structure analysis are evident to lead to the interpretation of uncharacterized proteins (Ngounou Wetie et al., 2014; Jez, 2017). In this study, we employed the

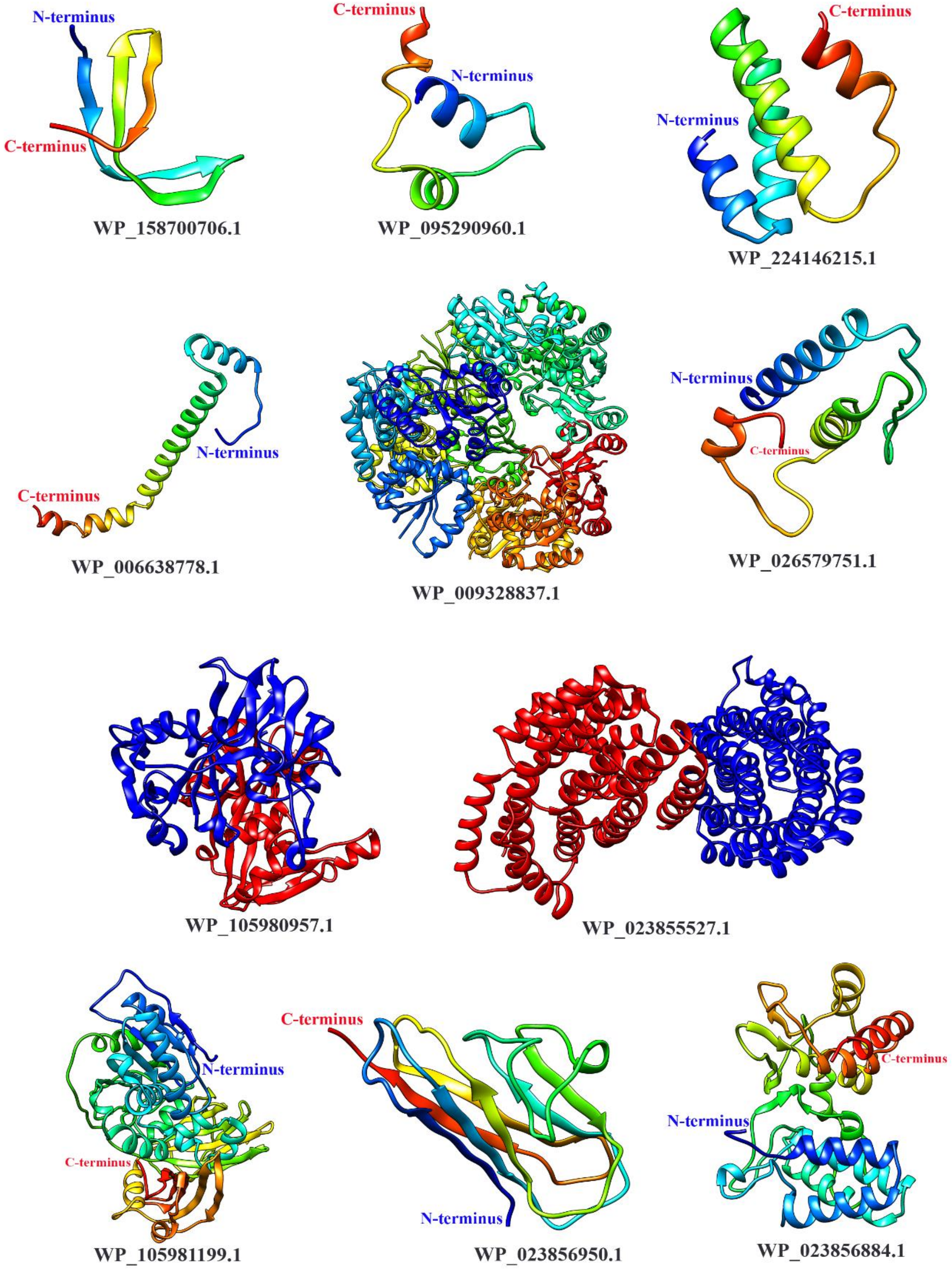

Figure 2: Tertiary structures of eleven proteins.

protein structure homology-modeling server SWISS-MODEL to have the tertiary structures and used the UCSF Chimera software to visualize and present them (**Figure 2**). We successfully build the three-dimensional models for 11 HPs with identity above 30% and the details were listed in the Supplementary Table S10. The structural data collected for several HPs has validated the precise functional annotation. For instance, WP_105981199.1 and WP_023856950.1 showed high identities and resolutions which were functionally annotated as Alpha/Beta hydrolase and BslA (Biofilm surface layer A) respectively. The structures built for these two proteins were determined by X-ray crystallography from two *Bacillus sp.* and those two template proteins have similar functions as we predicted in this study. In this way, proteins with similar sequences usually exhibit similar functions. Proteins dissimilar to current PDB entries may correspond to novel functions. We also checked the quality of the models with the Ramachandran values (Supplementary Table S10) and all the models have an excellent degree of reliability.

ROC Performance Measurement

The availability of genome sequences is increasing which is also allowing more scope to do the computational protein analysis. As these analysis methods are solely dependent on autonomic computing, the accuracy of these methods should be high. The ROC analysis is a broadly applied technique for evaluating the tool's accuracy. The employed pipeline had an average accuracy of 98 percent (**Table 2**), and the ROC analysis's findings supported the strong dependability of the tools used.

Table 2: ROC results of the tools used in this study.

Software	Accuracy (%)	Sensitivity (%)	Specificity (%)	ROC area
Pfam	99.0	98.0	100	0.99
InterPro	100.0	100.0	100.0	1
CATH	100.0	100.0	100.0	1
SUPERFAMILY	96.0	94.7	100.0	0.99
SCANPROSITE	97.0	93.8	100.0	0.99
SMART	98.0	97.0	100.0	1
CDD-BLAST	96.0	65.9	100.0	0.985
Average	**98.00**	**92.77**	**100**	**0.994**

Proteins with Biotechnological Potentials

We found several proteins that can be interesting targets for biotechnological applications.

WP_158700706.1 was predicted as a Metallo-dependent hydrolase (the amidohydrolase superfamily). This group includes numerous hydrolytic enzymes with a varied spectrum of substrates and reactions. The microbial obtained amidohydrolase possesses extensive biotechnological applications that include cosmetics, food, and therapeutics, especially as an anticancer/anti-proliferative agent (Durthi et al., 2020; Patel et al., 2021). This hydrolase group also contains amylases and α-amylase derived from *B. licheniformis*, *B. amyloliquefaciens* and *B. stearothermophilus* has been commercially used in fermentation, paper, and textiles industries (Pandey et al., 2000; Konsoula and Liakopoulou-Kyriakides, 2007).

Protein WP_020453622.1 is a Bacteriophage A118-like, holin that involves the lysis of bacterial membrane (Gründling et al., 2001). These holins can be utilized for controlled pore formation and can promote the release of the desired products. Microorganisms are used and improved for the industrial manufacture of a wide range of substances, including pharmaceuticals and biofuels. These target compounds can be sequestered inside the cell causing toxic effects to the chassis without an efficient active efflux system. In this case, Holin-mediated cell lysis offers an efficient releasing mechanism (Saier Jr and Reddy, 2015). One of the rate-limiting steps is releasing products from the microbial host for biotechnology-based chemical production on an industrial scale. Holins can provide an affordable and effective method of product release in many instances where the use of mechanical disruption or solvent extraction increases the cost of production (Gao et al., 2013). Liu and Curtiss applied phage holin/endolysin cassettes containing a nickel-inducible signal transduction system into the chromosome of *Synechocystis sp.* strain PCC6803 which is being developed for biofuel production (Liu and Curtiss III, 2009). They successfully eliminated the chemical or mechanical removal step by just adding nickel to the culture medium resulting in cell lysis. Another group utilized a light-inducible lytic mechanism in the same cyanobacterium for similar purposes (Miyake et al., 2014). Holins are currently being researched in this manner for numerous biotechnological uses.

The protein WP_009328837.1 was anticipated as Flavin-containing phosphopantothenoylcysteine decarboxylase which is involved in coenzyme A (CoA) biosynthesis (Strauss et al., 2001). CoA is a crucial cofactor involved in many metabolic processes including secondary metabolites production. These distinctive features make CoA an economically significant chemical compound in the cosmetic, and therapeutic industries (Suryatin Alim et al., 2021). Hence, the catalytic abilities of this enzyme make it of immense biotechnological significance.

The protein WP_020452371.1 is in the RmlC-like cupin superfamily and RmlC is a dTDP-sugar isomerase enzyme (dTDP - deoxythymidine diphosphates). This enzyme is involved in the L-rhamnose synthesis, commonly found in bacteria and plants (Kahraman, 1780; Giraud et al., 2000). This sugar getting more interest due to its wide range of substrate specificity and its excellent potential for various unique sugars syntheses such as D-allose, D-cellulose, L-mannose, L rhamnulose, L-spotose, and L-talose (Xu et al., 2016). Besides, rhamnose is combined with lipids to form rhamnolipids that can be used as potential biosurfactants (Kahraman, 1780).

The protein WP_105981199.1 contains an α/β-hydrolase fold that includes proteases, lipases, peroxidases, esterase, epoxide hydrolases, dehalogenases, and many others (Nardini and Dijkstra, 1999). Therefore, this protein can be studied further to uncover its actual functionality as several hydrolases are being used in industrial processes (Gurung et al., 2013). Additionally, an α/β-hydrolase fold protein was also studied which is involved in the cyclic oligopeptide antibiotic ‘thiostrepton’ biosynthesis (Zheng et al., 2016).

The protein WP_023857076.1 carries a structural domain found in numerous acyl-CoA acyltransferases including the N-acetyl transferase (NAT) (Burk, 2003). Several NATs from *Bacillus sp.* Have shown the capability to metabolize xenobiotic compounds that are highly toxic contaminants of groundwater and soils (Garefalaki et al., 2021). This study showed that a class of industrial contaminants or by-products of agrochemicals named “Arylamines” can be converted into less toxic states by *Bacillus* NATs. Hence, our WP_023857076.1 protein should be studied further to find out its bioremediation potential. Additionally, a synthetic N-acetyltransferase (MAT - methionine sulfone N-acetyltransferase) from a bacterial source was utilized to successfully design herbicide “Phosphinothricin” -resistant rice and Arabidopsis (Yun et al., 2009).

Different glycosyltransferases transfer sugar parts from donor molecules to acceptors to form glycosidic bonds and involve in disaccharides, oligosaccharides, and polysaccharides biosynthesis.

Several microbial glycosyltransferases are frequently applied in food processes such as in the shelf-life improvement of bakeries, production of glucose, fructose, or dextrins, lactose hydrolysis, food pectins modification, and many others (Bhatia et al., 2002; Viikari et al., 2007). In our study, protein WP_023856884.1 has the catalytic domain of the Six-hairpin glycosidase superfamily. To use this class of enzymes in different industrial conditions several enzymes functional in alkaline/acidic pH and/or at high temperatures have been discovered from various microorganisms (Thuan and Sohng, 2013; Schröder et al., 2015; Amin et al., 2021). In several studies, bacterial glycosidases were characterized to improve human health and the treatment of different diseases (Liu et al., 2007; Tiels et al., 2012).

The WP_020453535.1 was anticipated to be a prephenate dehydratase that is involved in the biosynthesis of phenylalanine and phenylalanine is an essential amino acid for animals. Recently, the interest in microbial production of L- phenylalanine has increased (Gerigk et al., 2002). It has been widely used in food and feeds as a taste and aroma enhancer, in pharmaceuticals as the drug's building block, as well as used in cosmetics as an ingredient (Sprenger, 2007; Zhou et al., 2010).

Proteins with Adaptational Functions to Extreme Environments

In this study, we identified 12 HPs that may have a significant role for *B. paralicheniformis* in the adaptation to extreme environments.

Sporulation aids bacterial survival in extreme environments by limiting active growth (Huang and Hull, 2017). We found protein WP_095290960.1 as RNA polymerase sporulation sigma factor SigK which is involved in the gene expression controlling during sporulation (Zheng et al., 1992). Similarly, two HPs (WP_224146215.1 and WP_023855527.1) were identified as the response regulator aspartate phosphatase which controls the phosphorelay for sporulation initiation by dephosphorylating Spo0F-P (Parashar et al., 2011). In this way, these HPs can be predicted to play crucial roles in adaption, and survival in extreme environments.

The protein WP_006638778.1 is a metal-responsive transcriptional regulator which can be engaged in the homeostasis and metabolism of any specific metal. These metal-responsive transcriptional regulators allow mechanisms for selective metal ion accumulation and utilization as well as tightly regulate intracellular metal trafficking mechanisms (Finney and O'Halloran, 2003). Metals can be limited in the environment or can be in high amounts that cause toxicity in extreme environments. Hence, a metal-responsive transcriptional regulator protein might be essential to the microorganism for the evolution and adaptation in that specific extreme environment (Musiani et al., 2015). Likewise, WP_026579751.1 is related to the transcription regulator DksA. It is an RNA polymerase-binding transcription factor and is involved in different stress conditions, including nitrosative stress, nutritional shortage, and other environmental stresses (Crawford et al., 2016; Łyżeń et al., 2016). So, this HP can be taken part in extreme environmental adaptations.

We detected a sigma-M inhibitor protein (WP_003180123.1). The sigma-M (yhdM) gene is essential for growth and survival in salt stress conditions (Horsburgh and Moir, 1999). Our predicted Sigma-M inhibitor WP_003180123.1 might play role in salt stress adaptation similarly to a previous study (Yoshimura et al., 2004).

Protein WP_105980957.1 contains a Nudix hydrolase domain that hydrolyzes intracellular nucleotides, regulates their levels, and removes potentially toxic derivatives (Bessman et al., 1996). Some superfamily members can degrade mutagenic, oxidized, and damaged nucleotides that may occur due to exposure to extreme environments (Fisher et al., 2004).

As mentioned earlier, WP_023857076.1 carries a structural domain found in numerous acyl-CoA acyltransferases including- GCN5-related N-acetyltransferases (GNAT) and Glycine N-acyltransferase (Trievel et al., 1999). The proteins from these classes were studied and found to be involved in the adaptation to diverse environmental stress conditions including high salinity, pH tolerance, nutrient stress, etc (Favrot et al., 2016; Dash and Modak, 2021).

Small Heat shock proteins are abundant molecular chaperones that counteract the aggregation of protein upon stress-induced unfolding (Bepperling et al., 2012). We identified protein WP_020451915.1 as a heat shock protein (Hsp20). Several studies showed that Hsp20 responds to different environmental stresses including severe heat, hydrogen peroxide, desiccation, and osmotic shocks (Ventura et al., 2007; Cocotl-Yanez et al., 2014; Singh et al., 2014; Khaskheli et al., 2015). Therefore, WP_020451915.1 might have adaptational functions to extreme environments.

The HesB-like domain is observed in several microbial nitrogen fixation proteins that are associated with FeS-cluster assembly (Zheng et al., 1998). Previous studies found that proteins having a HesB-like domain are involved in different metal resistance and thermal stress conditions (Braz and Marques, 2005; Crapoulet et al., 2006). HesB-like domain-containing protein WP_020452052.1 might also play role in survival in the extreme environment specifically in metal-rich or metal deficient conditions.

The WP_003185659.1 protein was identified as a swarming motility protein SwrA which is a transcription factor. It drives the fla/che operon, which encodes the components of the flagella, and causes swarming motility (Ogura and Tsukahara, 2012). Another study showed that SwrA is involved in bacterial motility (Ghelardi et al., 2012) and bacterial motility might be significant in extreme temperatures (Dall'Agnol et al., 2014).

The WP_023856950.1 protein was predicted as a biofilm surface layer A (BslA) protein which acts as a hydrophobin and participates in biofilm assembly (Kobayashi and Iwano, 2012). Certain microorganisms have great resistance to environmental challenges because of biofilm development (De Carvalho, 2018; Yin et al., 2019; Souza-Egipsy et al., 2021). Therefore, this protein might be crucial for adaptation to harsh environments.

MATERIALS AND METHODS

Sequence Retrieval

The genome of *Bacillus paralicheniformis strain Bac84* was used (CP023665.1). It has 4,376,831 bp in length containing 4413 genes. It encodes 4,237 proteins and 414 are HPs among those (https://www.ncbi.nlm.nih.gov/genome/). The HPs' sequences were obtained in FASTA format for the analyses (Supplementary Table S1).

Functional Annotation of Hypothetical Proteins

Functional annotation was applied to the HPs to reveal their functions (**Figure 3**). Firstly, several publicly available tools and databases (Pfam, InterPro, CATH, SUPERFAMILY, SMART, SCANPROSITE, and CDD-BLAST) are depicted in the Supplementary Table S2 were used. These bioinformatics tools and databases assist to find the conserved domains and afterward categorize the proteins. Pfam (Mistry et al., 2021), InterPro (Blum et al., 2021), SUPERFAMILY (Gough et al., 2001), and SCANPROSITE (De Castro et al., 2006) were employed to interpret the functional roles of the HPs based on similarity. Additionally, SMART and CATH were used to search for functions of our HPs based on the domain architecture and to categorize the domains within the structural hierarchy respectively (Sillitoe et al., 2015; Letunic et al., 2021). Conserved Domain Database (CDD) was utilized to search conserved domains (Lu et al., 2020). All these analyses were performed in the default parameters and the results are given in detail in Supplementary Table S3. These web tools showed distinctive results and to perform downstream analyses, 37 HPs were filtered as these HPs exhibited functional domains or motifs in at least three of the bioinformatic tools (Supplementary Table S4).

We also have predicted the gene ontology of all the HPs using Argot$^{2.5}$ (Annotation Retrieval of Genel Ontology Terms) (Lavezzo et al., 2016) (Supplementary Table S5) and the findings are illustrated in **Figure 1**.

We further used the fasta sequences of the selected 37 HPs for manual annotation utilizing the Basic Local Alignment Search Tool (BLAST) (Johnson et al., 2008). Here, the NCBI nonredundant database and hits with an identity ≥ 90% were employed (Supplementary Table S6). The DEG database was utilized to detect the essential genes with the screened 37 HPs (Luo et al., 2021). The search was performed against the available genomes of *Bacillus subtilis 168*, and *Bacillus thuringiensis BMB171* in the default parameters (Supplementary Table S7).

Prediction of Physicochemical Parameters and the Sub-Cellular Localization

The physicochemical parameters of the selected 37 HPs were theoretically measured using Expasy's Protparam server (Gasteiger et al., 2005). The predicted properties such as molecular mass, isoelectric point (pI), extinction coefficient, the total number of +/- residues, extinction coefficient, instability index, aliphatic index, and grand average of hydropathicity (GRAVY) were determined.

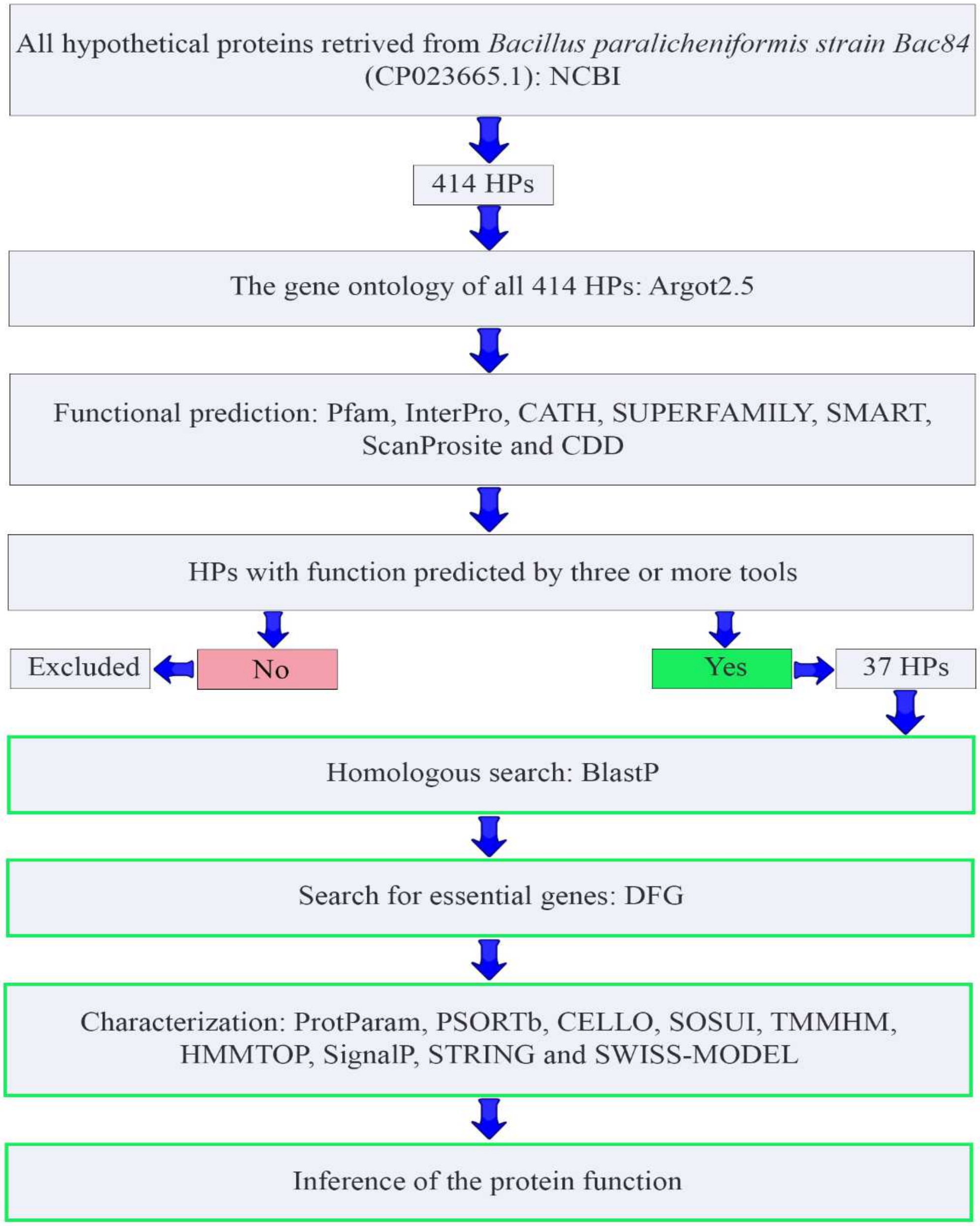
All hypothetical proteins retrived from *Bacillus paralicheniformis strain Bac84* (CP023665.1): NCBI
414 HPs
The gene ontology of all 414 HPs: Argot2.5
Functional prediction: Pfam, InterPro, CATH, SUPERFAMILY, SMART, ScanProsite and CDD
HPs with function predicted by three or more tools
Excluded
No
Yes
37 HPs
Homologous search: BlastP
Search for essential genes: DFG
Characterization: ProtParam, PSORTb, CELLO, SOSUI, TMMHM, HMMTOP, SignalP, STRING and SWISS-MODEL
Inference of the protein function

Figure 3: Workflow representing the overall design of the study.

Determination of the protein cellular localization of a helps to estimate its function. In this study, PSORTb (Yu et al., 2010) and CELLO (Yu et al., 2004) were used to identify the proteins' location in the cell. PSORTb includes both lab experimental data sets as well as in silico predictions. In contrast, CELLO employs a two-level support vector machine (SVM) based system.

Furthermore, SOSUI (Hirokawa et al., 1998), HMMTOP (Tusnady and Simon, 2001), TMHMM (Krogh et al., 2001), and SignalP (Nielsen et al., 2019) were utilized to predict the transmembrane helices as well as determine the presence of signal peptide cleavage sites. All the results of these characterization analyses were listed in the Supplementary Table S8.

Protein-Protein Interaction Analysis

In this study, STRING software (Szklarczyk et al., 2021) was used to predict interactive partners using a confidence score above 0.7 for ensuring the dependability of the predictions (Supplementary Table S9). We had to use the *Bacillus licheniformis DSM 13* reference genome to generate the interaction networks as the dataset for any strain of *B. paralicheniformis* has not been available yet. Both the physical and functional associations were applied to compute the networks. The Cytoscape was used to visualize the interaction networks (**Supplementary Figure S1**).

Tertiary Structure Prediction

Tertiary protein structures give significant insights into the molecular basis of protein function (Schwede et al., 2003). We used the SWISS-MODEL server (Waterhouse et al., 2018) for homology modeling of the target proteins where only templates with an identity ≥ 30% were considered (Supplementary Table S10). The UCSF Chimera-1.16 was used to visualize the 3D structures (**Figure 2**).

Performance Assessment

We performed a ROC- receiver operating characteristic analysis with 100 functionally characterized proteins (Supplementary Table S11) from the genome of the *Bacillus paralicheniformis strain Bac84* to check the accuracy of the anticipated functions of our studied HPs (Swets et al., 2000). These proteins were functionally checked using the seven databases used for our studied HPs.

For the interpretation, the binary numerals "1" and "0" were applied as the true positive and true negative respectively. The integers '2', '3', '4', and '5' were used to assess the prediction efficacy. After that, these datasets were submitted to the Web-based Calculator and calculated the specificity, sensitivity, accuracy, and the ROC area of each tool employed earlier for functional prediction of the HPs (**Table 2**).

CONCLUSIONS

Protein macromolecules are involved in numerous biological processes. Hence, functional annotation of proteins is crucial. An in silico approach was employed in this study to attribute functional annotation of HPs from the *Bacillus paralicheniformis strain Bac84* genome. We functionally annotated 37 HPs from this bacteria. The determination of physicochemical parameters and subcellular localization were effective to understand the specific properties of the annotated proteins. The PPI and tertiary structures of these proteins were also explored which assisted to obtain more understanding of the annotated proteins. We identified several proteins with biotechnological potentials as well as proteins having a high possibility to be involved in extreme environmental adaptation of the *Bacillus paralicheniformis strain Bac84.* Moreover, this strategy provided us with excellent results and it can be utilized to perform the functional annotations of unknown proteins. The combination of such in-silico analysis and lab experiments was successful to obtain functional annotations of HPs from different organisms (Zhang et al., 2006; Choi et al., 2013; Barta et al., 2014). Furthermore, the results also open prospects for further research of this bacterium for biotechnological applications.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories.

FUNDING

No funding sources.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

SUPPLEMENTARY MATERIAL

Supplementary Figure S1 - Protein-protein interaction networks obtained from STRING analysis. Networks are visualized using Cytoscape (v 3.9.1).
Supplementary Table S1 - All the hypothetical proteins from the *Bacillus paralicheniformis strain Bac84*.
Supplementary Table S2 - List of bioinformatics tools and databases used.
Supplementary Table S3 - Annotation dataset results for the 414 hypothetical proteins submitted to the workflow with Pfam, InterPro, CATH, SUPERFAMILY, ScanProsite, SMART, and CDD-Blast.
Supplementary Table S4 - List of selected HPs from the *Bacillus paralicheniformis strain Bac84*.
Supplementary Table S5 - GO terms by Argot$^{2.5}$ for all the HPs.
Supplementary Table S6 - Results of the Blastp search for similar sequences against the non-reduntant (nr) database.
Supplementary Table S7 - Result of essential gene prediction using DEG database.
Supplementary Table S8 - List of predicted physicochemical parameters, sub-cellular localization, and prediction of transmembrane helices for the selected 37 HPs.
Supplementary Table S9 - Protein-protein interactions analyses of the 37 HPs.
Supplementary Table S10 - Tertiary structural information of HPs from *B. Paralichenformis strain Bac84*.
Supplementary Table S11 - Dataset of functional annotation for 100 functionally known proteins from *Bacillus paralicheniformis strain Bac84* using the same pipeline used for the HP prediction.

REFERENCES

Amin, K., Tranchimand, S., Benvegnu, T., Abdel-Razzak, Z., and Chamieh, H. (2021). Glycoside hydrolases and glycosyltransferases from hyperthermophilic archaea: Insights on their characteristics and applications in biotechnology. *Biomolecules* 11(11)**,** 1557.

Barta, M.L., Thomas, K., Yuan, H., Lovell, S., Battaile, K.P., Schramm, V.L., et al. (2014). Structural and biochemical characterization of Chlamydia trachomatis hypothetical protein

CT263 supports that menaquinone synthesis occurs through the futalosine pathway. *Journal of Biological Chemistry* 289(46), 32214-32229.

Bepperling, A., Alte, F., Kriehuber, T., Braun, N., Weinkauf, S., Groll, M., et al. (2012). Alternative bacterial two-component small heat shock protein systems. *Proceedings of the National Academy of Sciences* 109(50), 20407-20412.

Bessman, M.J., Frick, D.N., and O'Handley, S.F. (1996). The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed,"housecleaning" enzymes. *Journal of Biological Chemistry* 271(41), 25059-25062.

Bhatia, Y., Mishra, S., and Bisaria, V.S. (2002). Microbial β-Glucosidases: Cloning, Properties, and Applications. *Critical Reviews in Biotechnology* 22(4), 375-407. doi: 10.1080/07388550290789568.

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., et al. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic acids research* 49(D1), D344-D354.

Brand, L.A., and Strauss, E. (2005). Characterization of a new pantothenate kinase isoform from Helicobacter pylori. *Journal of Biological Chemistry* 280(21), 20185-20188.

Braz, V.S., and Marques, M.V. (2005). Genes involved in cadmium resistance in Caulobacter crescentus. *FEMS Microbiology Letters* 251(2), 289-295.

Burk, D.L. (2003). X-ray structure of the AAC(6')-Ii antibiotic resistance enzyme at 1.8 A resolution; examination of oligomeric arrangements in GNAT superfamily members. *Protein Science* 12(3), 426-437. doi: 10.1110/ps.0233503.

Cabrera, M., and Blamey, J.M. (2018). Biotechnological applications of archaeal enzymes from extreme environments. *Biological research* 51.

Chance, M.R., Bresnick, A.R., Burley, S.K., Jiang, J.-S., Lima, C.D., Sali, A., et al. (2002). Structural genomics: a pipeline for providing structures for the biologist. *Protein science: a publication of the Protein Society* 11(4), 723.

Choi, H.-P., Juarez, S., Ciordia, S., Fernandez, M., Bargiela, R., Albar, J.P., et al. (2013). Biochemical characterization of hypothetical proteins from Helicobacter pylori. *PLoS One* 8(6), e66605.

Cocotl-Yanez, M., Moreno, S., Encarnacion, S., Lopez-Pliego, L., Castaneda, M., and Espín, G. (2014). A small heat-shock protein (Hsp20) regulated by RpoS is essential for cyst desiccation resistance in Azotobacter vinelandii. *Microbiology* 160(3), 479-487.

Crapoulet, N., Barbry, P., Raoult, D., and Renesto, P. (2006). Global transcriptome analysis of Tropheryma whipplei in response to temperature stresses. *Journal of bacteriology* 188(14), 5228-5239.

Crawford, M.A., Henard, C.A., Tapscott, T., Porwollik, S., McClelland, M., and Vázquez-Torres, A. (2016). DksA-dependent transcriptional regulation in Salmonella experiencing nitrosative stress. *Frontiers in microbiology* 7, 444.

da Costa, W.L.O., Araújo, C.L.d.A., Dias, L.M., Pereira, L.C.d.S., Alves, J.T.C., Araujo, F.A., et al. (2018). Functional annotation of hypothetical proteins from the Exiguobacterium antarcticum strain B7 reveals proteins involved in adaptation to extreme environments, including high arsenic resistance. *PloS one* 13(6), e0198965.

Dall'Agnol, H.P., Baraúna, R.A., de Sá, P.H., Ramos, R.T., Nóbrega, F., Nunes, C.I., et al. (2014). Omics profiles used to evaluate the gene expression of Exiguobacterium antarcticum B7 during cold adaptation. *BMC genomics* 15(1), 1-12.

Dash, A., and Modak, R. (2021). Protein acetyltransferases mediate bacterial adaptation to a diverse environment. *Journal of Bacteriology* 203(19)**,** e00231-00221.

De Carvalho, C.C. (2018). Marine biofilms: a successful microbial strategy with economic implications. *Frontiers in marine science* 5**,** 126.

De Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., et al. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic acids research* 34(suppl_2)**,** W362-W365.

Desvaux, M., Dumas, E., Chafsey, I., and Hebraud, M. (2006). Protein cell surface display in Gram-positive bacteria: from single protein to macromolecular protein structure. *FEMS microbiology letters* 256(1)**,** 1-15.

Dhakal, R., Chauhan, K., Seale, R.B., Deeth, H.C., Pillidge, C.J., Powell, I.B., et al. (2013). Genotyping of dairy Bacillus licheniformis isolates by high resolution melt analysis of multiple variable number tandem repeat loci. *Food microbiology* 34(2)**,** 344-351.

Doerks, T., Von Mering, C., and Bork, P. (2004). Functional clues for hypothetical proteins based on genomic context analysis in prokaryotes. *Nucleic acids research* 32(21)**,** 6321-6326.

Du, Y., Ma, J., Yin, Z., Liu, K., Yao, G., Xu, W., et al. (2019). Comparative genomic analysis of Bacillus paralicheniformis MDJK30 with its closely related species reveals an evolutionary relationship between B. paralicheniformis and B. licheniformis. *Bmc Genomics* 20(1)**,** 1-16.

Dunlap, C.A., Kwon, S.-W., Rooney, A.P., and Kim, S.-J. (2015). Bacillus paralicheniformis sp. nov., isolated from fermented soybean paste. *International journal of systematic and evolutionary microbiology* 65(Pt_10)**,** 3487-3492.

Durthi, C.P., Pola, M., Rajulapati, S.B., Kola, A.K., and Kamal, M.A. (2020). Versatile and valuable utilization of amidohydrolase L-glutaminase in pharma and food industries: A review. *Current Drug Metabolism* 21(1)**,** 11-24.

Favrot, L., Blanchard, J.S., and Vergnolle, O. (2016). Bacterial GCN5-related N-acetyltransferases: from resistance to regulation. *Biochemistry* 55(7)**,** 989-1002.

Finney, L.A., and O'Halloran, T.V. (2003). Transition metal speciation in the cell: insights from the chemistry of metal ion receptors. *Science* 300(5621)**,** 931-936.

Fisher, D.I., Cartwright, J.L., Harashima, H., Kamiya, H., and McLennan, A.G. (2004). Characterization of a Nudix hydrolase from Deinococcus radiodurans with a marked specificity for (deoxy) ribonucleoside 5'-diphosphates. *BMC biochemistry* 5(1)**,** 1-8.

Gao, Y., Feng, X., Xian, M., Wang, Q., and Zhao, G. (2013). Inducible cell lysis systems in microbial production of bio-based chemicals. *Applied microbiology and biotechnology* 97(16)**,** 7121-7129.

Garefalaki, V., Papavergi, M.-G., Savvidou, O., Papanikolaou, G., Felföldi, T., Márialigeti, K., et al. (2021). Comparative Investigation of 15 Xenobiotic-Metabolizing N-Acetyltransferase (NAT) Homologs from Bacteria. *Applied and environmental microbiology* 87(19)**,** e0081921-e0081921. doi: 10.1128/AEM.00819-21.

Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M.R., Appel, R.D., and Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. *The proteomics protocols handbook***,** 571-607.

Gazi, M., Mahmud, S., Fahim, S.M., Islam, M., Das, S., Mahfuz, M., et al. (2020). Questing functions and structures of hypothetical proteins from Campylobacter jejuni: a computer-aided approach. *Bioscience reports* 40(6).

Geer, L.Y., Domrachev, M., Lipman, D.J., and Bryant, S.H. (2002). CDART: protein homology by domain architecture. *Genome research* 12(10), 1619-1623.

Gerigk, M., Bujnicki, R., Ganpo-Nkwenkwa, E., Bongaerts, J., Sprenger, G., and Takors, R. (2002). Process control for enhanced L-phenylalanine production using different recombinant Escherichia coli strains. *Biotechnology and bioengineering* 80(7), 746-754.

Ghelardi, E., Salvetti, S., Ceragioli, M., Gueye, S.A., Celandroni, F., and Senesi, S. (2012). Contribution of surfactin and SwrA to flagellin expression, swimming, and surface motility in Bacillus subtilis. *Applied and environmental microbiology* 78(18), 6540-6544.

Giraud, M.-F., Leonard, G.A., Field, R.A., Berlind, C., and Naismith, J.H. (2000). RmlC, the third enzyme of dTDP-L-rhamnose pathway, is a new class of epimerase. *Nature structural biology* 7(5), 398-402.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of molecular biology* 313(4), 903-919.

Gründling, A., Manson, M.D., and Young, R. (2001). Holins kill without warning. *Proceedings of the National Academy of Sciences* 98(16), 9348-9352.

Gurung, N., Ray, S., Bose, S., and Rai, V. (2013). A broader view: microbial enzymes and their relevance in industries, medicine, and beyond. *BioMed research international* 2013.

Guruprasad, K., Reddy, B.B., and Pandit, M.W. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Engineering, Design and Selection* 4(2), 155-161.

Hashimoto, F., Horigome, T., Kanbayashi, M., Yoshida, K., and Sugano, H. (1983). An improved method for separation of low-molecular-weight polypeptides by electrophoresis in sodium dodecyl sulfate-polyacrylamide gel. *Analytical Biochemistry* 129(1), 192-199.

Hawkins, T., and Kihara, D. (2007). Function prediction of uncharacterized proteins. *Journal of bioinformatics and computational biology* 5(01), 1-30.

Hirokawa, T., Boon-Chieng, S., and Mitaku, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics (Oxford, England)* 14(4), 378-379.

Horsburgh, M.J., and Moir, A. (1999). σM, an ECF RNA polymerase sigma factor of Bacillus subtilis 168, is essential for growth and survival in high concentrations of salt. *Molecular microbiology* 32(1), 41-50.

Huang, M., and Hull, C.M. (2017). Sporulation: how to survive on planet Earth (and beyond). *Current genetics* 63(5), 831-838.

Ijaq, J., Bethi, N., and Jagannadham, M. (2020). Mass spectrometry-based identification and characterization of human hypothetical proteins highlighting the inconsistency across the protein databases. *Journal of Proteins and Proteomics* 11(1), 17-25.

Ijaq, J., Chandrasekharan, M., Poddar, R., Bethi, N., and Sundararajan, V.S. (2015). Annotation and curation of uncharacterized proteins-challenges. *Frontiers in genetics* 6, 119.

Ikai, A. (1980). Thermostability and aliphatic index of globular proteins. *The Journal of Biochemistry* 88(6), 1895-1898.

Jagannadham, M., Abou-Eladab, E.F., and Kulkarni, H.M. (2011). Identification of outer membrane proteins from an Antarctic bacterium Pseudomonas syringae Lz4W. *Molecular & Cellular Proteomics* 10(6).

Jagannadham, M.V., and Chowdhury, C. (2012). Differential expression of membrane proteins helps Antarctic Pseudomonas syringae to acclimatize upon temperature variations. *Journal of proteomics* 75(8)**,** 2488-2499.

Jaspard, E., Macherel, D., and Hunault, G. (2012). Computational and statistical analyses of amino acid usage and physico-chemical properties of the twelve late embryogenesis abundant protein classes. *PloS one* 7(5)**,** e36968.

Jeong, H., Qian, X., and Yoon, B.-J. (Year). "Effective comparative analysis of protein-protein interaction networks by measuring the steady-state network flow using a Markov model", in: *BMC bioinformatics*: BioMed Central), 15-27.

Jez, J.M. (2017). Revisiting protein structure, function, and evolution in the genomic era. *Journal of invertebrate pathology* 142**,** 11-15.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. (2008). NCBI BLAST: a better web interface. *Nucleic acids research* 36(suppl_2)**,** W5-W9.

Kahraman, H. (1780). The Importance of L-Rhamnose Sugar. *Biochimica et Biophysica Acta* 12**,** 1388-1394.

Karthik, L., Kumar, G., Keswani, T., Bhattacharyya, A., Chandar, S.S., and Bhaskara Rao, K. (2014). Protease inhibitors from marine actinobacteria as a potential source for antimalarial compound. *PloS one* 9(3)**,** e90972.

Khaskheli, G.B., Zuo, F., Yu, R., and Chen, S. (2015). Overexpression of small heat shock protein enhances heat-and salt-stress tolerance of Bifidobacterium longum NCC2705. *Current Microbiology* 71(1)**,** 8-15.

Kobayashi, K., and Iwano, M. (2012). BslA (YuaB) forms a hydrophobic layer on the surface of Bacillus subtilis biofilms. *Molecular microbiology* 85(1)**,** 51-66.

Konsoula, Z., and Liakopoulou-Kyriakides, M. (2007). Co-production of α-amylase and β-galactosidase by Bacillus subtilis in complex organic substrates. *Bioresource Technology* 98(1)**,** 150-157.

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology* 305(3)**,** 567-580.

Lavezzo, E., Falda, M., Fontana, P., Bianco, L., and Toppo, S. (2016). Enhancing protein function prediction with taxonomic constraints–The Argot2. 5 web server. *Methods* 93**,** 15-23.

Lee, B.-Y., Hefta, S., and Brennan, P. (1992). Characterization of the major membrane protein of virulent Mycobacterium tuberculosis. *Infection and immunity* 60(5)**,** 2066-2074.

Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: recent updates, new developments and status in 2020. *Nucleic acids research* 49(D1)**,** D458-D460.

Liu, Q.P., Sulzenbacher, G., Yuan, H., Bennett, E.P., Pietz, G., Saunders, K., et al. (2007). Bacterial glycosidases for the production of universal red blood cells. *Nature biotechnology* 25(4)**,** 454-464.

Liu, X., and Curtiss III, R. (2009). Nickel-inducible lysis system in Synechocystis sp. PCC 6803. *Proceedings of the National Academy of Sciences* 106(51)**,** 21550-21554.

Liu, Z., and Karmarkar, V. (2008). Groucho/Tup1 family co-repressors in plant development. *Trends in plant science* 13(3)**,** 137-144.

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., et al. (2020). CDD/SPARCLE: the conserved domain database in 2020. *Nucleic acids research* 48(D1)**,** D265-D268.

Luo, H., Lin, Y., Liu, T., Lai, F.-L., Zhang, C.-T., Gao, F., et al. (2021). DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. *Nucleic acids research* 49(D1), D677-D686.

Łyżeń, R., Maitra, A., Milewska, K., Kochanowska-Łyżeń, M., Hernandez, V.J., and Szalewska-Pałasz, A. (2016). The dual role of DksA protein in the regulation of Escherichia coli pArgX promoter. *Nucleic Acids Research* 44(21), 10316-10325.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L., et al. (2021). Pfam: The protein families database in 2021. *Nucleic acids research* 49(D1), D412-D419.

Miyake, K., Abe, K., Ferri, S., Nakajima, M., Nakamura, M., Yoshida, W., et al. (2014). A green-light inducible lytic system for cyanobacterial cells. *Biotechnology for biofuels* 7(1), 1-8.

Musiani, F., Zambelli, B., Bazzani, M., Mazzei, L., and Ciurli, S. (2015). Nickel-responsive transcriptional regulators. *Metallomics* 7(9), 1305-1318.

Nakashima, H., and Nishikawa, K. (1994). Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *Journal of molecular biology* 238(1), 54-61.

Naqvi, A.A.T., Shahbaaz, M., Ahmad, F., and Hassan, M.I. (2015). Identification of functional candidates amongst hypothetical proteins of Treponema pallidum ssp. pallidum. *PloS one* 10(4), e0124177.

Nardini, M., and Dijkstra, B.W. (1999). α/β Hydrolase fold enzymes: the family keeps growing. *Current opinion in structural biology* 9(6), 732-737.

Ngounou Wetie, A.G., Sokolowska, I., Woods, A.G., Roy, U., Deinhardt, K., and Darie, C.C. (2014). Protein–protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. *Cellular and molecular life sciences* 71(2), 205-228.

Nielsen, H., Tsirigos, K.D., Brunak, S., and von Heijne, G. (2019). A brief history of protein sorting prediction. *The protein journal* 38(3), 200-216.

Nielsen, J., Archer, J., Essack, M., Bajic, V.B., Gojobori, T., and Mijakovic, I. (2017). Building a bio-based industry in the Middle East through harnessing the potential of the Red Sea biodiversity. *Applied Microbiology and Biotechnology* 101(12), 4837-4851.

Ogura, M., and Tsukahara, K. (2012). SwrA regulates assembly of Bacillus subtilis DegU via its interaction with N-terminal domain of DegU. *The journal of biochemistry* 151(6), 643-655.

Othoum, G., Bougouffa, S., Razali, R., Bokhari, A., Alamoudi, S., Antunes, A., et al. (2018). In silico exploration of Red Sea Bacillus genomes for natural product biosynthetic gene clusters. *BMC genomics* 19(1), 1-11.

Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A., and Ghasemi, Y. (2018). A comprehensive review of signal peptides: Structure, roles, and applications. *European journal of cell biology* 97(6), 422-441.

Pandey, A., Nigam, P., Soccol, C.R., Soccol, V.T., Singh, D., and Mohan, R. (2000). Advances in microbial amylases. *Biotechnology and applied biochemistry* 31(2), 135-152.

Parashar, V., Mirouze, N., Dubnau, D.A., and Neiditch, M.B. (2011). Structural basis of response regulator dephosphorylation by Rap phosphatases. *PLoS biology* 9(2), e1000589.

Patel, N.Y., Baria, D.M., Yagnik, S.M., Rajput, K.N., Panchal, R.R., and Raval, V.H. (2021). Bio-prospecting the future in perspective of amidohydrolase L-glutaminase from marine habitats. *Applied Microbiology and Biotechnology* 105(13), 5325-5340.

Pellegrini, O., Mathy, N., Gogos, A., Shapiro, L., and Condon, C. (2005). The Bacillus subtilis ydcDE operon encodes an endoribonuclease of the MazF/PemK family and its inhibitor. *Molecular microbiology* 56(5), 1139-1148.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012). The Pfam protein families database. *Nucleic acids research* 40(D1), D290-D301.

Rao, V.S., Srinivas, K., Sujini, G., and Kumar, G. (2014). Protein-protein interaction detection: methods and analysis. *International journal of proteomics* 2014.

Rey, M.W., Ramaiya, P., Nelson, B.A., Brody-Karpin, S.D., Zaretsky, E.J., Tang, M., et al. (2004). Complete genome sequence of the industrial bacterium Bacillus licheniformis and comparisons with closely related Bacillusspecies. *Genome biology* 5(10), 1-12.

Roels, S., Driks, A., and Losick, R. (1992). Characterization of spoIVA, a sporulation gene involved in coat morphogenesis in Bacillus subtilis. *Journal of bacteriology* 174(2), 575-585.

Saier Jr, M.H., and Reddy, B.L. (2015). Holins in bacteria, eukaryotes, and archaea: multifunctional xenologues with potential biotechnological and biomedical applications. *Journal of bacteriology* 197(1), 7-17.

Schnoes, A.M., Brown, S.D., Dodevski, I., and Babbitt, P.C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology* 5(12), e1000605.

Schröder, C., Blank, S., and Antranikian, G. (2015). First glycoside hydrolase family 2 enzymes from Thermus antranikianii and Thermus brockianus with β-glucosidase activity. *Frontiers in bioengineering and biotechnology* 3, 76.

Schumacher, M.A., Lee, J., and Zeng, W. (2016). Molecular insights into DNA binding and anchoring by the Bacillus subtilis sporulation kinetochore-like RacA protein. *Nucleic Acids Res* 44(11), 5438-5449. doi: 10.1093/nar/gkw248.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic acids research* 31(13), 3381-3385.

Shahbaaz, M., ImtaiyazHassan, M., and Ahmad, F. (2013). Functional annotation of conserved hypothetical proteins from Haemophilus influenzae Rd KW20. *PloS one* 8(12), e84263.

Sillitoe, I., Lewis, T.E., Cuff, A., Das, S., Ashford, P., Dawson, N.L., et al. (2015). CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic acids research* 43(D1), D376-D381.

Singh, H., Appukuttan, D., and Lim, S. (2014). Hsp20, a small heat shock protein of Deinococcus radiodurans, confers tolerance to hydrogen peroxide in Escherichia coli. *Journal of Microbiology and Biotechnology* 24(8), 1118-1122.

Snider, J., Kotlyar, M., Saraon, P., Yao, Z., Jurisica, I., and Stagljar, I. (2015). Fundamentals of protein interaction network mapping. *Molecular systems biology* 11(12), 848.

Souza-Egipsy, V., Vega, J.F., González-Toril, E., and Aguilera, Á. (2021). Biofilm mechanics in an extremely acidic environment: microbiological significance. *Soft Matter* 17(13), 3672-3680.

Sprenger, G.A. (2007). From scratch to value: engineering Escherichia coli wild type cells to the production of L-phenylalanine and other fine chemicals derived from chorismate. *Applied microbiology and biotechnology* 75(4), 739-749.

Stormo, G.D. (2009). An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics* 27(1), 3.1. 1-3.1. 7.

Strauss, E., Kinsland, C., Ge, Y., McLafferty, F.W., and Begley, T.P. (2001). Phosphopantothenoylcysteine synthetase from Escherichia coli: identification and characterization of the last unidentified coenzyme A biosynthetic enzyme in bacteria. *Journal of Biological Chemistry* 276(17)**,** 13513-13516.

Suryatin Alim, G., Iwatani, T., Okano, K., Kitani, S., and Honda, K. (2021). In vitro production of coenzyme A using thermophilic enzymes. *Applied and environmental microbiology* 87(14)**,** e00541-00521.

Swets, J.A., Dawes, R.M., and Monahan, J. (2000). Better decisions through science. *Scientific American* 283(4)**,** 82-87.

Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research* 49(D1)**,** D605-D612.

Thuan, N.H., and Sohng, J.K. (2013). Recent biotechnological progress in enzymatic synthesis of glycosides. *Journal of Industrial Microbiology and Biotechnology* 40(12)**,** 1329-1356.

Tiels, P., Baranova, E., Piens, K., De Visscher, C., Pynaert, G., Nerinckx, W., et al. (2012). A bacterial glycosidase enables mannose-6-phosphate modification and improved cellular uptake of yeast-produced recombinant human lysosomal enzymes. *Nature biotechnology* 30(12)**,** 1225-1231.

Trievel, R.C., Rojas, J.R., Sterner, D.E., Venkataramani, R.N., Wang, L., Zhou, J., et al. (1999). Crystal structure and mechanism of histone acetylation of the yeast GCN5 transcriptional coactivator. *Proceedings of the National Academy of Sciences* 96(16)**,** 8931-8936.

Tusnady, G.E., and Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics* 17(9)**,** 849-850.

Ventura, M., Canchaya, C., Zhang, Z., Fitzgerald, G.F., and van Sinderen, D. (2007). Molecular characterization of hsp20, encoding a small heat shock protein of Bifidobacterium breve UCC2003. *Applied and environmental microbiology* 73(14)**,** 4695-4703.

Vickers, N.J. (2017). Animal communication: when i'm calling you, will you answer too? *Current biology* 27(14)**,** R713-R715.

Viikari, L., Alapuranen, M., Puranen, T., Vehmaanperä, J., and Siika-Aho, M. (2007). Biofuels. *Advances in biochemical engineering/biotechnology* 108.

Walian, P.J., Allen, S., Shatsky, M., Zeng, L., Szakal, E.D., Liu, H., et al. (2012). High-throughput isolation and characterization of untagged membrane protein complexes: outer membrane complexes of Desulfovibrio vulgaris. *Journal of proteome research* 11(12)**,** 5720-5735.

Wang, Y., Liu, H., Liu, K., Wang, C., Ma, H., Li, Y., et al. (2017). Complete genome sequence of Bacillus paralicheniformis MDJK30, a plant growth-promoting rhizobacterium with antifungal activity. *Genome Announcements* 5(25)**,** e00577-00517.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research* 46(W1)**,** W296-W303.

Wei, W., Ning, L.-W., Ye, Y.-N., and Guo, F.-B. (2013). Geptop: a gene essentiality prediction tool for sequenced bacterial genomes based on orthology and phylogeny. *PloS one* 8(8)**,** e72343.

Xu, W., Zhang, W., Zhang, T., Jiang, B., and Mu, W. (2016). L-Rhamnose isomerase and its use for biotechnological production of rare sugars. *Applied microbiology and biotechnology* 100(7)**,** 2985-2992.

Yin, W., Wang, Y., Liu, L., and He, J. (2019). Biofilms: the microbial "protective clothing" in extreme environments. *International journal of molecular sciences* 20(14)**,** 3423.

Yoshimura, M., Asai, K., Sadaie, Y., and Yoshikawa, H. (2004). Interaction of Bacillus subtilis extracytoplasmic function (ECF) sigma factors with the N-terminal regions of their potential anti-sigma factors. *Microbiology* 150(3)**,** 591-599.

Yu, C.S., Chen, Y.C., Lu, C.H., and Hwang, J.K. (2006). Prediction of protein subcellular localization. *Proteins: Structure, Function, and Bioinformatics* 64(3)**,** 643-651.

Yu, C.S., Lin, C.J., and Hwang, J.K. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein science* 13(5)**,** 1402-1406.

Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., et al. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26(13)**,** 1608-1615.

Yun, C.-S., Hasegawa, H., Nanamiya, H., Terakawa, T., and Tozawa, Y. (2009). Novel Bacterial<i>N</i>-Acetyltransferase Gene for Herbicide Detoxification in Land Plants and Selection Maker in Plant Transformation. *Bioscience, Biotechnology, and Biochemistry* 73(5)**,** 1000-1006. doi: 10.1271/bbb.80777.

Zhang, W., Culley, D.E., Gritsenko, M.A., Moore, R.J., Nie, L., Scholten, J.C., et al. (2006). LC–MS/MS based proteomic analysis and functional inference of hypothetical proteins in Desulfovibrio vulgaris. *Biochemical and biophysical research communications* 349(4)**,** 1412-1419.

Zheng, L., Cash, V.L., Flint, D.H., and Dean, D.R. (1998). Assembly of iron-sulfur clusters: identification of an iscSUA-hscBA-fdx gene cluster from Azotobacter vinelandii. *Journal of Biological Chemistry* 273(21)**,** 13264-13272.

Zheng, L., Halberg, R., Roels, S., Ichikawa, H., Kroos, L., and Losick, R. (1992). Sporulation regulatory protein GerE from Bacillus subtilis binds to and can activate or repress transcription from promoters for mother-cell-specific genes. *Journal of molecular biology* 226(4)**,** 1037-1050.

Zheng, Q., Wang, S., Duan, P., Liao, R., Chen, D., and Liu, W. (2016). An α/β-hydrolase fold protein in the biosynthesis of thiostrepton exhibits a dual activity for endopeptidyl hydrolysis and epoxide ring-opening/macrocyclization. . *Proceedings of the National Academy of Sciences* 113(50)**,** 14318-14323. doi: doi:10.1073/pnas.1612607113.

Zhou, H., Liao, X., Wang, T., Du, G., and Chen, J. (2010). Enhanced L-phenylalanine biosynthesis by co-expression of pheAfbr and aroFwt. *Bioresource technology* 101(11)**,** 4151-4156.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigureS1.pdf
- SupplementaryTableS1.xlsx
- SupplementaryTableS2.xlsx
- SupplementaryTableS3.xlsx
- SupplementaryTableS4.xlsx
- SupplementaryTableS5.xlsx
- SupplementaryTableS6.xlsx
- SupplementaryTableS7.xlsx
- SupplementaryTableS8.xlsx
- SupplementaryTableS9.xlsx
- SupplementaryTableS10.xlsx
- SupplementaryTableS11.xlsx