Hindawi Publishing Corporation
Comparative and Functional Genomics
Volume 2007, Article ID 47161, 23 pages
doi:10.1155/2007/47161

Research Article

Identification and Analysis of Novel Amino-Acid Sequence Repeats in *Bacillus anthracis* str. *Ames* Proteome Using Computational Tools

G. R. Hemalatha, D. Satyanarayana Rao, and L. Guruprasad

School of Chemistry, University of Hyderabad, Hyderabad 500 046, India

Received 7 August 2006; Revised 6 December 2006; Accepted 9 December 2006

Recommended by Stephen Oliver

We have identified four repeats and ten domains that are novel in proteins encoded by the *Bacillus anthracis* str. *Ames* proteome using automated in silico methods. A "repeat" corresponds to a region comprising less than 55-amino-acid residues that occur more than once in the protein sequence and sometimes present in tandem. A "domain" corresponds to a conserved region with greater than 55-amino-acid residues and may be present as single or multiple copies in the protein sequence. These correspond to (1) 57-amino-acid-residue PxV domain, (2) 122-amino-acid-residue FxF domain, (3) 111-amino-acid-residue YEFF domain, (4) 109-amino-acid-residue IMxxH domain, (5) 103-amino-acid-residue VxxT domain, (6) 84-amino-acid-residue ExW domain, (7) 104-amino-acid-residue NTGFIG domain, (8) 36-amino-acid-residue NxGK repeat, (9) 95-amino-acid-residue VYV domain, (10) 75-amino-acid-residue KEWE domain, (11) 59-amino-acid-residue AFL domain, (12) 53-amino-acid-residue RIDVK repeat, (13) (a) 41-amino-acid-residue AGQF repeat and (b) 42-amino-acid-residue GSAL repeat. A repeat or domain type is characterized by specific conserved sequence motifs. We discuss the presence of these repeats and domains in proteins from other genomes and their probable secondary structure.

1. INTRODUCTION

The anthrax is a disease of herbivores and other mammals including humans, caused by the *Bacillus anthracis* str. *Ames*, a Gram-positive, rod-shaped, nonmotile, spore-forming bacterium [1]. It is an endospore-forming bacterium that causes inhalational anthrax. During the course of disease, endospores are taken up by alveolar macrophages where they germinate in the phagolysosomal compartment. Vegetative cells then escape from the macrophage, eventually infecting blood. Expression of the major plasmid-encoded virulence determinants, tripartite toxin, and a poly-D-glutamic acid capsule is essential for full pathogenicity [2]. Key virulence genes found on plasmids are pXO1 and pXO2 [1]. The 60 MDa plasmid pXO2 carries genes required for the synthesis of an antiphagocytic poly-D-glutamic acid capsule [3]. The 110 MDa plasmid pXO1 [4] is required for the synthesis of the anthrax proteins, edema factor, lethal factor, and protective antigen. These proteins act in binary combinations to produce two anthrax toxins: edema toxin (a protective antigen and edema factor) and lethal toxin (a protective antigen and lethal factor) [5]. The chromosome encodes potential virulence factors that include haemolysins, enterotoxins, phospholipases, proteases, metalloproteases, and iron-acquisition proteins.

The chromosome of *B. anthracis* str. *Ames* contains three homologues of sortase transpeptidase that is responsible for attachment of secreted proteins to peptidoglycan on the cell surface of Gram-positive bacteria [6]. A range of important surface proteins, including enzymes and virulence-related MSCRAMMs (microbial surface components recognizing adhesive matrix molecules) are anchored to the cell wall in Gram-positive bacteria by sortase, a transpeptidase in *Staphylococcus aureus*, that cleaves polypeptides at a conserved LPxTG motif near the carboxyl terminus and covalently links them to penta-glycine crossbridges in peptidoglycan [7, 8]. Nearly 34 candidate surface proteins which have sortase attachment sites and SLH domains were identified. Two putative *B. anthracis* str. *Ames* sortase attached genes have internalin like repeats [9]. The chromosome of

B. anthracis str. *Ames* also contains the *csaAB* genes for binding of proteins with S-layer homology (SLH) domains to polysaccharide. The SLH domain is a repetitive modular element that is present in several bacterial cell surface proteins and is involved in noncovalent association with peptidoglycan associated polymers [10]. The SLH domain comprises 55-amino-acid residues [11] and the potential role of most proteins with SLH domains on the surface of *B. anthracis* str. *Ames* is unknown at present [12]. However, these surface proteins may mediate unknown interactions between *B. anthracis* str. *Ames* and its external environment and could be targets for vaccine and drug design. Read et al. [12] reported the complete genome sequence of *B. anthracis* str. *Ames*. It comprises 5 227 293 base pairs and 5508 genes with an overall G+C content of 35.4%. Of these, 2762 are functional genes, 1212 are conserved hypothetical genes, 657 genes are of unknown function, and 877 genes are annotated as hypothetical proteins.

As the complete genome sequence of *B. anthracis* str. *Ames* is available [12], we intended to systematically identify and analyze all the amino-acid sequence repeats in this proteome. In a general context, a "repeat" corresponds to a region comprising less than 55-amino-acid residues that occur more than once, sometimes in tandem along the primary sequence, examples are the YVTN repeats in various cell surface proteins and the WD repeats present in proteins that perform a variety of functions. On the other hand, a "domain" refers to a region of the protein comprising greater than 55-amino-acid residues and does not contain internal sequence repeats. According to the crystallographer definition, a domain represents a region of the protein capable of folding independently as a stable unit. A domain can also exist in multiple copies and there can be several different domains per protein, examples are the SH2, SH3, and PH domains present in signal transduction proteins. The repeats and domains are characterized by conserved sequence motifs that may be identified according to the conservation of individual amino-acid residues at equivalent positions derived from multiple sequence alignments. In the absence of experimental data, the structural information can be obtained from secondary structure or fold prediction studies in silico. Information about the identified domains and repeats is represented in databases such as SMART, INTERPRO and PFAM. SMART (simple modular architecture research tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures [13]. INTERPRO is a searchable database that provides information on sequence, function, and annotation. It is an integrated documentation resource for protein families, domains, and sites [14]. PFAM is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. This can be used to view the domain organization of proteins [15]. We believe that a systematic sequence analysis will provide information on the novel repeats and domains present in *B. anthracis* str. *Ames* proteome that are not identified so far.

The *B. anthracis* str. *Ames* proteome consists of several known repeats and domains. Some of these domains are as follows. (1) BRCT (breast cancer carboxy terminal) domain was first identified as 100-amino-acid tandem repeat at the C-terminus of the tumor suppressor gene product BRCA1, in which the germline mutations lead to nearly 50% familial breast cancer. Most BRCT domains containing proteins participate in DNA damage checkpoint or DNA repair pathways and transcription regulation [16]. The BRCT is an evolutionarily conserved module that exists in a large number of proteins from prokaryotes to eukaryotes. (2) Excalibur (extracellular calcium binding) domain consists of a conserved DxDxDGxxCE motif, which is strikingly similar to the Ca^{2+} binding loop of the calmodulin like EF hand domains, suggesting an evolutionary relationship. (3) Cna_B domain forms a stalk in *Streptococcus aureus* collagen-binding protein that presents the ligand binding domain away from the bacterial cell surface. (4) CBS (cystathionine beta synthase) domain is a small intracellular module with 60-amino-acid residues, mostly found in two or four copies within a protein and occurs in several proteins in all kingdoms of life. Tandem pairs of CBS domains can act as binding domains for adenosine derivatives. In some cases, CBS domains may act as sensors of cellular energy status by being activated by AMP and inhibited by ATP. (5) Par B (par B like nuclease) domain cleaves single stranded DNA, nicks supercoiled plasmid DNA, and exhibits $5'$-$3'$ exonuclease activity. (6) KH (K homology) domain comprises 70-amino-acids residues and is involved in RNA binding. (7) PAS and PAC domains comprising 300 and 45-amino-acid residues, respectively, mediate signal transduction. (8) PASTA domain is an extracellular module comprising 70-amino-acids residues that fold into a globular architecture consisting of 3β-strands and an α-helix which aids in penicillin binding. (9) NEAT (near transporter) domain is a 125-amino-acid residue conserved region consisting mainly β-strands. The NEAT domain appears to be associated with iron transport in several Gram-positive species, some of them are pathogenic. (10) SLH domain is present in several bacterial cell surface proteins and is involved in noncovalent association with peptidoglycan associated polymers. It comprises 55-amino-acid residues and the predicted secondary structure comprises two α-helices flanking a short β-strand [11].

The repeats present in *B. anthracis* str. *Ames* proteome are as follows. (1) RHS repeats are 21-amino-acids residues long and are involved in carbohydrate binding. (2) TPR (tetratricopeptide) repeats are 34-amino-acids residues long and are involved in protein-protein interactions. (3) EZ_HEAT repeats are 37–47-amino-acid residues long and occur in tandem in a number of cytoplasmic proteins that are involved in intracellular transport processes. Arrays of HEAT repeats consist of 3 to 36 units forming a rod-like helical structure and appear to function as protein-protein interaction surfaces. (4) Ankyrin repeats are about 33-amino-acid residues long and occur in at least four consecutive copies; the core of the repeat appears as a helix-loop-helix structure and is involved in protein-protein interactions. (5) LRR (lecuine rich repeats) are 20-amino-acids residues long, each repeat consists of a β-strand and α-helix, that are oriented in an antiparallel manner. The function of LRRs includes signal transduction, transmembrane receptors, DNA repair, cell adhesion, and extracelluarl matrix proteins [17].

Andrade et al. [18] reviewed methods to identify repeats in proteins and the relationship between repeat sequences and their associated functions. Repeats may be identified by manual examination, if the sequence similarity is very high and present in tandem. Repeats are thought to arise due to gene duplication and recombination events. Protein domains may exist either as single or multiple copies and repeats always exist as multiple copies [18, 19]. Programs such as BLASTP [20] are also useful in detecting internal and homologous repeats in a protein database. By using the BLAST program, the presence of repeats in a query protein sequence can be identified if (a) the same region of the query is aligned against two or more distinct regions of a second protein; and (b) different regions of the query are being aligned against the same region of a second protein [18].

Several web-based methods are available for ab initio identification of sequence repeats in proteins. For example, RADAR (rapid automatic detection and alignment of repeats) [21] uses an automatic algorithm, for segmenting a query sequence into repeats; it identifies short composition biased as well as gapped approximate repeats and complex repeat architectures involving many different types of repeats in a query sequence. Rep program [22] uses an iterative algorithm based on score distributions from profile analysis. This procedure allows the identification of homologues at alignment scores lower than the highest optimal alignment score for nonhomologous sequences. The PROSPERO program [23] is ideal for large scale self-comparison of protein sequences. It uses a formula that accurately assesses the significance of protein repeat similarities, allowing for existence of gaps, and also takes into account sequence length and composition. TRUST (tracking repeats using significance and transitivity) program [24] exploits the concept of transitivity of alignments as well as a statistical scheme optimized for the evaluation of repeat significance. Starting from significant local suboptimal alignments, the application of transitivity allows to (1) identify distant repeat homologues for which no alignments were found; (2) gain confidence about consistently well-aligned regions; and (3) recognize and reduce the contribution of nonhomologous repeats. This assessment step will enable to derive a virtually noise-free profile representing a generalized repeat with high fidelity. It has been demonstrated by the authors that TRUST is a useful and reliable tool for mining tandem and nontandem repeats in protein sequence databases, to predict multiple repeat types with varying intervening segments within a single sequence. Once statistically significant repeats are detected, construction of a multiple sequence alignment provides insight into the extent of sequence homology among members of the new protein family and identification of the conserved sequence motifs.

We have implemented TRUST on a personal computer in our laboratory and used it to identify amino-acid sequence repeats in the proteins of *B. anthracis* str. *Ames* proteome. We have identified four repeats and ten domains that are novel in the proteome of *B. anthracis* str. *Ames*. Further analysis corresponding to searches of the completed and unfinished genome databases identified some of these to be present in other bacterial genomes.

2. METHODS

We have downloaded the entire proteome of *B. anthracis* str. *Ames* from the website http://www.ncbi.nlm.nih.gov in the FASTA format. The TRUST program was downloaded from the website and installed on the local Pentium IV computers on the Linux platform. The TRUST server together with the source code is available at http://ibivu.cs.vu.nl/programs/trustwww. The TRUST program was run for all the sequences in this proteome. Based on the size of the TRUST output file, the protein sequences with no internal repeats were discarded automatically; that is, only those protein sequences which comprise repeats were retained. The lengths of repeats and domains currently annotated in the INTERPRO database often comprise greater than 25-amino-acid residues; therefore, in this work, we have considered the repeats with greater than 25-amino-acid residues alone for further analysis. Thus selected proteins were submitted to SMART online (http://smart.embl-heidelberg.de/smart/batch.pl) [13] program in batch mode. Manual inspections of the SMART results identified proteins comprising known repeats or domains and were therefore discarded. Only those repeats that are not identified by SMART database are retained for further analysis.

We have downloaded NCBI NR (release date: April 22, 2005) and UNIPROT (release date: April 23, 2005) databases and installed BLAST-2.2.10 on the local Linux computers (OS: Fedora Core-2, Pentium-IV 3.00 GHz, 1 GB RAM, 80 GB hard disk). Using automatic shell scripts, these protein sequences were then blasted using PSI-BLAST program [25] for three iterations against the NCBI NR database and using BLASTALL program against UNIPROT database. The proteins confirmed to comprise repeats by the BLAST program were retained and were tested for presence in the offline versions of INTERPRO (Database: iprscan_DATA_10.0, Applications: iprscan_V4.1, iprscan_binn4.x_Linux) and PFAM (release date: April 26, 2005) databases. A final check was made using online versions of INTERPRO and PFAM. These series of steps are given in the flowchart as shown in Figure 1.

The repeats which are not present in any of these databases were considered to be novel repeats or domains, depending upon (1) the number of times they occur in the protein sequences, and (2) length of the amino-acid sequence region. The novel repeats and domains thus identified in *B. anthracis* str. *Ames* proteome were subjected to PSI-BLAST analysis in order to identify other proteins from databases that comprise these repeats and domains. Multiple sequence alignment program, ClustalW [26], was used to detect the extent of sequence conservation and the secondary structure prediction was carried out using PHD [27] method.

3. RESULTS AND DISCUSSION

From the analysis of *B. anthracis* str. *Ames* proteome using TRUST program, we identified 905 proteins comprising

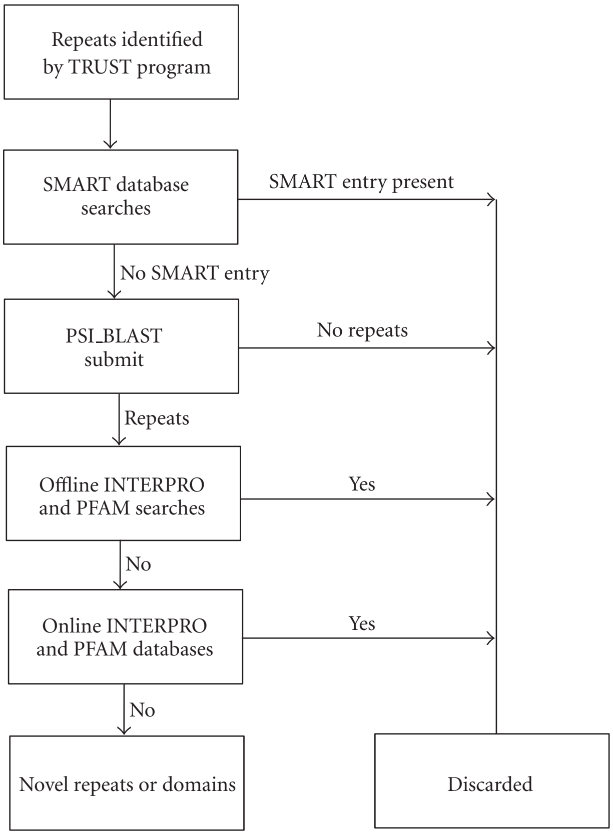

Figure 1: Flowchart for systematic analysis of repeats in proteins.

of amino-acid sequence repeats. SMART database analysis identified that 302 entries do not have a SMART description. Based on their absence in the INTERPRO and PFAM databases and the length of repeat sequence (greater than 25-amino-acid residues), we have identified about 120 proteins (data not shown) in the *B. anthracis* str. *Ames* proteome to comprise novel amino-acid sequence repeats. We have added an additional constraint that the repeats identified by TRUST program should also be identified as a repeat by the BLAST program. Subsequent online INTERPRO and PFAM searches confirmed that these domains and repeats have not been reported before. In this work, we have identified four repeats and ten domains, that are not within or part of previously reported repeats and our findings are therefore novel. Further analysis identified some of these in the proteins of other bacterial genomes. The conserved amino-acid residues observed from multiple sequence alignments using the CLUSTALW program were used to describe sequence motifs characteristic of these novel repeats and domains. Often, more than one sequence motif is associated with repeats or domains and the amino-acid sequence patterns characteristic of these repeats are represented according to the PROSITE description [28]. Ponting et al. [29], have earlier used a similar approach to identify novel domains and repeats in *Drosophila melanogaster*.

In this work, we identified four repeats and ten domains that have not been reported before in the *B. anthracis* str. *Ames* proteome. The repeats and domains described in 1 to 6 and 9 are also present in some bacterial organisms, 7, 8, 10 and 11 are *Bacillus*-specific, 12 and 13 are *Bacillus anthracis* str. *Ames* specific. Lists of the proteins containing these novel repeats and domains are shown in Tables 1a to 1k. These tables indicate the protein identifiers (Gene or Swall_ID), the number of amino-acid residues in the protein, a description of the protein, and other well-characterized repeats and domains present in the protein. Some sequences representing these repeats or domains share lower than 15% pairwise sequence identity. However, these sequences retain the conserved motifs and the positions of secondary structure elements in the multiple sequence alignment. For all the proteins, the amino-acid sequence corresponding to each representative repeat are shown in the multiple sequence alignments (see Figures from 2 to 14).[1] Conservation of the position of secondary structural elements is indicated from the multiple sequence alignment. The schematic figures used to represent these repeats and domains are shown in Figures 15 to 27. These figures (drawn to an approximate scale) reflect the relative proximity and location of individual repeats and domains along the primary sequence. We discuss each of these novel repeats and domains below.

3.1. 57-amino-acid-residue PxV domain

The 251-amino-acid-residue protein corresponding to the GENE_ID BA2292 and described as hypothetical protein comprises of a 57-amino-acid-residue region as two copies. Further BLAST searches using sequence corresponding to the region (65–121) as a query identified 24 proteins that are described as hypothetical (see Table 1(a)). This region occurs as four copies in proteins from *Shewanella amazonensis, and Haloarcula marismortui*, as two copies in proteins from *B. anthracis, B. cereus, B. halodurans, B. thuringiensis, B. thuringiensis serovar*, *Thermus thermopilus, Chloroflexus aurantiacus*, *Chloroflexus aggregans Exiguobacterium sp., Bacillus weihenstephanensis, Roseiflexus castenholzii, Clostridium novyi, Herpetosiphon aurantiacus*, and as single copy in *Anabaena variabilis*; we therefore describe this region as a

[1] The multiple sequence alignments corresponding to representative repeats and domains from various proteins along with their GENE or SWall identifiers. (a) PxV domain, (b) FxF domain, (c) YEFF domain, (d) IMxxH domain, (e) VxxT domain, (f) ExW domain, (g) NTGFIG domain, (h) NxGK repeat (i) VYV domain, (j) KEWE domain, (k) AFL domain, (l) RIDVK repeat, (m)(a) AGQF repeat and (b) GSAL repeat. The numbers given in brackets indicate the start and end of amino-acid-residue positions corresponding to either the repeat or domain. The 80% consensus is labeled according to the alignment to the alignment generated at the website http://www.bork.embl-heidelberg.de/Alignment/consensus.html: alcohol (o, ST); aliphatic (I, ILV); any (·, ACDEFGHIKLMNPQRSTVWY); aromatic (a, FHWY); charged (c, DEHKR); hydrophobic (h, ACFGHIKLM-RTVWY); negative (−, DE); polar (p, CDEHKNQRST); positive (+, HKR); small (s, ACDGNPSTV); tiny (u, AGS); turn-like (t, ACDEGHKN-QRST). A capital letter indicates 80% conservation of corresponding amino-acid residue. The secondary structure prediction indicated at the top was derived using the PHD program. Residues predicted with greater than 82% accuracy to form β-sheets are represented by "E" and α-helices are represented by "H."

Table 1: The proteins are represented by their corresponding Gene_ID along with the number of amino-acid residues indicated in brackets in the first column. The organism and corresponding phylogeny are indicated in the second column: (A) represents Archaea and (B) represents Bacteria. The third column contains the description of the proteins containing the repeats or the domains identified elsewhere, including those identified in the present work and the total number of such repeats or domains. The fourth column represents exclusively the total number of novel repeats or domains identified in this work.

(a) List of proteins containing the 57-amino-acid-residue PxV domain.

Gene ID (number of residues)	Organism	Description	Number of PxV domains
BA2292 (251)	*Bacillus anthracis* str. *Ames* (B)	Hypothetical protein	2
BAS2138 (249)	*Bacillus anthracis* Sterne (B)	Hypothetical protein	2
BT9727_2076 (249)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Hypothetical protein	2
BCZK2072 (249)	*Bacillus cereus* E33L (B)	Hypothetical protein	2
BCE2326 (249)	*Bacillus cereus* ATCC 10987 (B)	Hypothetical protein	2
BC2244 (249)	*Bacillus cereus* ATCC 14579 (B)	Hypothetical protein	2
BH1282 (222)	*Bacillus halodurans* C-125 (B)	BH1282 protein	2
BCE_G9241_2259 (249)	*Bacillus cereus* G9241 (B)	Hypothetical conserved protein	2
RBTH_03198 (251)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical protein	2
TT_P0044 (221)	*Thermus thermopilus* HB27 (B)	Hypothetical conserved protein	2
TTHB089 (221)	*Thermus thermophilus* HB8 (B)	Hypothetical protein	2
Chlo02001630 (262)	*Chloroflexus aurantiacus* J-10-fl (B)	Hypothetical protein	2
ExigDRAFT_0608 (264)	*Exiguobacterium* sibiricum 255-15 (B)	Hypothetical protein	2
SamaDRAFT_3539 (469)	*Shewanella amazonensis* SB2B (B)	Hypothetical protein	4
rrnAC0576 (488)	*Haloarcula marismortui* ATCC 43049 (A)	Unknown	4
Ava_3757 (292)	*Anabaena variabilis* ATCC 29413 (B)	Hypothetical protein	1
BcerKBAB4DRAFT_2942 (249)	*Bacillus weihenstephanensis* KBAB4 (B)	Conserved hypothetical protein	2
B14911_22687 (254)	*Bacillus* sp. NRRL B-14911 (B)	Hypothetical protein	2
Bcer98DRAFT_2673 (249)	*Bacillus cereus* subsp. cytotoxis NVH (B)	Conserved hypothetical protein	2
RcasDRAFT_0590 (259)	*Roseiflexus castenholzii* DSM 13941 (B)	Surface protein from Gram-positive cocci, anchor region	2
RoseRSDRAFT_1732 (259)	*Roseiflexus* sp. RS-1 (B)	Surface protein from Gram-positive cocci, anchor region	2
NT01CX_1619 (210)	*Clostridium novyi* NT (B)	Conserved hypothetical protein	2
HaurDRAFT_2803 (196)	*Herpetosiphon aurantiacus* ATCC 23779 (B)	Conserved hypothetical protein	2
CaggDRAFT_2922 (261)	*Chloroflexus aggregans* DSM 9485 (B)	Conserved hypothetical protein	2

Table 1: Continued.

(b) List of proteins containing the 122-amino-acid-residue FxF domain.

Gene ID (number of residues)	Organism	Description	Number of FxF domains
BA0881 (293)	*Bacillus anthracis* str. *Ames* (B)	Conserved domain protein	2
BCZK0785 (293)	*Bacillus cereus* E33L (B)	Hypothetical protein	2
BT9727_0783 (295)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Hypothetical protein	2
BCE_G9241_0886 (293)	*Bacillus cereus* G9241 (B)	Conserved protein, putative	2
GK3171 (297)	*Geobacillus kaustophilus* HTA426 (B)	Hypothetical conserved protein	2
CTC00525 (279)	*Clostridium tetani* E88 (B)	Hypothetical protein	2
Bcer98DRAFT_3031 (293)	*Bacillus cereus* subsp. cytotoxis NVH (B)	Conserved hypothetical protein	2
B14911_04439 (305)	*Bacillus* sp. NRRL B-14911 (B)	Hypothetical protein	2
DredDRAFT_0533 (262)	*Desulfotomaculum reducens* MI-1 (B)	Hypothetical protein	2
NT01CX_1557 (276)	*Clostridium novyi* NT (B)	Conserved protein, putative	2

(c) List of proteins containing the 111-amino-acid-residue YEFF domain.

Gene ID (number of residues)	Organism	Description and other known domains	Number of YEFF domains
BA3695 (510)	*Bacillus anthracis* str. *Ames* (B)	S-layer protein, putative, SLH-domain (3)	2
BCZK3337 (492)	*Bacillus cereus* E33L (B)	S-layer protein, SLH-domain (3)	2
BT9727_3386 (510)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	S-layer protein, SLH-domain (3)	2
Bant_01004347 (510)	*Bacillus anthracis* str. A2012 (B)	Hypothetical protein, SLH-domain (3)	2
BCE_G9241_3590 (492)	*Bacillus cereus* G9241 (B)	Lipoprotein, putative SLH-domain (3)	2
BA5326 (321)	*Bacillus anthracis* str. *Ames* (B)	Lipoprotein, putative	2
BT9727_4791 (321)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Hypothetical protein	2
BC5098 (321)	*Bacillus cereus* ATCC 14579 (B)	Hypothetical protein	2
BCZK4809 (321)	*Bacillus cereus* E33L (B)	Hypothetical protein	2
RBTH_06214 (321)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical protein	2
EF0374 (325)	*Enterococcus faecalis* V583 (B)	Lipoprotein, putative	2
EF0375 (321)	*Enterococcus faecalis* V583 (B)	Hypothetical protein	2
EF0376 (347)	*Enterococcus faecalis* V583 (B)	Hypothetical protein	2

(d) List of proteins containing the 109-amino-acid-residue IMxxH domain.

Gene ID (number of residues)	Organism	Description	Number of IMxxH domains
BA1021 (266)	*Bacillus anthracis* str. *Ames* (B)	Hypothetical protein	2
BAS0955 (283)	*Bacillus anthracis* Sterne (B)	Hypothetical protein	2
BCZK0933 (283)	*Bacillus cereus* E33L (B)	Hypothetical protein	2
BT9727_0941 (283)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Hypothetical protein	2
BC1029 (283)	*Bacillus cereus* ATCC 14579 (B)	Hypothetical protein	2
RBTH_03050 (283)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical protein	2
CAC3450 (307)	*Clostridium acetobutylicum* ATCC 824 (B)	Hypothetical protein	2
CPE0158 (303)	*Clostridium perfringens* str. 13 (B)	Hypothetical protein	2
CTC02189 (314)	*Clostridium tetani* E88 (B)	Conserved protein	2
CtheDRAFT_1311 (307)	*Clostridium thermocellum* ATCC 27405 (B)	Conserved hypothetical protein	2
DhafDRAFT_0725 (321)	*Desulfitobacterium hafniense* DCB-2 (B)	Conserved hypothetical protein	2

Table 1: Continued.

(d) Continued.

Gene ID (number of residues)	Organism	Description	Number of IMxxH domains
BCE_G9241_1042 (283)	*Bacillus cereus* G9241 (B)	Conserved protein	2
CbeiDRAFT_3331 (312)	*Clostridium beijerincki* NCIMB 8052 (B)	Conserved hypothetical protein	2
CphyDRAFT_3436 (305)	*Clostridium phytofermentans* ISDg (B)	Conserved hypothetical protein	2
ClosDRAFT_1658 (308)	*Clostridium* sp. OhILAs (B)	Conserved hypothetical protein	2
CdifQ_02001573 (254)	*Clostridium difficile* QCD-32g58 (B)	Hypothetical protein	2
BcerKBAB4DRAFT_3543 (283)	*Bacillus weihenstephanensis* KBAB4 (B)	Hypothetical protein	2
AmetDRAFT_1908 (272)	*Alkaliphilus metalliredigenes* QYMF (B)	Conserved hypothetical protein	2
CD1511 (304)	*Clostridium difficile* 630 (B)	Conserved hypothetical protein	2
CPF_0149 (303)	*Clostridium perfringens* ATCC 13124 (B)	Hypothetical protein	2
BcerKBAB4DRAFT_0307 (171)	*Bacillus weihenstephanensis* KBAB4 (B)	Conserved hypothetical protein	1
Bcer98DRAFT_1038 (303)	*Bacillus cereus* subsp. cytotoxis NVH 391-98 (B)	Conserved hypothetical protein	2

(e) List of proteins containing the 103-amino-acid-residue VxxT domain.

Gene ID (number of residues)	Organism	Description	Number of VxxT domains
BA4716 (349)	*Bacillus anthracis* str. *Ames* (B)	Germination protein gerM	2
gerM BT9727_4219 (349)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Germination protein	2
germ BCZK4235 (349)	*Bacillus cereus* E33L (B)	Germination protein	2
BCE4587 (349)	*Bacillus cereus* ATCC 10987 (B)	Germination protein gerM	2
BC4495 (349)	*Bacillus cereus* ATCC 14579 (B)	Germination protein germ	2
BSU28380 (366)	*Bacillus subtilis* subsp. subtilis str. 168 (B)	Germination protein gerM	2
BL00314 (369)	*Bacillus licheniformis* ATCC 14580 (B)	Spore germination protein GerM	2
BH3070 (365)	*Bacillus halodurans* C-125 (B)	Germination (Cortex hydrolysis) and sporulation	2
RBTH_05210 (349)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Germination protein germ	2
gerM (210)	*Bacillus subtilis* (B)	gerM	1
ABC2653 (377)	*Bacillus clausii* KSM-K16 (B)	Germination protein GerM	2
GK2667 (357)	*Geobacillus kaustophilus* HTA426 (B)	Germination (Cortex hydrolysis) and sporulation	2
OB2107 (352)	*Oceanobacillus iheyensis* HTE831 (B)	Germination (Cortex hydrolysis) and sporulation	2
SwolDRAFT_2302 (195)	*Syntrophomonas wolfei* str. Goettingen (B)	Hypothetical protein	1
MothDRAFT_0979 (200)	*Moorella thermoacetica* ATCC 39073 (B)	Similar to Spore germination protein	1
CtheDRAFT_0840 (299)	*Clostridium thermocellum* ATCC 27405 (B)	Hypothetical protein	1
gerM ABF83609 (349)	*Bacillus thuringiensis* serovar kurstaki (B)	Spore germination protein	2
Bcer98DRAFT_3179 (348)	*Bacillus cereus* subsp. cytotoxis NVH 391-98 (B)	Germination protein GerM	2
BcerKBAB4DRAFT_4089 (349)	*Bacillus weihenstephanensis* KBAB4 (B)	Germination protein gerM	2
B14911_06091 (361)	*Bacillus* sp. NRRL B-14911 (B)	Spore germination protein	2
GAA01614 (295)	*Pelotomaculum thermopropionicum* SI (B)	Unnamed protein product	1
AmetDRAFT_1640 (332)	*Alkaliphilus metalliredigenes* QYMF (B)	Hypothetical protein	2
Moth_0516 (200)	*Moorella thermoacetica* ATCC 39073 (B)	Spore germination protein-like	1

Table 1: Continued.

(f) List of proteins containing the 84-amino-acid-residue ExW domain.

Gene ID (number of residues)	Organism	Description	Number of ExW domains
BA4310 (246)	*Bacillus anthracis* str. *Ames* (B)	Hypothetical protein	2
BT9727_3829 (246)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Hypothetical protein	2
BCE4157 (246)	*Bacillus cereus* ATCC 10987 (B)	Hypothetical protein	2
BCZK3845 (246)	*Bacillus cereus* E33L (B)	Hypothetical protein	2
BC4088 (248)	*Bacillus cereus* ATCC 14579 (B)	IG hypothetical 17224	2
GK0969 (226)	*Geobacillus kaustophilus* HTA426 (B)	Hypothetical conserved protein	2
BSU30660 (145)	*Bacillus* subtilis subsp. str. 168 (B)	Hypothetical protein ytkA (PSPA8)	1
BL05305 (147)	*Bacillus licheniformis* ATCC 14580 (B)	Conserved protein YtkA	1
BH0983 (157)	*Bacillus halodurans* C-125 (B)	BH0983 protein	1
Bant_01004966 (252)	*Bacillus anthracis* str. A2012 (B)	Protein chain release factor A	2
RBTH_02670 (248)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical protein	2
BCE_G9241_4093 (246)	*Bacillus cereus* G9241 (B)	IG hypothetical protein	2
OB2488 (166)	*Oceanobacillus ihenyensis* HTE831 (B)	Hypothetical conserved protein	1
ABC0230 (158)	*Bacillus clausii* KSM-K16 (B)	Unknown conserved protein	1
BH0678 (246)	*Bacillus halodurans* C-125 (B)	BH0678 protein	2
ABC4088 (142)	*Bacillus clausii* KSM-K16 (B)	Hypothetical protein	1
ExigDRAFT_1796 (161)	*Exiguobacterium sibiricum* 255-15 (B)	Hypothetical protein	1
OB3282 (155)	*Oceanobacillus ihenyensis* HTE831 (B)	Hypothetical conserved protein	1
BcerKBAB4DRAFT_2040 (241)	*Bacillus weihenstephanensis* KBAB4 (B)	Conserved hypothetical protein	2
B14911_09907 (144)	*Bacillus* sp. NRRL B-14911 (B)	Hypothetical protein	1
B14911_05359 (273)	*Bacillus* sp. NRRL B-14911 (B)	Hypothetical protein	2
BAA83944 (267)	*Bacillus halodurans* (B)	Unnamed protein product	2
BH1853 (158)	*Bacillus halodurans* C-125 (B)	Hypothetical protein	1
Bcer98DRAFT_3614 (177)	*Bacillus cereus* subsp. cytotoxis NVH 391-98 (B)	IG hypothetical protein	2
ExigDRAFT_0574 (253)	*Exiguobacterium sibiricum* 255-15 (B)	Hypothetical protein	2

(g) List of proteins containing the 104-amino-acid-residue NTGFIG domain.

Gene ID (number of residues)	Organism	Description	Number of NTGFIG domains
BA2665 (232)	*Bacillus anthracis* str. *Ames* (B)	Hypothetical protein	2 tandem
BT9727_2444 (232)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Hypothetical protein	2 tandem
BCZK2413 (232)	*Bacillus cereus* E33L (B)	Group-specific protein	2 tandem
BCE2700 (234)	*Bacillus cereus* ATCC 10987 (B)	Hypothetical protein	2 tandem
BC2674 (234)	*Bacillus cereus* ATCC 14579 (B)	Hypothetical protein	2 tandem
Bant_01003317 (236)	*Bacillus anthracis* str. A2012 (B)	Hypothetical protein	2 tandem
BCE_G9241_CNI_0263 (234)	*Bacillus cereus* G9241 (B)	Conserved hypothetical protein	2 tandem
BcerKBAB4DRAFT_0535 (232)	*Bacillus weihenstephanensis* KBAB4(B)	Conserved hypothetical protein	2 tandem
Bcer98DRAFT_0128 (234)	*Bacillus cereus* subsp. cytotoxis NVH 391-98 (B)	Conserved hypothetical protein	2 tandem

Table 1: Continued.

(h) List of proteins containing the 36-amino-acid-residue NxGK repeat.

Gene ID (number of residues)	Organism	Description and other known domains	Numbre of NxGK repeats
BA3686 (193)	*Bacillus anthracis* str. *Ames* (B)	Hypothetical protein, SAP domain (1)	2
BT9727_3378 (193)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Hypothetical protein, SAP domain (1)	2
BCZK3328 (193)	*Bacillus cereus* E33L (B)	Hypothetical protein, SAP domain (1)	2
BC3626 (193)	*Bacillus cereus* ATCC 14579 (B)	Hypothetical protein, SAP domain (1)	2
BCE3645 (193)	*Bacillus cereus* ATCC 10987 (B)	Hypothetical protein, SAP domain (1)	2
RBTH_03615 (193)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical cytosolic protein, SAP domain (1)	2
BCE_G9241_3579 (193)	*Bacillus cereus* G9241 (B)	Hypothetical cytosolic protein, SAP domain (1)	2
BcerKBAB4DRAFT_0944 (193)	*Bacillus weihenstephanensis* KBAB4 (B)	Conserved hypothetical protein, SAP domain (1)	2
B14911_25780 (189)	*Bacillus* sp. NRRL B-14911 (B)	Hypothetical protein, SAP domain (1)	2

(i) List of proteins containing the 95-amino-acid-residue VYV domain.

Gene ID (number of residues)	Organism	Description	Number of VYV domains
BA1701 (225)	*Bacillus anthracis* str. *Ames* (B)	Hypothetical protein	2 tandem
BAS1577 (227)	*Bacillus anthracis* str. Sterne (B)	Hypothetical protein	2 tandem
RBTH_03882 (1004)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical exported protein	10 tandem
DSY3134 (1674)	*Desulfitobacterium hafniense Y51* (B)	Hypothetical protein	2 tandem

(j) List of proteins containing the 75-amino-acid-residue KEWE domain.

Gene ID (number of residues)	Organism	Description	Number of KEWE domains
BA3147 (262)	*Bacillus anthracis* str. *Ames* (B)	Hypothetical protein	3 tandem
BAS2924 (344)	*Bacillus anthracis* str. Sterne (B)	Hypothetical protein	4 tandem
RBTH_06405 (331)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical protein	4 tandem
pE33L466_0092 (328)	*Bacillus cereus* E33L (B)	Hypothetical protein	4 tandem
Bant_01003795 (178)	*Bacillus anthracis* str. A2012 (B)	Hypothetical protein	2 tandem
pBMB165 (247)	*Bacillus thuringiensis* serovar tenebrionis (B)	Hypothetical protein	3 tandem

(k) List of proteins containing the 59-amino-acid-residue AFL domain.

Gene ID (number of residues)	Organism	Description	Number of AFL domains
BA3065 (290)	*Bacillus anthracis* str. *Ames* (B)	Hypothetical protein.	2
BAS2851 (297)	*Bacillus anthracis* str. Sterne (B)	Hypothetical protein	2
Bant_01003715 (293)	*Bacillus anthracis* str. A2012 (B)	Hypothetical protein	2
RBTH_02124 (145)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical protein	1
BcerKBAB4DRAFT_1832 (291)	*Bacillus weihenstephanensis* KBAB4 (B)	Conserved hypothetical protein	2

domain. The length of proteins varied between 196 to 488-amino-acid residues. The multiple sequence alignment corresponding to this domain is associated with PxV sequence motif where x is any amino-acid residue and is shown in Figure 2. The pairwise identities between sequences corresponding to PxV domain varied between 15–96%. The secondary structure corresponding to PxV domain is predicted to comprise four β-strands as shown in Figure 2. The representative domain architecture corresponding to proteins comprising the PxV domain is shown in Figure 15.

```
Secondary structure               EEEEE          EE                          EEEEEE         EE
RcasDRAFT_0590_1(32-89)           VRVIHAS-PDAPAVDVIVNGNR--ALTNVPFFAASAYLDLPAGSYDIQVVPAGAT-S-PVVID  58
RoseRSDRAFT_1732_1(32-89)         VRVVHAS-PDAPAVDVIVNGNK--ALTNVPFFAASAYLDLPAGSYDIQVVPAGAT-S-PVVID  58
Chlo02001630_1(32-90)             VRVIHAS-PDAPAVDVFVNGNA--VLTNVGFFAASPYLDLPAGTYRVQVAPTGAG-AGSAVID  59
CaggDRAFT_2922_1(31-89)           VRVIHAS-PDAPAVDVFVNGNA--VLTNVGFFAASPYLDLPAGTYRVQVAPTGAG-AGSAVID  59
HaurDRAFT_2803_1(4-62)            VRVMHAS-PDAPAVDIFVDGKA--VLTSVPFFALSGQLALPDGTYTIDIAPAGAG-VAASVFE  59
B14911_22687_1(67-124)            VRVVHAS-PDAPNVDIYVNGNR--ILKDFPYKDVSGYLSLPAGKYQIDIYPAGDM-V-STVLS  58
HaurDRAFT_2803_2(105-162)         VRVIHGS-PDAPAVDIKIAGTQN-VVVKGAKFGDAATLEVPAGTYSFDISPAGSS-T--VLFT  58
rrnAC0576_1(67-124)               VRVAHMS-PNAPNVDVYLEGDA--VLEDVPFGAVSQYLDVPAGERSVEITAAGD--PDTSVFS  58
rrnAC0576_2(284-341)              VRVAHMS-PNAPNVDVYVDGSA--VLEDVPFGAVSDYLEVPAGARTVEITAAGD--PDTSVFE  58
BH1282_1(30-89)                   VRVLHAS-PDAPPVDVYIDGKK--QMEGVPFKQTSSYFNVPAGDHMITIFAAGDDPAETPVIE  60
ExigDRAFT_0608_1(29-86)           VRVIHAS-PDAPAVDIAVDGKK--AVSGAEFKAVTDYLTLPAGEHKVEVFAAGT--TKDPVLS  58
RBTH_03198_2(161-218)             IRFAHFS-PDTPVVNVDLKDGDH-LFENVLFKQITDFLQVSPGTADIEISLANNK---NVLLT  58
BC2244_2(159-216)                 IRFAHFS-PDTPVVNVDLKDGDH-LFENVLFKQITDFLQVSPGTADIEISLADNK---NVLLT  58
BcerKBAB4DRAFT_2942_2(159-216)    IRFAHFS-PDTPVVNVNLKDGDH-LFENVLFKQITDFLQVSPGTADIEVSLADTK---KVLLT  58
BCE_G9241_2259_2(159-217)         IRFAHFS-PDTPVVNVSLKGGDH-LFENVLFKQITDFLEVSPGTADIEVSLADNQ---NVLLT  58
BCE_2326_2(159-216)               IRFAHFS-PDTPVVNVSLKGGDH-LFENVLFKQITDFLEVSPGTADIEVSLADHQ---SVLLT  58
BA2292_2(161-218)                 IRFAHFS-PDTPVVNVSLKDGDH-LFENVLFKQITDFLEVSPGTADIEVSLADNQ---SVLLT  58
BAS2138_2(159-216)                IRFAHFS-PDTPVVNVSLKDGDH-LFENVLFKQITDFLEVSPGTADIEVSLADNQ---SVLLT  58
BT9727_2076_2(159-216)            IRFAHFS-PDTPVVNVSLKDGDH-LFENVLFKQITDFLEVSPGTADIEVSLADNQ---SVLLT  58
BCZK2072_2(159-216)               IRFAHFS-PDTPVINVSLKDGDH-LFENVLFKQITDFLEVSPGTADIEVSLADNQ---SILLT  58
Bcer98DRAFT_2673_2(159-216)       IRFAHFS-PDTSVVNVSLKNGDH-LFENVLFKQVTDYLQVSPGTADIEISLADTK---KNLVT  58
NT01CX_1619_2(113-170)            VKFVHLS-PGTPNVDITLPNGTI-LFKDVEFEEGTDYIPLKVGTYTIEAKPTGSD---KTVLT  58
B14911_22687_2(164-221)           ARFIHLS-PDAPAVDIAVKKGDV-IFPNISFRQATQYLGLTPMTVDLEVRVAGSS---NTVLS  58
RcasDRAFT_0590_2(131-188)         VRVIHFS-PDAPAVDIKVAGGPT-LISNLAFPNASNYLPVDAGSYDLQVTPAGGT---AVVLD  58
RoseRSDRAFT_1732_2(131-188)       VRVIHFS-PDAPAVDIKVAGGPT-LISNLAFPQASNYLPVDAGSYDLQVTPAGGT---AVVLD  58
Chlo02001630_2(131-188)           VRVYHFS-PDAPAVDVKLANGTT-LISNLAFPNASDYLEVPAGTYDLQVTPAGGS---AVVIN  58
CaggDRAFT_2922_2(130-187)         VRVYHFS-PDAPAVDVKLANGTT-LISNLAFPDASDYLEVPAGTYDLQVTPAGGD---AVVIN  58
ExigDRAFT_0608_2(126-183)         VRVAHFA-PDAPAVDVAPKGGDP-LFSDLEFSKVSDYGTLDAGTYDLEVRPAGAT---DVVKA  58
TTHB089_2(124-181)                IRVVHAS-PDAPAVDVAVKGGPV-LFAGLPFPRASAYASVPAGTYDLEVRAAGTA---TVALD  58
TTP0044_2(124-181)                IRVVHAS-PDAPAVDVAVKGGPV-LLAGLPFPRASAYASVPAGTYDLEVRAAGTA---TVALD  58
BH1282_2(130-187)                 LRAVHLS-PDTPAVQLHLSAANV-DMPSLSFENASRYIDLPAGAYDLDIRMIETD---DVATE  58
RBTH_03198_1(65-121)              IRIFHAD-PNIPAVDILVNGQKV--IKNISFKQFSPYLSLVQGKYRIDIVPVGNET---PIFS  57
BC2244_1(63-119)                  IRIFHAD-PNIPAVDILVNGQKV--IKNISFKQFSPYLSLVQGKYRIDIVPVGNET---PIFS  57
BCE_G9241_2259_1(63-119)          IRFFHSA-SNTPAVDILVNGQKV--IKNISFKQFSPYLTLVQGKYRIDIVPVGNET---PIFS  57
BAS2138_1(63-119)                 IRFFHSA-SNTPAVDILVNGQKV--IKNISFKQFSPYLTLVQGKYRIDIVPVGNET---PIFS  57
BCE_2326_1(63-119)                IRFFHSA-SNTPAVDILVNGQKV--IKNISFKQFSPYLTLVQGKYRIDIVPVGNET---PIFS  57
BA2292_1(65-121)                  IRFFHSA-SNTPAVDILVNGQKV--IKNISFKQFSPYLTLVQGKYRIDIVPVGNET---PIFS  57
BCZK2072_1(63-119)                IRFFHSA-SNTPAVDILVNGQKV--IKNISFKQFSPYLTLVQGKYRIDIVPVGNET---PIFS  57
BT9727_2076_1(63-119)             IRFFHSA-SNTPAVDILVNGQKV--IKNISFKQFSPYLTLVQGKYRIDIVPVGNET---PIFS  57
BcerKBAB4DRAFT_2942_1(63-119)     MRIFHTA-PHTPAVDIIINGQKV--IKNISFKQFSPYLSLVQGKYRIDIVPVGNET---PIFS  57
Bcer98DRAFT_2673_1(63-119)        MRIFHAS-PHTAPVDILINGQKV--IKNITFQQFSPYFSLMQGQYRLDIVPLDNET---PIFS  57
SamaDRAFT_3539(264-321)           IRVAHSA-ADVPQVDILANGTKVDALSGAAFGQASGYLNLAPGEYQVDTVLTSDNS---VVGI  59
Ava_3757(63-120)                  LRVINAAVPTASPVDVIVNGQRV--LENVNFRQASRYVNVTPGNIQVLFVTSGTNS---TIAS  58
TTHB089_1(24-81)                  VRVAHLS-PDAPAVDVLVNGQRA--ITGLAFKEVTPYIPLPAAKVRVQVVPAGQDAP--VVID  58
TTP0044_1(24-81)                  VRVAHLS-PDAPAVDVLVNGQRA--ITGLAFKEVTPYIPLPAAKVRVQVVPAGQDAP--VVID  58
NT01CX_1619_1(15-72)              MRLLNAS-PNAPAVDVYFNGQLI--TSNLAYKEFTEYMSTSPGLYNVKVFPHGKLSS--PIID  58

consensus/80%                     lRhhHhu.PssPsVsl.lpstt...hpsl.F.phosalpls.Gphplpl..ssst....slhs
```

Figure 2: BA2292 is homologous to protein GBAA2292 from *Bacillus anthracis* str. "*Ames* Ancestor." BAS2138 is homologous to proteins BT9727_2076 from *Bacillus thuringiensis* serovar konkukian str. 97-27 and Bant_01002917 from *Bacillus anthracis* str. A2012.

3.2. 122-amino-acid-residue FxF domain

The 293-amino-acid-residue protein corresponding to the GENE_ID BA0881 and described as conserved domain protein comprises a 122-amino-acid-residue region as two copies. Further BLAST searches using sequence corresponding to the region (55–176) as a query identified 10 proteins (see Table 1(b)). The proteins comprising this region are described as either conserved or hypothetical proteins. This region occurs as two copies in the proteins of *B. anthracis, B. cereus, B. thuringiensis*, *Geobacillus kaustophilus, Clostridium tetani*, *Clostridium novyi*, and *Desulfotomaculum reducens* genomes. The length of proteins varied between 262 to 305-amino-acid residues. The multiple sequence alignment corresponding to this domain is associated with characteristic sequence motif FxF (Figure 3) and we refer to this as the FxF domain. The pairwise sequence identities corresponding to this domain varies between 18–97%. The secondary structure corresponding to FxF domain is predicted to comprise one α-helix and five β-strands, and the representative domain architecture of proteins comprising this domain is shown in Figure 16.

3.3. 111-amino-acid-residue YEFF domain

The 510-amino-acid-residue protein corresponding to the GENE_ID BA3695 and described as a S-layer protein comprises a 111-amino-acid-residue region that is present as two copies. Further BLAST searches, using sequence corresponding to the region (247–357) as a query, identified 13 proteins (see Table 1(c)), that are described as S-layer proteins, hypothetical, or lipoproteins and correspond to the *B. anthracis* str. *Ames* and A2012, *B. cereus*, *B. thuringiensis, B. thuringiensis serovar israelensis*, and *Enterococcus faecalis* genomes. The length of proteins varied between 321 to 510-amino-acid residues. Five proteins corresponding to the GENE_ID BA3695 and Bant_01004347 of *B. anthracis*, BCE_G9241_3590, and BCZK3337 of *B. cereus* and BT9727_3386 of *B. thuringiensis* comprise three copies of SLH domain, indicating a cell surface role for these proteins.

```
Secondary structure          HHHHHHH         EEEEE          EEEEEE            EEEEE      EEEEE

BA0881_1(55-176)             IYQFLHKELPRLEEYQISLSGIEIEKRDNG-YDVAVFIRSTVPKPISFEEVTLILLNKEKKLCARKT
BCZK0785_1(55-176)           IYQFLHKELPRLEEYQISLSGIEIEERDNG-YDVAVFIRSTVPKPISFEEVTLILLNKEKKLCARKT
BCE_G9241_0886_1(55-176)     IYQFLHKELPRLEEYQISLSGIEIEKRDSG-YDVAVFIRSTVPKPISFEEVTLILLNKEKKLCARKT
Bcer98DRAFT_3031_1(55-176)   IYQFLHKELPRLQENQISLSGIEIEKREGS-YAVAAFIRSSISKPISFEEVTLLLLNKEDELCARKT
BT9727_0783_1(58-179)        IYQFLHKSLPTLQENQISLAGIESKKHENA-YYITTFIRSSVKHPIQFETLTLSLLNKNGETCARQT
GK3171_1(46-167)             VYRFYHEQLPPLQPNQISISGVKLVEYNDG-FVAVAILRNTLPKPVRFERIRLLLLDEDGTAIARKE
B14911_04439_1(59-182)       VLRFLNNELPPLLPNQISLAGIELQQDGGS-VTVAAFVRSSLSKAVEFKKTHLLLVGPDEEILARKE
BA0881_2(185-293)            ALRNFVDNLTPPNDGEINFLGLQAARKENGDLHTTLLIRNGCKDNIQLEQLPLHIEDATGAVVVKGA
BCZK0785_2(185-293)          ALRNFVDNLTPPNDGEINFLGLQAARKENGDLHTTLLIRNGCKDNIQLEQLPLHIEDATGAVVVKGA
BCE_G9241_0886_2(185-293)    ALRNFVDNLTPPNDGEINFLGLQAARKENGDLHTTLLIRNGCKENIQLEQLPLHIEDATGAVVVKGA
Bcer98DRAFT_3031_2(185-293)  ALRNFVESLTPPQNGELNFLGLQAAQKENGDLHATILIRNGCKRNIQLKQLPLHIEDASGEIVVKGA
BT9727_0783_2(188-295)       KLQEIIANLDPPEEDEINFRGLNAVVEENGDLNATILIRNGYNKNITLEQLPLHISDRSESTVAERI
B14911_04439_2(191-305)      KLKQMVEQMDPPKIGEINFMGIQAKVADNEDLQVTLLIRNGNDQNVMLQQLPLQVEDATSEVIAKGG
GK3171_2(176-297)            QLQALVDSVPPPAPGEVNFMGIEAKQLPSGELGVTLLIRNGSDKHIHFEQIPLEVRDYAGDIVARGL
NT01CX_1557_2(164-276)       QYEKFLKELPLLREGQVTMNAYDVYTNEDDGIAVELVIRNGRHNGVDIKRIPLSIYDKDKKLVASGT
DredDRAFT_0533_2(156-262)    QFTTFLKKLPSVQEGSINIDTYSIEKNNDGSLTVAIVLRHRLAKPTVLSRFQFGIVDTNKSIVARAA
CTC00525_2(170-279)          VFKEFLESLPKLERGQGSISVFTITQYENGDLLMTLLVRNATDEAVTMTKMPITLKTQKGETILSGV
CTC00525_1(36-159)           LEEELREVIPKVEEGKINIAGIYAFDQGDK-VEVKAYLANGLSQKINFEDVPIYIINSKEEKLAYQV
NT01CX_1557_1(31-154)        CLEEELEALPAIKEGELDVN-VDFFFDLGDRYEASIFIRNGLSTGVNLEKIPFIVLDKDEKEVGRKI
DredDRAFT_0533_1(25-147)     LMQEEINNLPQITDGTVAIDSIYTVNWEDK-IEIGFYLRNVTSHKICFTQTPLKILNPKGEVLASVT

consensus/80%                hhp.hhcpLs..ppsplsh.ulph.ptpss.htsshhlRsshtcslphcplsLhl.stptphhscth

Secondary structure            EEEEE

BA0881_1(55-176)             FNLSALGDIPSNVNMPFIFTFEQETI-TDAALSQTDWELAFELESK--HTLDLDPSWEA   122
BCZK0785_1(55-176)           FNLSALGDIPANVNMPFIFTFEQETI-TDAALSQTDWELAFELESK--HALDLDPSWEA   122
BCE_G9241_0886_1(55-176)     FNLSALGDIPANVNMPFIFTFEQETI-TDADLSQTDWELAFELESK--HVLDLDPSWEA   122
Bcer98DRAFT_3031_1(55-176)   FNLSDIGDIPANVNMPWVFTFDEETI-TDAELSQTDWQLAFELEGE--HRLDLDPTWET   122
BT9727_0783_1(58-179)        FDLSHLEGIPSNVNMPWTFVFEENSI-TEATLSNEDWQLVFELQGK--HSLDLDPIWQE   122
GK3171_1(46-167)             FDMSPFGELPPMTARPWRFLFAAEDK-LVDQLPADGWKIAFELTPR--HRLDLEESWEQ   122
B14911_04439_1(59-182)       FDLTEIGEIPAKSSRPWNFTFNSSDL-LTDSIPAEGWKLAFEIRNNEEHRLDLDEAWEN   124
BA0881_2(185-293)            FTLPNLEIKAN-TTKPWSFVFPASSI-LKEDMDLSSWKALVPQD---------------   109
BCZK0785_2(185-293)          FTLPNLEIKAN-TTKPWSFVFPASSI-LKEDMDLSSWKALVPQD---------------   109
BCE_G9241_0886_2(185-293)    FTLPNLEIKAN-TTKPWSFVFPASSI-LKEDMELSSWKALVPQD---------------   109
Bcer98DRAFT_3031_2(185-293)  FTLPNLEIKAN-STKPWSFIFPVSFV-LKKEMDLSTWKAIVPQD---------------   109
BT9727_0783_2(188-295)       FVLKDFQIKAN-STKPWTFTFPADSV-SKEPIDLSKWKAFIPQ----------------   108
B14911_04439_2(191-305)      FQLDKFELKAN-TSKPWTFIFPKSLL-LKDNPDLSSWKAYPLQQQVQTEI---------   115
GK3171_2(176-297)            FPCH-LEVKAH-TSKPWTFLFPPELL-HKAEPDWTSWKVTIPSSPAQSEKQETPSSDE-   122
NT01CX_1557_2(164-276)       FYLEDASLNPI-SAKVYLFTFSKDEL-LREDYNLKNWTIQFLLNSNVN-----------   113
DredDRAFT_0533_2(156-262)    FVIEQYILEPG-MFLLRSFKFTPETI-VNSDADINQCSIAFL-----------------   107
CTC00525_2(170-279)          FDIENFTVNPY-KARVLSLIFKKEVVNIEEDFDLSTCKIIFERE---------------   110
CTC00525_1(36-159)           FDLSEEGDIPSGKAIPVKLNFNKQNI-LVDKIPQDDWKVVFGGNDVKGVRYVNIELESI   124
NT01CX_1557_1(31-154)        FNLREVGEIPARSVRPWKIYFEKDEL-NVEGINLKDLKIVFDSRIKAAGVVNVQYENLP   124
DredDRAFT_0533_1(25-147)     INLSDMGDIPAYSVRPWRFYLGKEDL--TLDNSLKDLKIAFNSRNIPPYMLVIEDRLPE   123

consensus/80%                FsLptht.hss.sshPa.FhF.tppl.hptphs.psWchhh..p...............
```

Figure 3: BA0881 is homologous to proteins GBAA0881 *Bacillus anthracis* str. "*Ames* Ancestor," BAS0837 from *Bacillus anthracis* str. Sterne and Bant_01001534 from *Bacillus anthracis* str. A2012.

This domain is characterized by conserved sequence motifs; YEFF, RGD, FTY, GKD, and FVEH. We refer to this 111-amino-acid region as the YEFF domain. The pairwise sequence identities corresponding to the YEFF domain varied between 36–96%. The consensus secondary structure predicted for this domain suggests mainly β-strands and the conserved sequence motifs, that is, YEFF and FTY are associated with β-strands; see Figure 4. The representative domain architecture of proteins comprising this domain is shown in Figure 17. It is intriguing that each domain comprises RGD sequence motif which is found in the proteins of extracellular matrix. Many viruses enter their host cells via the RGD motif—integrin interaction and synthetic peptides containing this RGD motif are active modulators of cell adhesion [30]. The RGD motif was originally identified as the sequence within fibronectin that mediates cell attachment. This motif has now been found in numerous other proteins and supports cell adhesion. The integrins, a family of cell surface proteins, act as receptors for cell adhesion molecules. A subset of the integrins recognizes the RGD motif within their ligands, the binding of which mediates both cell substratum and cell-cell interactions [31]. The presence of RGD motif and SLH domain implies that the YEFF domain comprising proteins is also present on the cell surface and mediates protein-protein interactions.

3.4. 109-amino-acid-residue IMxxH domain

The 266-amino-acid-residue protein corresponding to the GENE_ID BA1021 and described as hypothetical protein comprises a 109-amino-acid-residue region as two copies. Further BLAST searches using sequence corresponding to the region (4–112) as a query identified 22 proteins (see Table 1(d)) that are described as either conserved or hypothetical proteins. This domain region occurs as two copies in all the proteins of *B. anthracis*, *B. cereus, B. thuringiensis, Bacillus weihenstephanensis C. acetobutylicum, C. perfringens, C. tetani, C. thermocellum, Desulfitobacterium hafniense, Clostridium phytofermentans*, and *Alkaliphilus metalliredigenes*, and as single domain in the 171-amino-acid-residue protein BcerKBAB4DRAFT_0307. The length of proteins varied between 171 to 321-amino-acid residues. The multiple sequence alignment corresponding to this domain identified the characteristic sequence motifs; IMxxH, REA, and we refer to this as the IMxxH domain. The IMxxH sequence motif occurs at the N-terminal region of the domain. The

```
Secondary structure                  EEEE                 EEEE     EEEE             EEE

EF0374(62-172)            ILSS--TDWQGTKVYDKNNNNLTAENANFIGLAKYDGETGFYEFFDKETGETRGDEGTFFVTD---DGE
EF0375(58-168)            ILSG--TDWQGTRVYDAAGNDLTAENANFIGLAKYDGETGFYEFFDKNTGETRGDEGTFFVTG---DGT
EF0376(59-172)            GLSE--KDWAGTRVYDRNGNDLTDENQNLLHAIKFDATTSFYEFFDKETGESTGDEGTFFMTAGITDVS
BA5326(58-168)            ILSD--TNWQGTRVYDKDKNDVTKENANFIGLAKYDAKSGRYEFFDAKTGASRGDKGTFFITN---DGK
BCZK4809(58-168)          ILSD--TNWQGTRVLDKDKNDLTKENANFIGLAKYDAKSGRYEFFDAKTGASRGDKGTFFITN---DGK
BT9727_4791(58-168)       ILSD--TNWQGTRVYDKDKNDVTKENANFIGLAKYDAKSGRYEFFDAKTGASRGDKGTFFITN---DGK
BC5098(58-168)            ILSE--TNWQGTRVYDKDKNDLTKENANFIGLAKYDAKSGRYEFFDAKTGASRGDKGTFFVTN---DGK
RBTH_06214(58-168)        ILSK--TNWQGTRVYDKDKNDLTKENANFIGLAKYDAKSGRYEFFDAKTGASRGDKGTFFVTN---DGK
BA3695(247-357)           ILGE--TNWQGTKVYDKDHNDVTKENQNFIGLAKYDAKTARYEFFNASTGESRNDSGTFFITN---DGK
Bant_01004347(247-357)    ILGE--TNWQGTKVYDKDHNDVTKENQNFIGLAKYDAKTARYEFFNASTGESRNDSGTFFITN---DGK
BT9727_3386(247-357)      ILGE--TNWQGTKVYDKDHNDVTKENQNFIGLAKYDAKMARYEFFNASTGESRNDSGTFFITN---DGK
BCZK3337(229-339)         ILGE--TNWQGTKVYDKDHNDVTKENQNFIGLAKYDAKTARYEFFNASTGESRNDSGTFFITN---DGK
BCE_G9241_3590(229-339)   ILGE--TNWQGTKVYDKDHNDVTKENQNFIGLAKYDAKTARYEFFNAKTGESRNDSGTFFITN---DGK
EF0376(223-336)           FDGTPQLLWNGTKVVDKDGNDVTSANQNFISLAKFDQDSSKYEFFNLQTGETRGDYGYFKVGN---QNK
EF0375(199-310)           ILGT--TLWNGTKVVDKNGNDVTAANQNFISLAKFDPNTSKYEFFNLQTGETRGDFGYFQVVD---NNK
EF0374(203-314)           ILGA--TLWNGTKVLDEDGNDVTEANKMFISLAKFDNKTSKYEFFDLETGKTRGDFGYFQVID---NNK
BA3695(388-499)           ILSS--TLWNGTVVLDEQGNNVTKYNSNLISLAKYDKNTNKYEFFNVNTGESRGDYGFFDVVH---DNK
BT9727_3386(388-499)      ILSS--TLWNGTVVLDEQGNNVTKYNSNLISLAKYDKNTNKYEFFNVNTGESRGDYGFFDVVH---DNK
Bant_01004347(388-499)    ILSS--TLWNGTVVLDEQGNNVTKYNSNLISLAKYDKNTNKYEFFNVNTGESRGDYGFFDVVH---DNK
BCZK3337(370-481)         ILSS--TLWNGTVVLDEQGNNVTKYNSNLISLAKYDENTNKYEFFNVNTGESRGDYGFFDVVH---GNK
BCE_G9241_3590(370-481)   ILSS--TLWNGTVVLDDQGNDVTKYNSNLISLAKYDKNTNKYEFFNVNTGESRGDYGFFDVVH---GNK
BA5326(199-310)           ILGG--TLWHGTKVLDEAGNDVTQFNSNFISLAKFDDKSNKYEFFNSETGQSRGDYGYFDVLH---ENK
BCZK4809(199-310)         ILGG--TLWHGTKVLDEAGNDVTQFNSNFISLAKFDDKFNKYEFFNSETGQSRGDYGYFDVLH---ENK
BT9727_4791(199-310)      ILGG--TLWHGTKVLDETGNDVTQFNSNFISLAKFDDKSNKYEFFNSETGQSRGDYGYFDVLH---ENK
BC5098(199-310)           ILGG--TLWHGTKVLDEAGNDVTQFNSNFISLAKFDDKSNKYEFFNSETGQSRGDYGYFDVVH---ENK
RBTH_06214(199-310)       ILGG--TKWHGTKILDEAGNDVTQFNSNFISLAKFDDKSNKYEFFNSETGQSRGDYGYFDVVH---ENK

consensus/80%             ILut..T.WpGT+VhDcstNDlTp.NtNhIuLAKaDtpos+YEFFshpTGpSRGD.GhF.lsp...-sK


Secondary structure        EEEEE       EEEE        EEE             EEEEEE

EF0374(62-172)            KRILISDTQN-YQAVVDLTEVTKDKFTYKRMGKDKDGKDVEVFVEHIP   111
EF0375(58-168)            KRILISRTQN-YQAVVDLTEVSKDKFTYKRLGKDKLGNDVEVYVEHIP   111
EF0376(59-172)            RLVIISETKN-YQGVYPLRTLYQDTFTYRQMGKDKNGNDIEVFVENKA   114
BA5326(58-168)            KRILISESMK-YQAVVDMTKLNKNVFTYKRMGKDANGNDVEVFVEHVP   111
BCZK4809(58-168)          KRILISESMK-YQAVVDMTKLNKNVFTYKRMGKDANGNDVEVFVEHVP   111
BT9727_4791(58-168)       KRILISESMK-YQAVVDMTKLNKNIFTYKRMGKDANGNDVEVFVEHVP   111
BC5098(58-168)            KRILISESMK-YQAVIDMTKLNKNVFTYKRMGKDANGKDVEVFVEHVP   111
RBTH_06214(58-168)        KRILISESMK-YQAVVDMTKLNKNVFTYKRMGKDANGKDVEVFVEHVP   111
BA3695(247-357)           KRVLISETQN-YQAVVELTQLDKEKFTYKRMGKDAKRNDVEVFVEHIP   111
Bant_01004347(247-357)    KRVLISETQN-YQAVVELTQLDKEKFTYKRMGKDAKRNDVEVFVEHIP   111
BT9727_3386(247-357)      KRVLISETQN-YQAVVELTQLDKEKFTYKRMGKDAKGNDVEVFVEHIP   111
BCZK3337(229-339)         KRVLISETQN-YQAVVELTQLDKEKFTYKRMGKDAKGNDVEVFVEHVP   111
BCE_G9241_3590(229-339)   KRVLISETQN-YQAVVELTQLDKEKFTYKRMGKDVKGNDVEVFVEHIP   111
EF0376(223-336)           FRAHVSIGTNRYGAVLELTELNDNRFTYTRMGKDNEGNDIQVYVEHEP   114
EF0375(199-310)           IRAHVSIGTNRYGAALELTELNNDRFTYTRMGKDNAGNDIQVFVEHEP   112
EF0374(203-314)           IRAHVSIGDNKYGAALELTELNDKRFTYTRMGKDNNGKEIKVFVEHEP   112
BA3695(388-499)           IRAHVSLGNNKYGAVLELTELNKEKFTYTRMGKDANGKDIKIFVEHEP   112
BT9727_3386(388-499)      IRAHVSLGNNKYGAVLELTELNKEKFTYTRMGKDANGKDIKIFVEHEP   112
Bant_01004347(388-499)    IRAHVSLGNNKYGAVLELTELNKEKFTYTRMGKDANGKDIKIFVEHEP   112
BCZK3337(370-481)         IRAHVSLGNNKYGAVLELTELNKAKFTYTRMGKDANGKDIKIFVEHEP   112
BCE_G9241_3590(370-481)   IRAHASLGNNKYGAVLELTELNKEKFTYTRIGKDANGKDIKIFVEHEP   112
BA5326(199-310)           IRAHVSIGNNKYGAALELTELNKNKFTYKRTGKDQAGNDITIFVEHEP   112
BCZK4809(199-310)         IRAHVSIGNNKYGAALELTELNKNKFTYKRTGKDQAGNDITIFVEHEP   112
BT9727_4791(199-310)      IRAHVSIGNNKYGAALELTELNKNKFTYKRTGKDQAGNDITIFVEHEP   112
BC5098(199-310)           IRAHVSIGNNKYGAALELTELNKNKFTYKRTGKDQAGKDITIFVEHEP   112
RBTH_06214(199-310)       IRAHVSIGNNKYGAALELTELNKNKFTYKRTGKDQAGKDITIFVEHEP   112

consensus/80%             hRhhlS.spN.YtAsl-LTpLsKppFTYpRhGKDttGpDlplFVEH.P
```

Figure 4: BA3695 is homologous to proteins GBAA3695 from *Bacillus anthracis* str. "*Ames* Ancestor" and BAS342 from *Bacillus anthracis* str. *Sterne*. BA5326 is homologous to proteins GBAA5326 from *Bacillus anthracis* str. "*Ames* Ancestor," BAS4948 from *Bacillus anthracis* str. Sterne and Bant_01000199 from *Bacillus anthracis* str. A2012.

pairwise sequence identities corresponding to the IMxxH domain varies between 5–98%. The secondary structure corresponding to IMxxH domain is predicted to comprise four α-helices as shown in Figure 5. The representative domain architecture corresponding to proteins comprising this domain is shown in Figure 18.

3.5. 103-amino-acid-residue VxxT domain

The 349-amino-acid-residue protein corresponding to the GENE_ID BA4716 and described as germination protein comprises a 103-amino-acid-residue region as two copies. Further BLAST searches using sequence corresponding to the region (67–169) as query identified 23 proteins (see Table 1(e)). The proteins comprising this domain are described as germination proteins as the *Bacillus anthracis* is an endospore-forming bacterium. This domain region occurs twice in proteins of *B. anthracis* str. *Ames, B. cereus, B. clausii, B. thuringiensis*, *B. thuringiensis serovar israelensis*, *Alkaliphilus metalliredigene*, and *Bacillus weihenstephanensis* genomes and only once in the proteins of *Syntrophomonas wolfei str. Goettingen, Moorella thermoacetica, Clostridium thermocellum, B. subtilis*, and *Pelotomaculum thermopropionicum* genomes. The length of proteins varied between 195 to 377-amino-acid residues. The multiple sequence alignment corresponding to this domain identified VxxT as sequence motif. This sequence motif occurs in the N-terminal region of each protein and the pairwise sequence identity

```
Secondary structure                  HHHHHHHHHHHHHHHHHH         HHHHHHHHHHHHHHHHHHHH

BCE_G9241_1042_1(21-129)          ERSLNEIRFWSRIMKEHSLFLRLGFRCEDTQLIEEANQFYRLFEHIEQIAHSYTNETDPEQ-----IKRF
BCZK0933_1(21-129)                ERSLNEIRFWSRIMKEHSLFLRLGFRCEDTQLIEEANQFYRLFEHIEQIAHSYTNETDPEQ-----IKRF
BT9727_0941_1(21-129)             ERSLNEIRFWSRIMKEHSLFLRLGFRCEDTQLIEEANQFYRLFEHIEQIAHSYTNETDPEQ-----IKRF
BA1021_1(4-112)                   ERSLNEIRFWSRIMKEHSLFLRLGFRCEDTQLIEEANQFYRLFEHIEQIAHSYTNETDPEQ-----IKRF
BAS0955_1(21-129)                 ERSLNEIRFWSRIMKEHSLFLRLGFRCEDTQLIEEANQFYRLFEHIEQIAHSYTNETDPEQ-----IKRF
RBTH_03050_1(21-129)              ERSLNEIRFWSRIMKEHSFFLRLGFRCEDTQLIEEANQFYRLFEHIEQIAHSYTNETDPEQ-----IKRF
BC1029_1(21-129)                  ERSLNEIRFWSRIMKEHSFFLRLGFRCEDTQLIEEANQFYRLFEHIEQIAHSYTNETDPEQ-----IKRF
BcerKBAB4DRAFT_3543_1(21-129)     ERSLNEIRFWSRIMKEHSFFLRLGFRCEDTQLIEEANQFYRLFEHIEQIAYSYTNETDPGQ-----IKRF
Bcer98DRAFT_1038_1(42-147)        EKSLTENRFWLRIMKEHALFLGEGFNRKDTNLIQQVDQFFHLFDRHLQKAFSIP--QTVQA-----VRQL
CTC02189(189-294)                 RYAYEQETFWNRIMAEHAKFIRGLLDPTEDALIDTANNFGKEFDELTR---EAKRAMYKTM----PISKV
CbeiDRAFT_3331(190-295)           REAYEQEAFWNRIMAEHSKFIRGLLDPTEDELINTANNFGHQFDILTR---EARAAMNKSI----PISKV
ClosDRAFT_1658(189-294)           KEIYEQELFWNRIMAEHSKFIRGLLDPTEDELIHIANDFAKEFDALTA---AVEEAIEKCL----PIDKI
CtheDRAFT_1311(189-294)           KEAYELQFFWNRQMAEHAKFIRGLLDPTENDLINQANDFGNEFDQLTA---EAKAAMDATS----PMAKV
CdifQ_02001573(138-241)           KNAKEIELFWDHIMMEHALFMRGLLDPSEGELINTSNDFAIKFNELIE---KTN--EMTDS----NIKNI
CD1511(189-291)                   KNAKEIELFWDHIMMEHALFMRGLLDPSEGELINTSNDFAIKFNELIE---KTN--EMTDS----NIKNI
CPE0158_2(188-291)                VNISKTEAFWNEIMMEHSLFIRGLLDPSEYELINTAHEFAFEFNELIQ---QLN--NVTNV----TIDNV
CPF_0149(188-291)                 VNISKTEAFWNEIMMEHSLFIRGLLDPSEYELINTAHEFAFEFNELIQ---QLN--NVTNV----TIDNV
CphyDRAFT_3436(189-292)           EDLKDDELFWNQIMMEHALFIRGLLDPTENDLIMQADDFASVYADLLD---EAS--TMTER----TMGDL
DhafDRAFT_0725_2(197-302)         CHMVEMQMFWDHIMKEHAEVISHLLDPKEKAMITRADHFAQAYEQLLN---QLGNGTVPDQ----SFRRI
BCZK0933_2(149-260)               DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDFDALMYQAIDLESMKPQSQ-TVPLLDQF
BT9727_0941_2(149-260)            DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDFDALMYQAIDLESMKPQSQ-TVPLLDQF
BA1021_2(132-243)                 DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDFDALMYQAIDLESMKPQSQ-TVPLLDQF
BAS0955_2(149-260)                DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDFDALMYQAIDLESMKPQSQ-TVPLLDQF
BCE_G9241_1042_2(149-260)         DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDFDALMYQAIDLESMKPQSQ-TVPLLDQF
BC1029_2(149-260)                 DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDFDELMYQAIDLESMKPQSQ-TAPLLDQF
RBTH_03050_2(149-260)             DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDFDELMYQAIDLESMKPQSQ-TVPLLDQF
BcerKBAB4DRAFT_3543_2(149-260)    DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDFDELMYQAIDLESMKPQSQ-TVPLLDQF
BcerKBAB4DRAFT_0307(35-147)       DAIISENVFWLRIMMEHSRFIGSLLDQSERNLVHTALKFGDDFEILLNQARDVESMLYQKEPTYPIIGKM
Bcer98DRAFT_1038_2(167-279)       DAIISENVFWLRIMMEHSRFIASLLDQSERNLVHTALKFGDDFEVLLSQARDVESMLYQKQPTYPIIGKM
CAC3450_1(190-295)                QGIIRQEIFWNDIMEDHAEFIRGYLDPSQTSLFNTANNFVRRFDDIEN---ATESLTNNPS----NLNNI
CPE0158_1(9-119)                  TSSLELHLFFMRVMKEHAIFLEAGLGPKNSKLAKELDKCKGNLEKLLFDVVKLSKGRVRQSIVD-SGEVF
DhafDRAFT_0725_1(12-122)          RESLELHLFWARIIKEHLIFLESGFMCKDADWMQEADALKCSFEEILHEANCLADGKVGIEVMK-SGELF
CAC3450_2(9-121)                  RLSLELNLFFLRIVKEHNVIAGASLPPKYAPTLMEILAVNKKLDMLLSKTVALSKGNISREAMN-SSTLI
AmetDRAFT_1908_1(11-115)          NVALFEHQFWLQVLGDHARFILNALSPEEREEIQRAQYFIHIFDQLLE----ESRKSPRGS----ALSKL
AmetDRAFT_1908_2(133-245)         TQPIHYHMVWLLDAAGHSAGIMGDLDMVEKELIRKSGKFTQRFEEFYIKAVEIAGYTRTTLDQFPAFTRF

consensus/80%                     ct.hp..hFa.+IMt-HuhFlthhhcsp-ppLlppAppF.p.F-tl.....phpt.p..pp.....lpph


Secondary structure                HHHHHHH  HHHHHHHHH

BCE_G9241_1042_1(21-129)          NAEVQQAATNIWGFKRKILGLILTCKLPGQNNFPLLVDHTSREA   109
BCZK0933_1(21-129)                NAEVQQAATNIWGFKRKILGLILTCKLPGQNNFPLLVDHTSREA   109
BT9727_0941_1(21-129)             NAEVQQAATNIWGFKRKILGLILTCKLPGQNNFPLLVDHTSREA   109
BA1021_1(4-112)                   NAEVQQAATNIWGFKRKILGLILTCKLPGQNNFPLLVDHTSREA   109
BAS0955_1(21-129)                 NAEVQQAATNIWGFKRKILGLILTCKLPGQNNFPLLVDHTSREA   109
RBTH_03050_1(21-129)              NAEVQQAATNIWGFKRKILGLILTCKLPGQNNFPLLVDHTSREA   109
BC1029_1(21-129)                  NAEVQQAATNIWGFKRKILGLILTCKLPGQNNFPLLVDHTSREA   109
BcerKBAB4DRAFT_3543_1(21-129)     NSEVQQAATNIWGFKRKILGLILTCKLPGQNNFPLLVDHTSREA   109
Bcer98DRAFT_1038_1(42-147)        NEESIQLVYAFRNYKRNLLILIINCKVSGFN-FPLLVDHIAREA   106
CTC02189(189-294)                 TNRSLRATRRIRNFKKQGTEGILDCKIRSII-IPLLADHTLREA   106
CbeiDRAFT_3331(190-295)           TDESLEATKSIRNFKAQGTQGLVECKIKSII-IPLLGDHTLREA   106
ClosDRAFT_1658(189-294)           TDKSLEATKEVRNFNTQGTEGLLDCKIRSII-IPLLGDHVLRES   106
CtheDRAFT_1311(189-294)           TDESLKATEDFRNFKAQGTQAILECKVKSII-IPLLGDHVLREA   106
CdifQ_02001573(138-241)           TEETLNETVEFKDFKEAGASGIEQCKIKSII-LPLLADHVLREA   104
CD1511(189-291)                   TEETLNETVEFKDFKEAGASGIEQCKIKSII-LPLLADHVLREA   104
CPE0158_2(188-291)                THEILKETTRLRDFKEEGTKGIMNCNIKSLI-LPLLSDHVLREA   104
CPF_0149(188-291)                 THETLKETTRLRDFKEEGTKGIMNCNIKSLI-LPLLSDHVLREA   104
CphyDRAFT_3436(189-292)           TCRTLEETIKYRDFKLAGTKGINDCEIRSII-LPLLADHVLREA   104
DhafDRAFT_0725_2(197-302)         TSETIRVTGEFKDFKAAGTDAILCCQLRSLI-LPLLADHVLREA   106
BCZK0933_2(149-260)               LDQNRVSVASLRDFKKTARDLIEQCKIKSII-HPLLADHVFREA   112
BT9727_0941_2(149-260)            LDQNRVSVASLRDFKKTARDLIEQCKIKSII-HPLLADHVFREA   112
BA1021_2(132-243)                 LDQNRVSVASLRDFKKTARDLIEQCKIKSII-HPLLADHVFREA   112
BAS0955_2(149-260)                LDQNRVSVASLRDFKKTARDLIEQCKIKSII-HPLLADHVFREA   112
BCE_G9241_1042_2(149-260)         LDQNRVSVASLRDFKKTARDLIEQCKIKSII-HPLLADHVFREA   112
BC1029_2(149-260)                 LDQNRVSVASLRDFKKTARDLIEQCKIKSII-HPLLADHVFREA   112
RBTH_03050_2(149-260)             LDQNRVSVTSLRDFKKTARDLIEQCKIKSII-HPLLADHVFREA   112
BcerKBAB4DRAFT_3543_2(149-260)    LDQNRVSVTSLRDFKKTARDLIEQCKIKSII-HPLLADHVFREA   112
BcerKBAB4DRAFT_0307(35-147)       NKDSENATVELRNFKKAGLELIQTCQIRSVI-NPLLADHVTREA   113
Bcer98DRAFT_1038_2(167-279)       NKDSENATVELRNFKKAGLELIQTCQIRNVI-NPLLADHVVREA   113
CAC3450_1(190-295)                TRNIYSLVTEFRNFKSTATKGLLACKIKAIM-APLLADHVTREA   106
CPE0158_1(9-119)                  TDYTLETEKKTEHYTGININSKITTMEKDLMC--APKKGIDSKV   111
DhafDRAFT_0725_1(12-122)          TNKTLKAEQKTQELTCIPINSQLTVETMSLHP--YMGVGMGMVP   111
CAC3450_2(9-121)                  TPLTLPSEKVTSALTGVPINTAITSKEISLGYRDYYRTGINMVT   113
AmetDRAFT_1908_1(11-115)          TDQAYGCAQEIRTFKLHLIKRHLVGKIEIGL-PPTFLNHMVNEV   105
AmetDRAFT_1908_2(133-245)         NYQVEGELLLFKKFLRELEALELNQKVLGTL-SALMLDHMAREE   113

consensus/80%                     .tps..tstphhsFKpthhthl.pCcl.u...hPLLsDHs.REA
```

Figure 5: BAS0955 is homologous to proteins BT9727_0941 from *Bacillus thuringiensis* serovar konkukian str. 97-27, BCZK0933 from *Bacillus cereus* E33L, and BCE_G9241_1042 from *Bacillus cereus* G9241. BA1021 is homologous to protein GBAA1021 from *Bacillus anthracis* str. "*Ames* Ancestor." BA0807 is homologous to proteins GBAA0807 from *Bacillus anthracis* str. "*Ames* Ancestor" and BAS0770 from *Bacillus anthracis* str. Sterne.

```
Secondary Structure                    EE                  HHHHHHHHHH

BT9727_4219_1(67-169)             VDKNGYVVPQTLAIPTPKANE----VIQQTLEYLVKDGPVTNLLPN-GFRAVIPANTSMT--LDLKKDG
BCZK4235_1(67-169)                VDKNGYVVPQTLAIPTPKANE----VIQQTLEYLVKDGPVTNLLPN-GFRAVIPANTSMT--LDLKKDG
BA4716_1(67-169)                  VDKNGYVVPQTLAIPTPKANE----VIQQTLEYLVKDGPVTNLLPN-GFRAVIPANTSMT--LDLKKDG
BCE4587_1(67-169)                 VDKNGYVVPQTLAIPTPKANE----VIQQTLEYLVKDGPVTNLLPN-GFRAVIPANTSMT--LNLKKDG
RBTH_05210_1(67-169)              VDKNGYVVPQTLAIPTPKANE----TVKQTLEYLVKDGPVTNLLPN-GFRAVIPANTTMT--LDLKKDG
ABF83609_1(67-169)                VDKNGYVVPQTLAIPTPKANE----TVKQTLEYLVKDGPVTNLLPN-GFRAVIPANTTMT--LDLKKDG
BC4495_1(67-169)                  VDKNGYVVPQTLAIPTPKANE----TVKQTLEYLVKDGPVTNLLPN-GFRAVIPANTTMT--LDLKKDG
BcerKBAB4DRAFT_4089_1(67-169)     VDKNGYVVPQTIAMPTPKANE----VVQQTLEYLVKDGPVTNLLPN-GFRAVLPANTTMT--LNLKKGG
Bcer98DRAFT_3179_1(67-169)        VDKNGYVVPQTLALPIPKQSE----VVKQTLEYLVKDGPVENILPN-GFRAVLPADTTMT--VDLKKDG
gerM(84-184)                      IDKNGYVVAQTLPLPKSES------TAKQALEYLVQGGPVSEILPN-GFRAVLPADTTVN--VDIKKDG
BSU28380_1(84-184)                IDKNGYVVAQTLPLPKSES------TAKQALEYLVQGGPVSEILPN-GFRAVLPADTTVN--VDIKKDG
BL00314_1(87-187)                 IDKNGYVTAQTLPLPKQEG------TAKQALEYLVEGGPVSNILPN-GFRAVLPADTTVN--VDIKEDG
GK2667_1(76-177)                  IDKNGFVVPQTVELPKTQA------VAKQVLEYLVEDGPVSEMLPN-GFRAVIPAGTTVL-GTKLEKDG
BH3070_1(87-186)                  LDENGMVVPQTLPLPKSDG------VLKQSLEYLVEGGPVTNLLPN-GFQAVLPPDTEMS--VNL-EDG
OB2107_1(69-172)                  LDANGMVASQTLELPVPDTNE----VAAQVLEHLVKGGPVTPLLPN-GFQAVLPEGTEVL-GVNLQEDG
B14911_06091_1(82-181)            VDKNGYVVPQTLTLPKTES------VATQALEYLMQNGPVTDMLPN-DFRAVLPADTKIS-VN--VKDK
ABC2653_1(99-200)                 IDSNGLVVPQTLTLPKTDS------VMKQALEYLVEGGPINDILPN-GFRAVLPAGTEVD-IDHLKEEK
SwolDRAFT_2302(77-173)            ADKEELVMERR-EITRTEG------IARSTLQELLK-GPDN---P--AYRNVFPEGTRLL-DINLKPDG
Moth_0516(72-172)                 DSSGNYLVAEKRSIPAVEG------IARATIEELIKGPAPDSK-----LLPTIPKGTVLK-DINIRPDG
MothDRAFT_0979(72-172)            DSSGNYLVAEKRSIPAVEG------IARATIEELIKGPAPDSK-----LLPTIPKGTVLK-DINIRPDG
CtheDRAFT_0840(63-168)            NEDNSKLKLEIRYIPVSETTKSVNHLAEIIVNELIKGPKVAG------LKPTIPEGTKLRSAIKIEGD-
AmetTDRAFT_1640_1(62-164)         RDDKGLLIPVMRRIPWQEG------IAKAALEQLVDQPVLRDDLATIGLLPVLPPGTEVI-GISINEG-
GAA01614(67-167)                  TGSDAYLVREVHQVPFTRE------VAKAALEELINTAPSTPG-----AVRVLPPATKIR-GISIKDG-
BCE4587_2(220-319)                NNKQQYYVPVTRRVVEGKE-----NDYAAIVDELVKGPIHQS------LLNDFNPGVKLI-TNPKLQDG
BA4716_2(220-319)                 NNKQQYYVPVTRRVVEGKE-----NDYAAIVDELVKGPIHQS------LLNDFNPGVKLI-TNPKLQDG
BT9727_4219_2(220-319)            NNKQQYYVPVTRRVVEGKE-----NDYAAIVDELVKGPIHQS------LLNDFNPGVKLI-TNPKLQDG
BCZK4235_2(220-319)               NNKQQYYVPVTRRVVEGKE-----NDYAAIVDELVKGPIHQS------LLNDFNPGVKLI-TNPKLQDG
ABF83609_2(20-319)                NNKQQYYVPVTRRVAEGKE-----NDYAAIIDELVKGPIHQS------LLNDFNPGVKLI-TNPKLQDG
BC4495_2(220-319)                 NNKQQYYVPVTRRVAEGKE-----NDYATIIDELVKGPIHQS------LLNDFNPGVKLI-TNPKLQDG
RBTH_05210_2(220-310)             NNKQQYYVPVTRRVAEGKE-----NDYAAIIDELVKGPIHQS------LLNDFNPGVKLI-TNPKLQDG
BcerKBAB4DRAFT_4089_2(220-319)    NNKQQYYVPVTRRVAEGKE-----NDYSAIVDELVKGPIQGS------LLNDFNPGAKLI-TNPKVENG
Bcer98DRAFT_3179_2(219-318)       NNKRQYYVPVTRRVAEEKE-----NEVETIINELVKGPSHSS------LLNDFNPGVKLV-SEPKIQDG
BSU28380_2(234-336)               NEDSEYYVPVTKRIDNSEK-----DDITAAINELAKGPSKVSG-----LLTDFSEDVKLV-SKPKIKDG
BL00314_2(237-339)                SDKGTYYVPVTKRTSAKEK-----DQVTAAIKELTEGPDNKSG-----LLSDFQGDVKLE-NKPKIEDG
GK2667_2(227-327)                 QGNSTYYVPVTRRVSNKEK-----DDIAAAVNELIQGPEQGSG-----LVGVFQPDAKLV-DAPKYEDG
B14911_06091_2(231-331)           EEGAYYYVPVTKRISAQED-----NQVEAVVKELVKGPSFTSN-----LFTDFMPEVELL-GDPKIENG
BH3070_2(236-335)                 SGDQTYYVPVTRRVNVKD------NSFATAVEELLNGPMVTSP-----LVTDFRNGVELL-DEPKYENG
ABC2653_2(250-349)                NDEDTYYVPVTKRVENVD------NELEAAINELIDGPSLMTN-----LLTEMSGDVELL-NEPKLQNG
OB2107_2(222-322)                 QENNRYYVPVTQYIETNED-----EAIANIIKELIDGPGHQSK-----VVNVFNPEAGLA-SEPTLNNG
AmetDRAFT_1640_2(210-309)         NGEDDFFIPITRGLNVLKA------DTKSVLTALVEGAPVGSG-----LHSEIPYGASIN--DVYVRDG

consensus/80%                     .scptYhVs.Thtlstsct.......htthlc.Llcss.hps......hhsshssssphh...shhp-G


Secondary Structure               EEEEEE                 HHHHHHHHHH         EEEE

BT9727_4219_1(67-169)             TAVIDFSKEMKNYA----KEEERQIVESIAWTLTQFK-EVKQVQFQ   103
BCZK4235_1(67-169)                TAVIDFSKEMKNYA----KEEERQIVESIAWTLTQFK-EVKQVQFQ   103
BA4716_1(67-169)                  TAVIDFSKEMKNYA----KEEERQIVESIAWTLTQFK-EVKQVQFQ   103
BCE4587_1(67-169)                 TAVIDFSKEMKNYA----KEEERQIVESIAWTLTQFK-EIKQVQFQ   103
RBTH_05210_1(67-169)              TAVIDFSKEMKNYA----KEEERQIVESIAWTLTQFT-EIKQVQFQ   103
ABF83609_1(67-169)                TAVIDFSKEMKNYA----KEEERQIVESIAWTLTQFT-EIKQVQFQ   103
BC4495_1(67-169)                  TAVIDFSKEMKNYA----KEEERQIVESIAWTLTQFT-EIKQVQFQ   103
BcerKBAB4DRAFT_4089_1(67-169)     TAVIDFSKEMKNYS----KEEERQIVESVAWTLTQFT-EIKQVQFQ   103
Bcer98DRAFT_3179_1(67-169)        TAVIDFSKEMQNYK----KEEERQIVESVAWTLTQFK-DIKQVKFQ   103
gerM(84-184)                      TAIADFSNEFKNYK----KEDEQKIVQSVTWTLTQFS-SIDKVKLR   101
BSU28380_1(84-184)                TAIADFSNEFKNYK----KEDEQKIVQSVTWTLTQFS-SIDKVKLR   101
BL00314_1(87-187)                 TAIADFSNEFKNYK----AEDEQKIVQAITWTLTQFN-SIDKVKLR   101
GK2667_1(76-177)                  TLIADFSPEFKNYK----PEDEKRILQSITWTLTQFD-NIKRVKIR   102
BH3070_1(87-186)                  VAVVDFSKEFTEYD----GEKEQQILQSITWTLTQFE-NVEKVKLQ   100
OB2107_1(69-172)                  TIIVDLSEEFTQYE----ENQEVQILESVTHTLTQFE-SVHKVKLR   104
B14911_06091_1(82-181)            VATVDFSKEFGDYQ----AEDEEKILESITWTLTQFD-SIEKVKLQ   100
ABC2653_1(99-200)                 LAIVNFSSEFNDYN----LADEKQIFEAVTWTLTQFP-DVEEVKVE   102
SwolDRAFT_2302(77-173)            TCILDFSSELRRLEN---EVEEKQMLDAVCQTLAQFP-AVKQLVFM    97
Moth_0516(72-172)                 LARVDFSKELVANHS-GGSLGESLTVYSIVNTLTQFP-TIKQVQFL   101
MothDRAFT_0979(72-172)            LARVDFSKELVANHS-GGSLGESLTVYSIVNTLTQFP-TIKQVQFL   101
CtheDRAFT_0840(63-168)            VAIVDFTKEFRDNHP-GGKAEERMTIYSVVNSLTELK-EINKVKFL   106
AmetDRAFT_1640_1(62-164)          LSKVDFNEQLLAYQS---EIDENAIVKSIVYTLTEFD-SIDQVQIM   103
GAA01614(67-167)                  LATVDFSRDVLRANT-G-ASGEALGIQSIVNTLTEFP-EVQKVSFL    99
BCE4587_2(220-319)                NLTLNFNENIFINP--DKNMISNYVLKSLVLSLTEKK-GVKSVSIE   100
BA4716_2(220-319)                 NLTLNFNENIFINP--DKNMISNYVLKSLVLSLTEKK-GVKSVSIE   100
BT9727_4219_2(220-319)            NLTLNFNENIFINP--DKNMISNYVLKSLVLSLTEKK-GVKSVSIE   100
BCZK4235_2(220-319)               NLTLNFNENIFINP--DKNMISNYVLKSLVLSLTEKK-GVKSVSIE   100
ABF83609_2(20-319)                NLTLNFNENIFVNP--DKNMISNYVLKSLVLSLTEKK-GVKNISIE   100
BC4495_2(220-319)                 NLTLNFNENIFVNP--DKNMISNYVLKSLVLSLTEKK-GVKNISIE   100
RBTH_05210_2(220-310)             NLTLNFNENIFVNP--DKNMISNYVLKSLVLSLTEKK-GVKNVSIE   100
BcerKBAB4DRAFT_4089_2(220-319)    NITLNFNENIFVNP--DKNMISNYVLKSLVLSLTEKQ-GVKNVSIE   100
Bcer98DRAFT_3179_2(219-318)       KVTLNFNENIYANK--DKNMISNYVLQSLVLSLTEKQ-GVKNVSVE   100
BSU28380_2(234-336)               RVTLDFNQSIFGSADEKTKMISSEVLNSIVLTLTEQP-DVKSVSVK   103
BL00314_2(237-339)                HVTLDFNEAIYGSADGQKKVISDEVLNSIVLTLTELP-DVKSVSVT   103
GK2667_2(227-327)                 KVTLNFNEGIYGSN--KKNVISDVVLNSLVLSLTEQK-GVESVAIT   101
B14911_06091_2(231-331)           LATLDFNESVYGSF--EEKIISQHLLNSLVLSLTEQK-GIESVAVT   101
BH3070_2(236-335)                 VVTLNFNEALLSQM--QATAVSDEIINMLALTLTEQD-GVEKVAIQ   100
ABC2653_2(250-349)                EVVLDFNEAIQSAN--EGSAIPTSVLESLALTLTEQG-GIEKVSIQ   100
OB2107_2(222-322)                 ILEVVFNKEILADS--EQGIIADEVMETMVRTLTEQP-NIDAVDVK   101
AmetDRAFT_1640_2(210-309)         IAYIDFTEEIRNVP--VNEKHQQSLVYELGLTLREVEPSIHQVRIL   100

consensus/80%                     .hhlsFscphhs......p..pp.llpSlshoLTpht.tlcpVphp
```

Figure 6: BA4716 is homologous to proteins GBAA4716 from *Bacillus anthracis* str. "*Ames* Ancestor," BAS4378 from *Bacillus anthracis* str. Sterne, and Bant_01005366 from *Bacillus anthracis* str. A2012. BT9727_4219 is homologous to protein BCZK4235 from *Bacillus cereus* E33L. BA4716 is homologous to protein BL02986 from *Bacillus licheniformis* ATCC 14580.

```
Secondary structure                    EEEEEE              EEEEE                    EEEEE

BC4088_1(47-130)                IKPGEKTEVQALVTQGKEKVTDADDVKFEIWKDGD--EKHEMLDGKHKGKGVYAVEKTFETDG
RBTH_02670_1(47-130)            IKPGEKTEVQALVTQGKEKXTDADDVKFEIWKDGD--EKHEMLDGKHKGKGVYAVEKTFETDG
BCE_G9241_4093_1(45-128)        IKPGEKTEVQALVTQGKERVTDADDVKFEIWKDGD--EKHEMLDGKHKGKGVYAVEKTFETDG
BA4310_1(45-128)                IKPGEKTEVQALVTQGKEKVTDADDVKFEVWKAGD--EKHEMLEGKHKGKGVYAVEKTFETDG
BT9727_3829_1(45-128)           IKPGEKTEVQALVTQGKEKVTDADDVKFEVWKAGD--EKHEMLEGKHKGKGVYAVEKTFETDG
Bant_01004966_1(51-134)         IKPGEKTEVQALVTQGKEKVTDADDVKFEVWKAGD--EKHEMLEGKHKGKGVYAVEKTFETDG
BCE4157_1(45-128)               IKPGEKTEVQALVTQGKEKVTDADDVKFEIWKAGD--EKHEMLEGKHKGKGVYAVEKTFETDG
BCZK3845_1(45-128)              IKPGEKTEVQALVTQGKEKVTDADDVKFEIWKAGD--EKHEMLEGKHKGKGVYAVEKTFETDG
BcerKBAB4DRAFT_2040_1(46-128)   IKPGEKTEVQALVTQGKEKVTDADDVKFEIWKAGD--EKHEMLNAKHKGKGVYAVEKTFETDG
GK0969(45-128)                  IDLNKPTKLACVVTYGGEKVDDANEVKFEVWKHGS--DEREMLEAKHDGDGRYSVEKTFTEAG
BL05305(45-129)                 AAKNEKAVIKATVLYGEEPVADADEVEFECWKAGSK-EDSELIKAKNEGKGVYSMEKAFPEDG
BSU30660(44-127)                VNPGESAAYEAAVSYGDEAVTDADEVEFEVWKEGEK-DASQMFKVKQE-KGVYRLETTFKEDG
OB2488(50-134)                  VETGETIDLTAHVTYGDAPVEDADEVIFEVWTQGNS-DQSVELEGKHQENGTYTASYTFEEEK
B14911_05359_1(53-137)          VELNEEITLSVEVVQGEEAVEDADEVKFEIWQEGNQ-EESEMLPAEHTGKGIYQAAKTFGKDG
BH0678_1(45-129)                LASGENMTFDVLVTQNEAPVEDAREVIVEFWQEGAK-EESDMIESTNEGGGVYRVTYEFPEDG
ABC0230(45-129)                 IEIGEEILLSVQLAQGEVQVEDADEVVFEVWKDQER-DNGTLQEATHQENGVYEITHTFDEDG
ABC4088(44-127)                 LEL-ENIVLEAKVMQGDEPVDDAEEVVFEVWPYDDR-EESEFHEASYAESGLYQAPLALEEAG
BH0983(47-131)                  LIPNTPHELAIHVTQGDENVTDATDIQFEIWQGHDR-EQGELIEASHVEDGIYLVEYEFPEDG
B14911_09907(34-118)            FAAGEDVPIRAVLTQNGEKVAGADYVHFEIWKRDGS-VHYPMEEAADEGEGVYQLTKKFEQDG
ExigDRAFT_1796(51-135)          ADQEKQYRFGATLWQDQKAVKEAEYVHFEIWKADGT-LRYSMEPADETKPGVYSIEKKLPKEG
BAA83944_1(46-130)              LVTDQEESLTVSLSHNGEILSKVDSLHVHIWKHDHT-VAYHFEQLETDQDGAFNLPLTFESDG
BH1853(46-130)                  LVTDQEESLTVSLSHNGEILSKVDSLHVHIWKHDHT-VAYHFEQLETDQDGAFNLPLTFESDG
OB3282(48-131)                  IEAKENTEVTFELSQNGESVSTLDDLSVTTWMVDSE-TTKQLVAENVG-NGEYSVETSFDQDG
BCE_G9241_4093_2(163-245)       IKANAESTMKVHLKQKE-EALTGAEVQLEIWKDGV--EKHEFIPAKEGNKGEYETKHTFKENG
BC4088_2(165-247)               IKANAESTMKVHLKQKE-EALTGAEVQLEIWKDGV--EKHEFIPAKEGNKGEYETKHTFKENG
RBTH_02670_2(165-247)           IKANAESTMKVHLKQKE-EALAGAEVQLEIWKDGV--EKHEFIPAKEGNKGEYETKHTFKENG
BcerKBAB4DRAFT_2040_2(158-240)  IKANAESTMKVHLKQKE-EALSGAEVQLEIWKDGV--EKHEFIPAKEGNKGEYESKHTFKENG
BT9727_3829_2(163-245)          IKANAESTMKVHLKQKE-EALTGAEVQLEIWKDGV--EKHEFIPAKEGNKGEYETKHTFKENG
Bant_01004966_2(169-251)        IKANAESTMKVHLKQKE-EALTGAEVQLEIWKDGV--EKHEFIPAKEGNKGEYETKHTFKENG
BA4310_2(163-245)               IKANAESTMKVHLKQKE-EALTGAEVQLEIWKDGV--EKHEFIPAKEGNKGEYETKHTFKENG
BCE4157_2(163-245)              IKANAESTMKVHLKQKE-EALTGAEVQLEIWKDGV--EKHEFIPAKEGNKGEYETKHTFKENG
BCZK3845_2(163-245)             IKANAESTMKVHLKQKE-EALTGAEVQLEIWKDGV--EKHEFIPAKEGNKGEYETKYTFKEKG
Bcer98DRAFT_3614(94-176)        VKANAESTLKAHVKQKE-EALTKAEVQFEIWKDGV--EKHTFITAKEDNKGEYVGKYTFKESG
B14911_05359_2(187-271)         IHMKQAAGLDVQVDKKDGAPLEKALVKLEIMKEGK--DTPEWVNLKESGEGKYSAEHSFAEAG
BH0678_2(159-242)               IQAGEETTLLIVVEHKD-KPFTGGVLTLEVWQHED--EAHTWLDTEETDVGQYEVSHTFADAG
ExigDRAFT_0574(52-137)          KTMENQKVVFQATALENKKAVNLENVAFEVWKADEKEAVHQKFKAALKKTGTYQAEAKLA-EG

consensus/80%                   lp.stptphpshlpptc.ts.sus-VphElWKtss..-ppphh.ucpttpG.YtsphoFtpsG


Secondary structure              EEEEEEE            EE

BC4088_1(47-130)                VYHIIAHTNARE-MHVMPEVKVAV  84
RBTH_02670_1(47-130)            VYHIIAHTNARE-MHVMPEVKVAV  84
BCE_G9241_4093_1(45-128)        VYHIIPHTNARD-MHVMPEHKVAV  84
BA4310_1(45-128)                VYHIIAHTNARE-MHVMPEVKVAV  84
BT9727_3829_1(45-128)           VYHIIAHTNARE-MHVMPEVKVAV  84
Bant_01004966_1(51-134)         VYHIIAHTNARE-MHVMPEVKVAV  84
BCE4157_1(45-128)               VYHIIAHTNARE-MHVMPEVKVAV  84
BCZK3845_1(45-128)              VYHIIAHTNARE-MHVMPEVKVAV  84
BcerKBAB4DRAFT_2040_1(46-128)   VYHVIAHTNARE-MHVMPEVKVAV  84
GK0969(45-128)                  TYSVVAHVTARD-MHNMPKKDIVA  84
BL05305(45-129)                 HYKVQVHVTAKK-QHTMPVADIKV  85
BSU30660(44-127)                VYTVQSHVTAKK-QHSMPTLKVQV  84
OB2488(50-134)                  VYEMYAHTTAEA-IHSMPFKTVIV  85
B14911_05359_1(53-137)          DYIVQVHVTARD-MHTMPKAEVQA  85
BH0678_1(45-129)                LYFVQPHVTARD-MHRMPLYELTI  85
ABC0230(45-129)                 IYIVQTHVTARD-MHVMPKQMIVA  85
ABC4088(44-127)                 IYMVQVHVTARG-MHVMPTQPLFA  84
BH0983(47-131)                  IYFVQAHVTARG-LHVMPTERLIV  85
B14911_09907(34-118)            VYIIKVHASSGG-SLIMPQKQFVV  85
ExigDRAFT_1796(51-135)          LYYIKVHASSNG-AMIMPTRQFIV  85
BAA83944_1(46-130)              LYYMKVDVTHNG-DTIMPTAQLIV  85
BH1853(46-130)                  LYYMKVDVTHNG-DTIMPTAQLIV  85
OB3282(48-131)                  IYHMKVTASKNN-ATIMPTKQFIV  84
BCE_G9241_4093_2(163-245)       AYKVKVHVRKGE-LHEHKEETIEV  83
BC4088_2(165-247)               AYKVKVHVRKGE-LHEHKEETIEV  83
RBTH_02670_2(165-247)           AYKVKVHVRKGE-LHEHKEETIEV  83
BcerKBAB4DRAFT_2040_2(158-240)  AYKVKVHVRKGE-LHEHKEETVEV  83
BT9727_3829_2(163-245)          SYKVKVHVKKGE-LHEHKEETVEV  83
Bant_01004966_2(169-251)        SYKVKVHVKKGE-LHEHKEETVEV  83
BA4310_2(163-245)               SYKVKVHVKKGE-LHEHKEETVEV  83
BCE4157_2(163-245)              SYKVKVHVKKGE-LHEHKEEXVEV  83
BCZK3845_2(163-245)             SYKVKVHVKKGE-LHEHKEETVEV  83
Bcer98DRAFT_3614(94-176)        KYKVKVHVRKGD-LHEHKEETVEV  83
B14911_05359_2(187-271)         SYTVTVHVENSEGLHEHSDFPLTV  85
BH0678_2(159-242)               EYHVVFHIEDDTGLHEHIHEALIV  84
ExigDRAFT_0574(52-137)          EYEGLYHINDKNGLHHMDKISFVV  86

consensus/80%                   .YhlhsHspttp.hH.h.p.pl.V
```

Figure 7: BA4310 is homologous to proteins GBAA4310 from *Bacillus anthracis* str. "*Ames* Ancestor," BAS3998 from *Bacillus anthracis* str. Sterne, and BT9727_3829 from *Bacillus thuringiensis* serovar konkukian str. 97-27.

```
Secondary structure                    EEEEE            HH   H               HHHHH

BCZK2413_2(120-222)               VYNTGFIGVVFADLCSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLRHNVPAYISTSHLEEQNK
BT9727_2444_2(120-222)            VYNTGFIGVVFADLCSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLRHNVPAYISTSHLEEQNK
BA2665_2(120-222)                 VYNTGFIGVVFADLCSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLRHNVPAYISTSHLEEQNK
Bant_01003317_2(124-226)          VYNTGFIGVVFADLCSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLRHNVPAYISTSHLEEXNK
BCE2700_2(122-224)                VFNTGFIGVVFADLCSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLRHNVPAYISTSHLEEQNK
BCE_G9241_CNI_0263_2(122-224)     VFNTGFIGVVFADLCSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLRHNVPAYISTSHLEEQNK
BcerKBAB4DRAFT_0535_2(120-222)    VFNTGFIGVVFADLSSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLRHNVPAYISTSHLEEQNK
BC2674_2(122-224)                 VFNTGFIGVVFADLSSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLRHNVPAYISTSHLEEQNK
Bcer98DRAFT_0128_2(122-224)       VFNTGFIGVVFADLSSIDRFNFEF---EMNMLFKLMKDMIIPVKELFLRHNIPAYISTSHLETQNK
BA2665_1(16-119)                  ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSSLFKEHNISAYMSCVYLHKQKK
Bant_01003317_1(20-123)           ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSSLFKEHNISAYMSCVYLHKQKK
BCZK2413_1(16-119)                ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSSLFKEHNISAYMSCVYLHKQKK
BT9727_2444_1(16-119)             ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSSLFKEHNISAYMSCVYLHKQKK
BcerKBAB4DRAFT_0535_1(16-119)     ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSALFKEHNTAAYMSCVYLHKQKK
BCE2700_1(16-121)                 ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSSLFKEHNISAYMSCVYLHKQKK
BCE_G9241_CNI_0263_1(16-121)      ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSSLFKEHNISAYMSCVYLHKQKK
BC2674_1(16-121)                  ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSALFKEHNISAYMSCVYLHKQKK
Bcer98DRAFT_0128_1(16-121)        ISNTGFIGSVFIDTLELQKKSYYFSRKKLQIVHHVLDGLAEATSSLFHEHEVAAYISCVYLHKQKK

consensus/80%                     l.NTGFIGsVFhDhhplp+hsa.F...chthlp+lhcsh.hssppLFhcHNlsAYhSssaLccQpK

Secondary structure                 EEEE          HHHHHHHHHHH

BCZK2413_2(120-222)               LGFVLSIKPYDERAEADLYFEAYLKERGLFIG-DEEDDIDK   103
BT9727_2444_2(120-222)            LGFVLSIKPYDERAEADLYFEAYLKERGLFIG-DEEDDIDK   103
BA2665_2(120-222)                 LGFVLSIKPYDERAEADLYFEAYLKERGLFIG-DEEDDIDK   103
Bant_01003317_2(124-226)          LGFVLSIKXYDERAEADLYFEAYLKERGLFIG-DEEDDIDK   103
BCE2700_2(122-224)                LGFVLSVKPYDERAEADLYFEAYLKERGLFIG-DEEDDIDK   103
BCE_G9241_CNI_0263_2(122-224)     LGFVLSVKPYDERAEADLYFEAYLKERGLFIG-DEEDDIDK   103
BcerKBAB4DRAFT_0535_2(120-222)    LGFVLSVKPYDERAEADLYFETYLKERGLFIG-DEEDDIDK   103
BC2674_2(122-224)                 LGFVLSVKPYDERAEADLYFEAYLKERGLFIG-DEEDDIDK   103
Bcer98DRAFT_0128_2(122-224)       VGFVLSIKPYDERAEADLYFETYLKERGLFIG-DEEDEMDK   103
BA2665_1(16-119)                  IGFVLSTKPFEQ-SDGVAYFINYLIEKNFYG--NEEVEYQE   104
Bant_01003317_1(20-123)           IGFVLSTKPFEQ-SDGVAYFINYLIEKNFYG--NEEVEYQE   104
BCZK2413_1(16-119)                IGFVLSTKPFEQ-SDGVAYFINYLIEKNFYG--NEEVEYQE   104
BT9727_2444_1(16-119)             IGFVLSTKPFEQ-SDGVAYFINYLIEKNFYG--NEEVEYQE   104
BcerKBAB4DRAFT_0535_1(16-119)     IGFVLSTKPFEQ-SDGVSYFINYLIEKNFYG--NEEVEYQE   104
BCE2700_1(16-121)                 IGFVLSTKPFEQ-SDGVAYFVNYLIEKNFYGNHDEDVEYQE   106
BCE_G9241_CNI_0263_1(16-121)      IGFVLSTKPFEQ-SDGVAYFVNYLIEKNFYGNHDEDVEYQE   106
BC2674_1(16-121)                  IGFVLSTKPFEQ-SDGVAYFVNYLIEKNFYGGHDEDVEYQE   106
Bcer98DRAFT_0128_1(16-121)        IGFVLSTKLFEQ-TDGIAYFKNYLIEKNFYGKTDQEVEYQE   106

consensus/80%                     lGFVLShKPa-p.u-ushYF.sYLhE+shah..sEEs-hpc
```

Figure 8: BA2665 is homologous to proteins GBAA2665 from *Bacillus anthracis* str. "*Ames* Ancestor," BAS2482 from *Bacillus anthracis* str. Sterne. BT9727_2444 is homologous to protein BCZK2413 from *Bacillus cereus* E33L.

varied between 11–98%. The secondary structure is predicted to comprise two α-helices and three β-strands as shown in Figure 6. The representative domain architecture corresponding to proteins comprising this domain is shown in Figure 19.

3.6. 84-amino-acid-residue ExW domain

The 246-amino-acid-residue protein corresponding to the GENE_ID BA4310 and described as hypothetical protein comprises an 84-amino-acid-residue region as two copies. Further BLAST searches using sequence corresponding to the domain (45–128) as a query identified 25 proteins (Table 1(f)) that are described as either conserved or hypothetical proteins. This domain region occurs as two copies in proteins of *B. anthracis* str. *Ames, B. cereus, B. halodurans* (GENE_ID BH0678), *B. thuringiensis*, *B. thuringiensis serovar israelensis*, *Geobacillus kaustophil*us, *Bacillus weihenstephanensis*, and *Exiguobacterium sibiricum* genomes and as single copy in proteins of *B. clausii*, *B. halodurans* (GENE_ID BH0983), *B. licheniformis, B. subtilis, Exiguobacterium sp.*, and *Oceanobacillus ihenyensis* genomes. The length of proteins varied between 142 to 273-amino-acid residues. The multiple sequence alignment corresponding to this domain identified ExW sequence motif. The pairwise sequence identities corresponding to the ExW domain varied between 14–98%. The secondary structure of this domain is predicted to comprise five β-strands and the conserved sequence motif is associated with one of the β-strands as shown in Figure 7. The representative domain architecture corresponding to proteins comprising this domain is shown in Figure 20.

3.7. 104-amino-acid-residue NTGFIG domain

The 232-amino-acid-residue protein corresponding to the GENE_ID BA2665 and described as a hypothetical protein comprises a 104-amino-acid-residue region as two copies in tandem. Further BLAST searches using sequence corresponding to the region (16–119) as query identified 9 hypothetical proteins comprising this domain from organisms such as *B. anthracis*, *B. thuringiensis, Bacillus weihenstephanensis*, and *B. cereus*. The protein corresponding to the GENE_ID BCZK2413 of *B. cereus* is described as group-specific protein. The list of 9 proteins comprising this domain is shown in Table 1(g). The length of proteins varied between 232 to 236-amino-acid residues. This domain

```
Secondary structure                                         HHHHHHHH

BT9727_3378_2(139-176)          YKTTISAQFEYNRFTRDFFEDPNNKGKSKADAIAAWNE  38
BcerKBAB4DRAFT_0944_2(139-176)  YKTTISAQFEYNRFTRDFFEDPNNKGKSKADAIAAWNE  38
BA3686_2(139-176)               YKTTISAQFEYNRFTRDFFEDPNNKGKSKADAIAAWNE  38
BCZK3328_2(139-176)             YKTTISAQFEYNRFTRDFFEDPNNKGKSKADAIAAWNE  38
BCE_G9241_3579_2(139-176)       YKTTISSQFEYNRFTRDFFEDPNNKGKSKADAIAAWNE  38
BCE3645_2(139-176)              YKTTISPQFEYNRFTRDFFEDPNNKGKTKADVIAAWNE  38
RBTH_03615_2(139-176)           YKTTIGAQFEYNRFTRDFFEDPNNKGKAKADAIAAWNE  38
BC3626_2(139-176)               YKTTIGTQFEYNRFTRDFFEDPNNKGKAKADAIAAWNE  38
B14911_25780_2(138-175)         YKSEIGRQFEYNQFIRDYYADQKNQGKSRAEAIAAWML  38
RBTH_03615_1(94-129)            FKEKIGTNFRFTVALQKFFK--ENVGKTYEDAVAFWHE  36
BC3626_1(94-129)                FKEKIGTNFRFTVALQKFFK--ENVGKTYEDAIAFWHE  36
BT9727_3378_1(94-129)           FKEKIGANFRFTVALQKFFK--ENIGKTYEDAVAFWHE  36
BCZK3328_1(94-129)              FKEKIGANFRFTVALQKFFK--ENIGKTYEDAVAFWHE  36
BA3686_1(94-129)                FKEKIGANFRFTVALQKFFK--ENIGKTYEDAVAFWHE  36
BCE_G9241_3579_1(94-129)        FKEKIGANFRFTVALQKFFK--ENVGKTYEDAITFWYE  36
BCE3645_1(94-129)               FKEKIGANFRFTVALQKFFK--ENVGKTYEDAITFWYE  36
BcerKBAB4DRAFT_0944_1(94-129)   FKEKIGANFRFTVALQKFFK--ENVGKTYEDAITFWYE  36
B14911_25780_1(93-128)          FKSVIGSHFHFSTYIQDYFK--HNPGKTYNDAVSAWHE  36

consensus/80%                   aKppIuspFcashhhpcFFc..pNhGKohtDAluhW.E
```

Figure 9: BT9727_3378 is homologous to protein BCZK3328 from *Bacillus cereus* E33L. BA3686 is homologous to proteins GBAA3686 from *Bacillus anthracis* str. "*Ames* Ancestor," BAS3417 from *Bacillus anthracis* str. Sterne, and Bant_01004341 from *Bacillus anthracis* str. A2012.

```
Secondary structure                                       EEE         EEEE           EEE

BAS1577_2(128-220)      AKKYDTQVSLAPAVKNIVILNND-DADDIVRVTGLESGDVVKVYGEATGG-EVIEKATVQG
BA1701_2(126-218)       AKKYDTQVSLAPAVKNIVILNND-DADDIVRVTGLESGDVVKVYGEATGG-EVIEKATVQG
RBTH_03882_10(898-990)  AVKYESQVTAEPVGGNIVVLNND-GAADIVRVTGLTAGDKVSVYNEETVQ-EAIGTATVAE
BAS1577_1(33-127)       AAEVAIVKTKAVTVDAITVANNEKEAEDTIKVTGLVTGDIVKVYDAASKGKELGTTK-VAE
BA1701_1(31-125)        AAEVAIVKTKAVTVDAITVANNEKEAEDTIKVTGLVTGDIVKVYDAASKGKELGTTK-VAE
RBTH_03882_2(610-705)   VKYEAEPTTVAPAVEKITVSNNKVEAEDTITVSELKKGDIVRVYEASKGGEAIVTSEAVAE
RBTH_03882_3(802-897)   VKYEAEPTTVAPAVEKITVSNNKVEAEDTITVSELKKGDIVRVYEASKGGEAIVTSEAVAE
RBTH_03882_1(418-513)   VKYEAEPTTVAPAVEKITVSNNKVEAEDTITVSELKKGDIVRVYEASKGGEAIVTSEAVAE
RBTH_03882_8(226-320)   VKYEAEPTTVAPAVEKITVSNNKVGNADAITVSKLKKGDIVRVYEASKGGAAIAASEAVAE
RBTH_03882_6(33-128)    AAEVTSAKTAALSVEKANIINNKKGETDTITVSELKKGDIVRVYEASKGGEAIATSEAVAE
RBTH_03882_4(321-416)   AVKYESQVTVAPAVDTVKVANNKAGDADTITVSGVAEGDLVRVYDASTEG-KELGNATVAK
RBTH_03882_5(706-800)   AVKYESQVTVAPAVDTVKVANNKAGDADTITVSEVTEGDVVKVYDASTEG-KELGNATVAK
RBTH_03882_7(514-608)   AVKYESQVTVAPAVDTVKVANNKAGDADTITVSGVAEGDLVRVYDASTEG-KELGNATVAK
RBTH_03882_9(129-224)   AMKYESEVTVAPAVDTVKVANNKAGDADTITVSELAPGDIVKIYDASTGGNLKATSAAVAE
DSY3134_1(51-142)       VPFSEPLKTTTP--SAIEVRNYIEGIRDRVTVSSLEEGDIVKIYPSEESN-TPSGTEAVKA
DSY3134_2(150-240)      PIPWLIYGHTGNWGEDVKLPRTPFDQSK-ASYPAY-PIDANGISDDNPLGIIYNQHIIIKG

consensus/80%           sh..t..hThAssVcplpl.NNc.tstDhlpVotltpGDlV+VYpsuptG.t.hsstsVtt

Secondary structure         EEE            EEEEEE

BAS1577_2(128-220)      NKTAVNVKIPQLGIEAG-KVYVTVTKPNKDESKRV  93
BA1701_2(126-218)       NKTAVNVKIPQLGIEAG-KVYVTVTKPNKDESKRV  93
RBTH_03882_10(898-990)  NKTAVNVVIPQLGEVAG-KIYVSVTKVNKDESKRV  93
BAS1577_1(33-127)       NATDATITGKDLLAVAGGTVYVSVQSKDQLESPRT  95
BA1701_1(31-125)        NATDATITGKDLLAVAGGTVYVSVQSKDQLESPRT  95
RBTH_03882_2(610-705)   GKTEATILGKDLLKVTGGTVYVSVQSENELESART  96
RBTH_03882_3(802-897)   GKTEATILGKDLLKVTGGTVYVSVQSENELESART  96
RBTH_03882_1(418-513)   GKTEATILGKDLLKVTGGTVYVSVQSENELESART  96
RBTH_03882_8(226-320)   GKTEATILGKDLLKVTGGTVYVSVQSENELESART  96
RBTH_03882_6(33-128)    GKVEVTITKKDLLKATGGTVYVSVQSESELESTRT  96
RBTH_03882_4(321-416)   DAKEATITGKDLLVSTGGTVYVTVTKPNKDESKRV  95
RBTH_03882_5(706-800)   DAKEATITGKDLLVSTGGTVYVTVTKPNKDESKRV  95
RBTH_03882_7(514-608)   EATEVKIEKTDLLVSTGGTVYVTVTKPNKDESKRV  95
RBTH_03882_9(129-224)   GKKEATITGKDLLVSTGGTVYVTVTKPNKDESKRV  96
DSY3134_1(51-142)       GQTSVTIEIDQLSEVYG-EIYVTVTRSGYEESDRV  92
DSY3134_2(150-240)      NGSRVTFYG---YAQNAYKDFILLPSESVAKKTIE  91

consensus/80%           stspsslhh.pLhhssG.pVYVoVpp.sp.EStRs
```

Figure 10: BA1701 is homologous to proteins GBAA1701 from *Bacillus anthracis* str. "*Ames* Ancestor," and Bant_01002313 from *Bacillus anthracis* str. A2012.

occurs twice in every protein of the bacillus species as shown in Table 1(g). We refer to this as the NTGFIG domain based on the conserved sequence motif that is present at the N-terminal part. The pairwise sequence identities between sequences corresponding to this domain varied between 31–99%. The secondary structure corresponding to this domain is predicted to comprise three α-helices and two β-strands as shown in Figure 8. The representative domain architecture corresponding to proteins comprising this domain is shown in Figure 21.

3.8. 36-amino-acid-residue NxGK repeat

The 193-amino-acid-residue protein corresponding to GENE_ID BA3686 and described as hypothetical cytosolic protein comprises a 36-amino-acid-residue region as two

```
Secondary structure            HHHHHHHHHHHHHHHH   HH                           HHHHHH

RBTH_06405_4(259-331)         REKALDALQWTIEEKEKLTDNQLLQQYTMQWLKNHRLWTPVVRYWNGSPYAMINDLYPNKY
pBMB165_3(175-247)            REKALEALQWTIEEKEKLIDNQLLQQYTMKWLKRHRLWTPVVRYWNGSPYAMINDLYPNKY
pE33L466_0092_4(259-328)      KRKALEALRWTIEEKEKLDEKQLLKVFNQKWLIKQKLWTPLKRYWKGSPYEMLIALYPNRF
RBTH_06405_3(109-183)         KEKALQLLKWLIEEEEKLPPQKLLQIYGQKWLIEHRLSAPLRVIWNGSPYAMINDLYPNRF
pBMB165_2(25-99)              KEKALQLLKWIIEEEEKVSPQKLLQIYGQKWLNERRLSAPLRVIWDGSPYAMINDLYPNRF
BA3147_2(109-183)             KEKALEALKWTVEEKEKLSKVELLKFYSKKWLEKNKLSAPLVMYWNGSPYAMINSLYPNKF
Bant_01003795_1(25-99)        KEKALEALKWTVEEKEKLSKVELLKFYSKKWLEKNKLSAPLVMYWNGSPYAMINSLYPNKF
BAS2924_2(116-190)            KEKALEALKWTVEEKEKLSKVELLKFYSKKWLEKNKLSAPLVMYWNGSPYAMINSLYPNKF
BAS2924_3(191-265)            KEKALEALKWTVEEKEKLSKVELLKFYSKKWLEKNKLSAPLVMYWNGSPYAMINSLYPNKF
pE33L466_0092_2(109-183)      KEKALTILKWIIEEKEGLSQEKLLELYGKKWLEKNKLGAPLAMYWNSSPYAMINDLYPRRF
RBTH_06405_2(184-258)         KDKTLQALKWTIEKKEKLNVDQLKNIYDNKWLVQSGLSGACQLYWNDSPYAMINDLYPGQF
pBMB165_1(100-174)            KEKALQALKWTIEEKEKLNPDQLKNIYENKWLTQLGLRGACQLYWNDSPYAMINDLYPNQF
BAS2924_4(266-340)            KEKALVALRWTIEEKEKLTSFQLLQVYSVKWLTIHNLISPCQIFWNNSPYSMINELYPGQN
Bant_01003795_2(100-174)      KEKALVALRWTIEEKEKLTSFQLLQVYSVKWLTIHNLISPCQIFWNNSPYSMINELYPGQN
BA3147_3(184-258)             KEKALVALRWTIEEKEKLTSFQLLQVYSVKWLTIHNLISPCQIFWNNSPYSMINELYPGQN
pE33L466_0092_3(184-258)      KEKALEALKWTIEEKEGLTPKQLLDVYNIKWLQTHRLASACQIIWGNSPFRMINDLYIDRF
BA3147_1(34-108)              RELSKRVTKYLIETILKWNEEDIKQKWNTPLIIKYRLLGALKHGYDNSPYKMIEDLYPNRF
BAS2924_1(41-115)             RELSKRVTKYLIETILKWNEEDIKQKWNTPLIIKYRLLGALKHGYDNSPYKMIEDLYPNRF
RBTH_06405_1(34-108)          NQLARRVTKYLVTKILNWNEEEIKQNWNNKLIAKYRLRGVLKHKYNNSPYAMINDLYPNQF
pE33L466_0092_1(34-108)       NKMARRVLTYLLNSILKWNKEDIRKKWNTKLLVKYRLRGLLKHRYENSPFKAINDLYPNQF

consensus/80%                 +EKALpsL+WhlEccEKls..pLhphas.KWL.p.pL.ush.hhWssSPYtMINsLYPspa
```

```
Secondary structure

RBTH_06405_4(259-331)         IKSSFSGYINKF--  73
pBMB165_3(175-247)            LKSSFRGYINKS--  73
pE33L466_0092_4(259-328)      SKNMLKGYM-----  70
RBTH_06405_3(109-183)         KEWEFNKAPNKFWT  75
pBMB165_2(25-99)              KEWEFTKAPNKFWT  75
BA3147_2(109-183)             KEWEFSMTPNNFWT  75
Bant_01003795_1(25-99)        KEWEFSMTPNNFWT  75
BAS2924_2(116-190)            KEWEFSMTPNNFWT  75
BAS2924_3(191-265)            KEWEFSMTPNNFWT  75
pE33L466_0092_2(109-183)      KEWEFGMTPNNFWT  75
RBTH_06405_2(184-258)         KEWEFKMTPNGFWT  75
pBMB165_1(100-174)            KEWEFKMTPSGFWT  75
BAS2924_4(266-340)            KEWEYKFTPTGFWT  75
Bant_01003795_2(100-174)      KEWEYKFTPTGFWT  75
BA3147_3(184-258)             KEWEYKFTPTGFWT  75
pE33L466_0092_3(184-258)      KEWEFRVTPVGYWS  75
BA3147_1(34-108)              KEWEFGMAPLNFWT  75
BAS2924_1(41-115)             KEWEFGMAPLNFWT  75
RBTH_06405_1(34-108)          KEWEFRMTPLNFWT  75
pE33L466_0092_1(34-108)       KEWEFGMTPLNFWT  75

consensus/80%                 KEWEFphsP.tFWT
```

Figure 11: BA3147 is homologous to protein GBAA3147 from *Bacillus anthracis* str. "*Ames* Ancestor."

```
Secondary structure                      EEE                  HHHHHH              EEE

BAS2851_1(20-78)                LEYQQSRFYVTRIPKDFLSIARKRFSIPTDDQIIAFLSCNLFG---SGKYGVYFTSSGLYWK  59
BA3065_1(13-71)                 LEYQQSRFYVTRIPKDFLSIARKRFSIPTDDQIIAFLSCNLFG---SGKYGVYFTSSGLYWK  59
Bant_01003715_1(16-74)          LEYQQSRFYVTRIPKDFLSIARKRFSIPTDDQIIAFLSCNLFG---SGKYGVYFTSSGLYWK  59
BcerKBAB4DRAFT_1832_1(14-72)    LEYQQSRFYVTRIPKDFLSVAKKRFSIPIDDRIFAFLSCNLFG---SGKYGVYFTSSGLYWK  59
RBTH_02124_1(13-71)             LEFQQSRFYVTRIPKDFLSIAQKRFSIPTEDQIVAFLSCNLLG---SGKYGVYFTSSGLYWK  59
Bant_01003715_2(164-225)        LEPDNGLFVETHISDKKLKAIEVRFIIPIEEQIIAFLDTSVLGNMGKGSDGVLICQSGIYFR  62
BAS2851_2(168-229)              LEPDNGLFVETHISDKKLKAIEVRFIIPIEEQIIAFLDTSVLGNMGKGSDGVLICQSGIYFR  62
BA3065_2(161-222)               LEPDNGLFVETHISDKKLKAIEVRFIIPIEEQIIAFLDTSVLGNMGKGSDGVLICQSGIYFR  62
BcerKBAB4DRAFT_1832_2(162-223)  LEPDNGLFVDTHISHKKLKEIGAKYIIPKEEKIIAFLDTSVLGNLGKGSDGVLICEPGIYFR  62

consensus/80%                   LE.ppuhFh.T+IscchLphhphRF.IPh--pIlAFLssslhG...pGp.GVhhspSGlYa+
```

Figure 12: BA3065 is homologous to protein GBAA3065 from *Bacillus anthracis* str. "*Ames* Ancestor."

```
Secondary structure EEEE      EEE                 EEEEE

BA0482(4-56)        IEIHTQGGLKHKVQTEVYNAEALNTKLNDNDLITVLIGDFIIQRIDVKRIIPL  53
BA0482(67-119)      VEVHTNAGKVIEITTNDYDPIYLNEQLNNNNTITVVIGDYIFSRIDVKQVVPV  53

consensus/80%       lElHTpuGhhhclpTpsYss.hLNppLNsNshITVlIGDaIhpRIDVKpllPl
```

Figure 13: BA0482 is homologous to proteins GBAA0482 from *Bacillus anthracis* str. "*Ames* Ancestor," BAS0458 from *Bacillus anthracis* str. Sterne, and Bant_01001108 from *Bacillus anthracis* str. A2012.

```
Secondary structure                   HHHH
BA4081(10-50)       SIGMYLSELQKGTESSRLLAESMAKEIDGKMKIDLGPAGQF   41
BA4081(172-212)     NIQTLINGMQIGALSLPQVAQTMGLDIKSNVQVDLGEAGQF   41

consensus/80%       sIthhlsthQhGs.S...lApoMuh-IcuphplDLG.AGQF
```

(a)

```
Secondary structure            EEE          HHH
BA4081(292-333)     GSKSGSELGQGIISQDGYIKGSALQVVGSAHNAFSTINGSPA   42
BA4081(334-375)     GNQGGQGFGSGIVNQKGYIRGSALEAVTPAHTGFNTINGTPQ   42

consensus/80%       GspuGpthGpGIlsQcGYI+GSALpsVssAHsuFsTINGoPt
```

(b)

FIGURE 14: BA4081 is homologous to proteins GBAA4081 from *Bacillus anthracis* str. "*Ames* Ancestor," BAS3792 from *Bacillus anthracis* str. Sterne, and Bant_01004731 from *Bacillus anthracis* str. A2012.

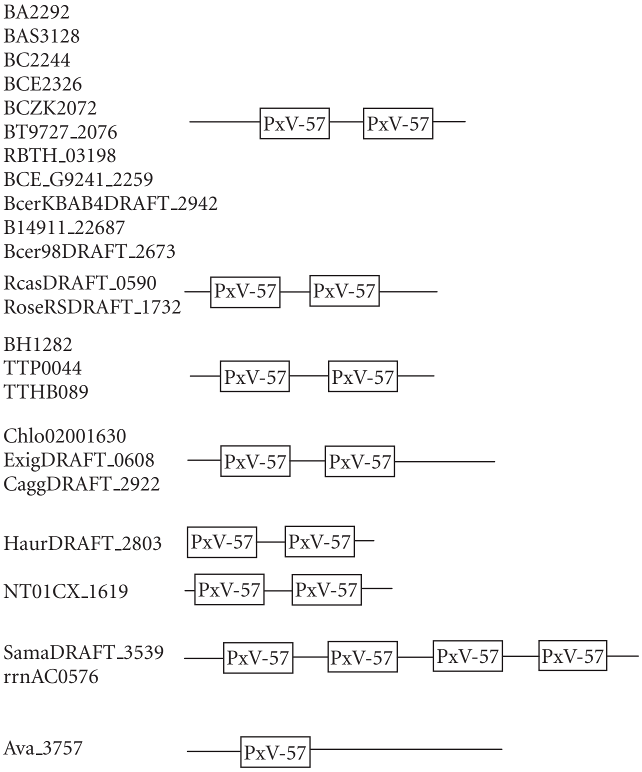

FIGURE 15: PxV-57 aa domain.

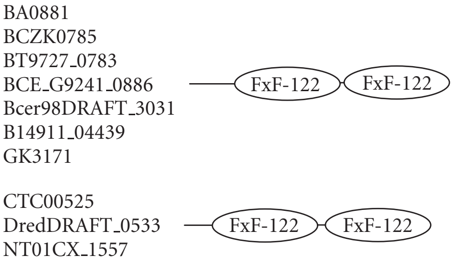

FIGURE 16: FxF-122 aa domain.

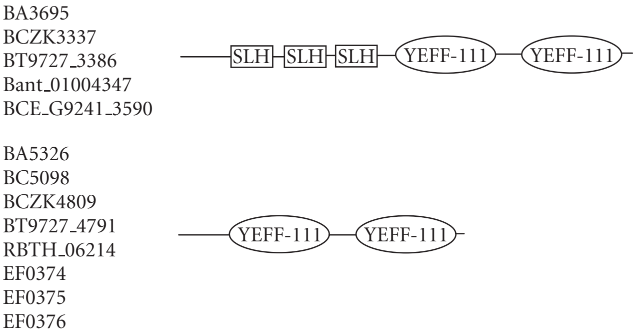

FIGURE 17: YEFF-111 aa domain.

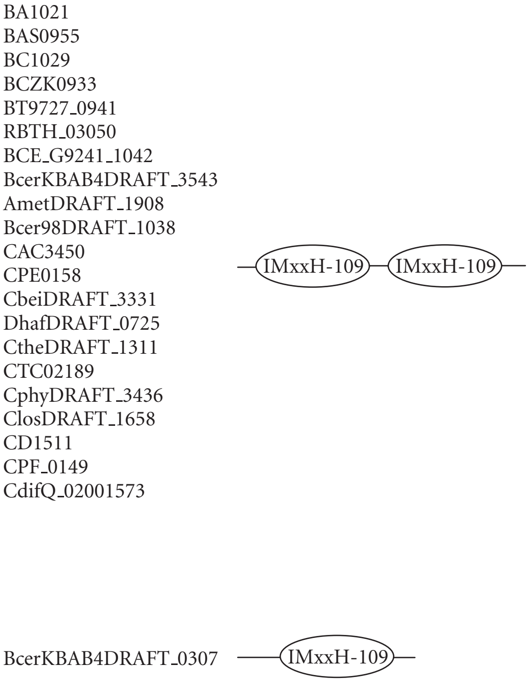

FIGURE 18: IMxxH-109 aa domain.

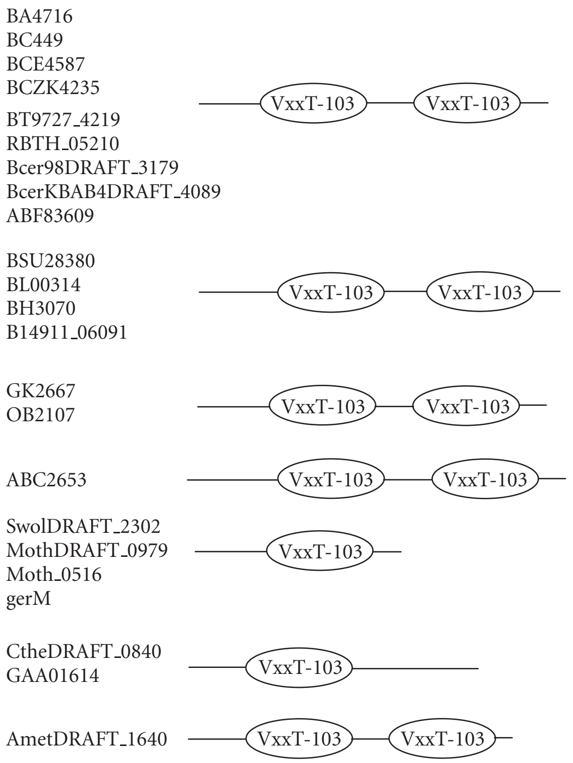

Figure 19: VxxT-103 aa domain.

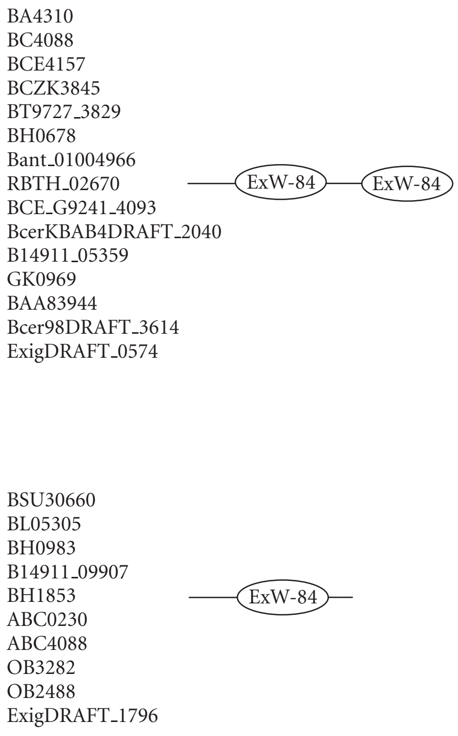

Figure 20: ExW-84 aa domain.

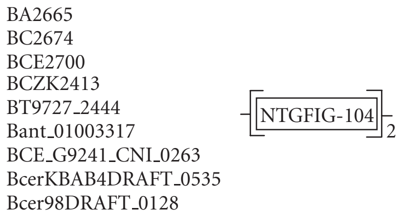

Figure 21: NTGFIG-104 aa domain.

copies. Further BLAST searches using sequence corresponding to the region (94–129) as query identified 9 hypothetical proteins comprising this repeat region from the organisms *B. anthracis*, *B. thuringiensis*, *B. thuringiensis serovar israelensis, Bacillus weihenstephanensis*, and *B. cereus* (see Table 1(h)). The length of proteins varied between 189 to 193-amino-acid residues, and also consists a SAP domain at the N-terminus, in addition to the novel repeat described here. A SAP domain consists of two α-helices and is a DNA-binding motif that is involved in chromosomal organization [32]. Therefore, we believe that these repeats might also participate in a similar function. The multiple sequence alignment corresponding to this repeat identified NxGK sequence motif (Figure 9). The pairwise sequence identities between sequences corresponding to NxGK repeats varied between 36–97%. The secondary structure is predicted to comprise a α-helix and the conserved sequence motif described above is also associated with α-helix. The representative domain architecture corresponding to proteins comprising the NxGK repeats is shown in Figure 22.

3.9. 95-amino-acid-residue VYV domain

The 225-amino-acid-residue protein corresponding to the GENE_ID BA1701 and described as a hypothetical protein comprises a 95-amino-acid-residue region, as two copies in tandem. Further BLAST searches using sequence corresponding to the region (31–125) as query identified BAS1577 protein of *B. anthracis*, RBTH_03882 protein of *Bacillus thuringiensis serovar israelensis*, and DSY3134 of *Desulfitobacterium hafniense* Y51 that are described as hypothetical proteins. The length of proteins varied between 227 to 1674-amino-acid residues (see Table 1(i)). In RBTH_03882, this region occurs ten times and in tandem. The multiple sequence alignment corresponding to this domain identified characteristic sequence motifs; GDxV, VYV (see Figure 10). For the sake of simplicity, we refer to this 95-amino-acid region as VYV domain. The pairwise sequence identities between sequences corresponding to VYV domains varied between 29–95%. The secondary structure corresponding to VYV domain is predicted to comprise five β-strands. The representative domain architecture corresponding to proteins comprising the VYV domains is shown in Figure 23.

3.10. 75-amino-acid-residue KEWE domain

The 262-amino-acid-residue protein corresponding to the GENE_ID BA3147 and described as a hypothetical protein comprises a 75-amino-acid-residue region as three copies in

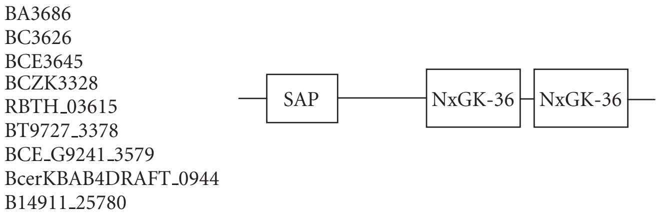

Figure 22: NxGK-36 aa repeat.

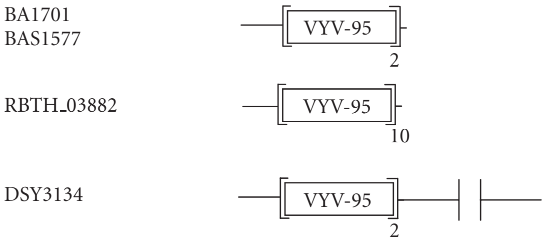

Figure 23: VYV-95 aa domain.

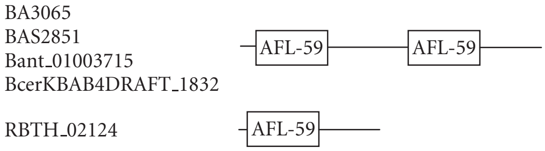

Figure 25: AFL-59 aa domain.

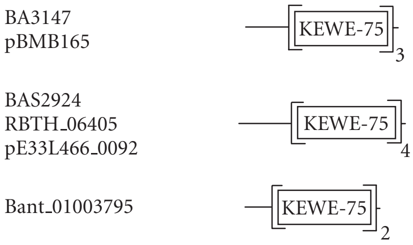

Figure 24: KEWE-75 aa domain.

tandem. Further BLAST searches using the sequence corresponding to the region (34–108) as query identified this domain in 6 proteins that are described as hypothetical proteins (see Table 1(j)). This domain may exist as 2, 3, or 4 copies in these proteins. The length of proteins identified varied between 178 to 344-amino-acid residues. The pairwise sequence identities between sequences corresponding to these regions varied between 22–69%. These domains are present in tandem and associated with SPY, MIN, LYP, KEWE, and FWT conserved sequence motifs as indicated in the multiple sequence alignment (see Figure 11). We refer to these as the KEWE domain, and this sequence motif occurs at the C-terminus of the domain. The secondary structure corresponding to KEWE domain is predicted to comprise three α-helices as shown in Figure 11. The representative domain architecture corresponding to proteins comprising the KEWE domain is shown in Figure 24.

3.11. 59-amino-acid-residue AFL domain

The 290-amino-acid-residue protein corresponding to the GENE_ID BA3065 and described as hypothetical protein comprises a 59-amino-acid-residue region as two copies. Further BLAST searches using sequence corresponding to the region (13–71) as query identified that this region occurs twice in the proteins with GENE_ID's: BAS2851 and Bant_01003715 of *B. anthracis* strains, the protein with GENE_ID: BcerKBAB4DRAFT_1832 of *Bacillus weihenstephanensis,* and once in the protein with GENE_ID: RBTH_02124 of *Bacillus thuringiensis serovar israelensis* (see Table 1(k)). The lengths of the proteins varied between 145 to 297-amino-acid residues and are described as hypothetical proteins. The multiple sequence alignment corresponding to this domain identified two characteristic sequence motifs: RFxI and AFL (see Figure 12). We refer to this as the AFL domain. The sequence identities shared between AFL domains varied between 38–91%. The secondary structure corresponding to the AFL domain is predicted to comprise of one α-helix and two β-strands and the conserved sequence motif AFL is a part of the α-helix. The representative domain architecture corresponding to protein comprising the AFL domain is shown in Figure 25.

3.12. 53-amino-acid-residue RIDVK repeat

The 159-amino-acid-residue protein corresponding to the GENE_ID BA0482 and described as a conserved domain protein comprises a 53-amino-acid region as two copies. BLAST did not identify this repeat in any other proteins; therefore this repeat is unique to *B. anthracis* str. *Ames*. The multiple sequence alignment corresponding to this repeat identified three characteristic sequence motifs: ITV, IGD, and RIDVK (Figure 13). We refer to this as the RIDVK repeat. The sequence identity shared between this RIDVK repeats in BA0482 is 45%. The secondary structure corresponding to the RIDVK repeat is predicted to comprise three β-strands. The representative domain architecture corresponding to protein comprising the RIDVK repeat is shown in Figure 26.

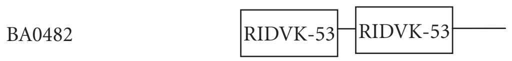

Figure 26: RIDVK-53 aa repeat.

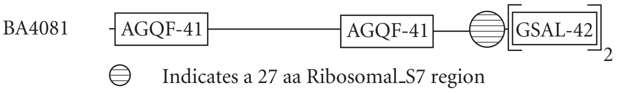

Figure 27: AGQF-41 aa repeat; GSAL-42 aa repeat.

3.13. (a) 41-amino-acid-residue AGQF repeat and (b) 42-amino-acid-residue GSAL repeat

The protein corresponding to the GENE_ID BA4081 comprises 462-amino-acid residues and described as conserved domain protein contains two novel repeat types. The sequence length corresponding to repeat types are 41 and 42-amino-acid residues and are present as two copies in BA4081. BLAST searches identified these repeats to be specific to this protein alone.

(a) The sequence alignment corresponding to 41-amino-acid-residue repeat identified two characteristic sequence motifs: DLG and AGQF (Figure 14(a)). We refer to this as the AGQF repeat. The motif occurs at the C-terminal part of the repeat region. The sequence homology shared between this AGQF repeats is about 34%. The secondary structure corresponding to the AGQF repeat is predicted to comprise one α-helix. The representative domain architecture corresponding to protein comprising the AGQF repeat is shown in Figure 27.

(b) The sequence alignment corresponding to the 42-amino-acid-residue tandem repeat identified three characteristic sequence motifs: GYI, GSAL, and TING (Figure 14(b)) and is a glycine-rich repeat. We refer to this as the GSAL repeat. The sequence homology shared between this GSAL repeats is 52%. The secondary structure corresponding to the GSAL repeat is predicted to comprise one α-helix and one β-strand. The representative domain architecture corresponding to protein comprising the GSAL repeat is shown in Figure 27. This protein is associated with a 27-amino-acid-residue Ribosomal_S7 region that is sandwiched between the 41-amino-acid-residue AGQF repeat and the 42-amino-acid-residue GSAL repeat. These two repeats are specific to this protein alone and are therefore *B. anthracis* str. *Ames* specific.

From the analysis of the *B. anthracis* proteome, we observed that the novel repeats and domains are present in all the strains, such as *Ames*, *Ames* ancestor, Sterne, and A2012, that have been sequenced so far. This indicates that these strains of *B. anthracis* have diverged recently. We also observed that the domains PxV, FxF, YEFF, VxxT, ExW, and VYV are present in proteins from several bacterial organisms. The domains NTGFIG, KEWE, AFL, and the repeats NxGK are specific to bacillus. It is interesting to note that the domains VYV and AFL are present in all the *B. anthracis* species while absent in *B. cereus* genomes. The repeats RIDVK, AGQF, and GSAL are also specifically present only in all the strains of *B. anthracis*. This analysis explains some differences in the closely related *B. anthracis* and *B. cereus* genomes. The identification of these novel domains and repeats in subsequently sequenced genomes will add value to their annotation.

4. CONCLUSIONS

A systematic analysis using computational tools identified four novel repeats and ten domains corresponding to the *B. anthracis* str. *Ames* proteome. Further database searches identified that some novel repeats and domains are also present in other bacterial genomes. The NxGK repeats are associated with SAP domain. The SAP domain is a DNA-binding motif that is involved in chromosomal organization. Therefore, we believe that these repeats also participate in similar function. The YEFF domain-containing proteins are associated with RGD motif and may be involved in cell adhesion. The identification of novel repeats and domains corresponding to *B. anthracis* proteome may be useful for annotation. From the presence of VYV and AFL domains in all the *B. anthracis* species and their absence in *B. cereus* genomes, we identified some differences in these two genomes that are otherwise closely related.

ACKNOWLEDGMENTS

The first author thanks UGC, New Delhi, for a JRF fellowship. The second author thanks DBT, New Delhi, for project funding. The third author thanks DBT, New Delhi, for research funding.

REFERENCES

[1] R. Okinaka, K. Cloud, O. Hampton, et al., “Sequence, assembly and analysis of pX01 and pX02,” *Journal of Applied Microbiology*, vol. 87, no. 2, pp. 261–262, 1999.

[2] T. C. Dixon, M. Meselson, J. Guillemin, and P. C. Hanna, “Anthrax,” *New England Journal of Medicine*, vol. 341, no. 11, pp. 815–826, 1999.

[3] I. Uchida, T. Sekizaki, K. Hashimoto, and N. Terakado, “Association of the encapsulation of Bacillus anthracis with a 60 megadalton plasmid,” *Journal of General Microbiology*, vol. 131, no. 2, pp. 363–367, 1985.

[4] I. Uchida, K. Hashimoto, and N. Terakado, “Virulence and immunogenicity in experimental animals of Bacillus anthracis strains harbouring or lacking 110 MDa and 60 MDa plasmids,” *Journal of General Microbiology*, vol. 132, no. 2, pp. 557–559, 1986.

[5] S. H. Leppla, “Anthrax toxins,” in *Bacterial Toxins and Virulence Factors in Disease*, J. Moss, B. Iglewski, M. Vaughn, and A. T. Tu, Eds., pp. 543–572, Marcel Dekker, New York, NY, USA, 1995.

[6] M. J. Pallen, A. C. Lam, M. Antonio, and K. Dunbar, “An embarrassment of sortases - a richness of substrates?” *Trends in Microbiology*, vol. 9, no. 3, pp. 97–101, 2001.

[7] J. M. Patti, B. L. Allen, M. J. McGavin, and M. Hook, “MSCRAMM-medicated adherence of microorganisms to host tissues,” *Annual Review of Microbiology*, vol. 48, pp. 585–617, 1994.

[8] W. W. Navarre and O. Schneewind, "Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope," *Microbiology and Molecular Biology Reviews*, vol. 63, no. 1, pp. 174–229, 1999.

[9] D. M. Guttmann and D. J. Ellar, "Phenotypic and genotypic comparisons of 23 strains from the Bacillus cereus complex for a selection of known and putative *B. thuringiensis* virulence factors," *FEMS Microbiology Letters*, vol. 188, no. 1, pp. 7–13, 2000.

[10] A. Lupas, H. Engelhardt, J. Peters, U. Santarius, S. Volker, and W. Baumeister, "Domain structure of the *Acetogenium kivui* surface layer revealed by electron crystallography and sequence analysis," *Journal of Bacteriology*, vol. 176, no. 5, pp. 1224–1233, 1994.

[11] A. Lupas, "A circular permutation event in the evolution of the SLH domain?" *Molecular Microbiology*, vol. 20, no. 4, pp. 897–898, 1996.

[12] T. D. Read, S. N. Peterson, N. Tourasse, et al., "The genome sequence of *Bacillus anthracis Ames* and comparison to closely related bacteria," *Nature*, vol. 423, no. 6935, pp. 81–86, 2003.

[13] I. Letunic, R. R. Copley, S. Schmidt, et al., "SMART 4.0: towards genomic data integration," *Nucleic Acids Research*, vol. 32, pp. D142–D144, 2004.

[14] N. J. Mulder, R. Apweiler, T. K. Attwood, et al., "The InterPro database, 2003 brings increased coverage and new features," *Nucleic Acids Research*, vol. 31, no. 1, pp. 315–318, 2003.

[15] A. Bateman, E. Birney, L. Cerruti, et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 30, no. 1, pp. 276–280, 2002.

[16] X. Yu, C. C. S. Chini, M. He, G. Mer, and J. Chen, "The BRCT domain is a phospho-protein binding domain," *Science*, vol. 302, no. 5645, pp. 639–642, 2003.

[17] B. Kobe and J. Deisenhofer, "The leucine-rich repeat: a versatile binding motif," *Trends in Biochemical Sciences*, vol. 19, no. 10, pp. 415–421, 1994.

[18] M. A. Andrade, C. Perez-Iratxeta, and C. P. Ponting, "Protein repeats: structures, functions, and evolution," *Journal of Structural Biology*, vol. 134, no. 2-3, pp. 117–131, 2001.

[19] M. A. Andrade, F. D. Ciccarelli, C. Perez-Iratxeta, and P. Bork, "NEAT: a domain duplicated in genes near the components of a putative Fe^{3+} siderophore transporter from Gram-positive pathogenic bacteria," *Genome Biology*, vol. 3, no. 9, pp. 0047.1–0047.5, 2002.

[20] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[21] A. Heger and L. Holm, "Rapid automatic detection and alignment of repeats in protein sequences," *Proteins: Structure, Function and Genetics*, vol. 41, no. 2, pp. 224–237, 2000.

[22] M. A. Andrade, C. P. Ponting, T. J. Gibson, and P. Bork, "Homology-based method for identification of protein repeats using statistical significance estimates," *Journal of Molecular Biology*, vol. 298, no. 3, pp. 521–537, 2000.

[23] R. Mott, "Accurate formula for P-values of gapped local sequence and profile alignments," *Journal of Molecular Biology*, vol. 300, no. 3, pp. 649–659, 2000.

[24] R. Szklarczyk and J. Heringa, "Tracking repeats using significance and transitivity," *Bioinformatics*, vol. 20, supplement 1, pp. i311–i317, 2004.

[25] S. F. Altschul, T. L. Madden, A. A. Schäffer, et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[26] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[27] B. Rost, C. Sander, and R. Schneider, "PHD—an automatic mail server for protein secondary structure prediction," *Computer Applications in the Biosciences*, vol. 10, no. 1, pp. 53–60, 1994.

[28] L. Falquet, M. Pagni, P. Bucher, et al., "The PROSITE database, its status in 2002," *Nucleic Acids Research*, vol. 30, no. 1, pp. 235–238, 2002.

[29] C. P. Ponting, R. Mott, P. Bork, and R. R. Copley, "Novel protein domains and repeats in *Drosophila melanogaster*: insights into structure, function, and evolution," *Genome Research*, vol. 11, no. 12, pp. 1996–2008, 2001.

[30] S. M. Akula, N. P. Pramod, F.-Z. Wang, and B. Chandran, "Integrin $\alpha3\beta1$ (CD 49c/29) is a cellular receptor for Kaposi's sarcoma-associated herpesvirus (KSHV/HHV-8) entry into the target cells," *Cell*, vol. 108, no. 3, pp. 407–419, 2002.

[31] S. E. D'Souza, M. H. Ginsberg, and E. F. Plow, "Arginyl-glycyl-aspartic acid (RGD): a cell adhesion motif," *Trends in Biochemical Sciences*, vol. 16, no. 7, pp. 246–250, 1991.

[32] L. Aravind and E. V. Koonin, "SAP—a putative DNA-binding motif involved in chromosomal organization," *Trends in Biochemical Sciences*, vol. 25, no. 3, pp. 112–114, 2000.