The proteome of spore surface layers in food spoiling bacteria
Abhyankar, W.R.

Citation for published version (APA):
Abhyankar, W. R. (2014). The proteome of spore surface layers in food spoiling bacteria

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Abstract Analysis of bacterial spore surface proteomes has led to identification of > 90 different proteins however functional characterization has not been achieved for most of these proteins. In addition to their discovery, a few proteins have been utilized for biotechnological applications. Spore longevity and their escape from immune surveillance still remain challenges for the food security and medical fields. This chapter reasons out the peculiarities of spore surface proteins that might be decisive in these respects. Emerging bioinformatic tools have made protein structure and function predictions possible. In this context, using such tools the study of amino acid distributions of proteins highlighted some characteristics about structures of these proteins whereas molecular mass, pI and GRAVY distributions focussed on the potential role of these proteins in spore surface adhesion. Finally, the potential of certain peptide sequences, from the identified proteins, in drug development against resistant pathogens was examined.
Introduction

Spores are of major concern when it comes to food safety and pathogenic infections where they act as mediators. Thus their persistence and the germination mechanisms are considerably important for human infections. As discussed extensively in Chapter 1, spore structure plays a crucial role in their survival. In spite of the many efforts to unravel the details behind spore integrity, longevity and survival, especially with regards to the proteins from the spore outer layers, the functional and structural studies of these proteins have been scarce. X-ray crystallographic structures (with resolutions < 3Å and R-values < 20%) of only 12 proteins are available in the Protein Data Bank (http://www.rcsb.org/pdb/home/home.do). From *B. subtilis*, structures of proteins CotA [1], CotI (YtaA [2]), GerBC [3], GerE [4], PdaA [5], SpoVT [6], SpsA [7] and YkuD [8] are available in the PDB data bank while from *B. cereus* and *B. anthracis* structures of proteins SleB (catalytic domain [9, 10]), Alr (BA_0252 [11]), BclA [12] are accessible. Recently the structure of CspB protease involved in *Clostridium* spore germination was also published [13]. Evidently, the knowledge of the molecular mechanisms indispensable for the assembly of the bacterial spore coat is incomplete and more detailed studies are needed to understand the means through which proteins are targeted to the basic coat layer. Along with the fundamental knowledge of the spore proteins and the spore structure, the resistant as well as the adhesive properties of spores also need to be focused upon in order to culminate spores. In an attempt to uncover the structural details of these proteins, their amino acid compositions and other discernible molecular properties that were marked in our studies are discussed below. Furthermore, the potential therapeutic applications of the spores in view of their component coat or exosporium proteins are also discussed.

1. Fundamentals of spore surface proteins
   
   (a) Amino acid compositions

   Since 1950s the relation between the protein amino acid sequence and the protein structure is well known. Subsequently, the propensities of different amino acids in formation of the secondary protein structures namely, α helices, β sheets, turns were studied and it was found that, for example, amino acids such as alanine, glutamate, leucine tend to be present in α helices, amino acids such as valine, isoleucine tend to prefer β sheet structures whereas amino acids glycine, asparagine and proline have preference towards turns. More recently it was also shown that the amino acid sequence also determines the intrinsic protein disorder (IPD) that renders a particular region of the protein unstructured thereby enlightening the functional variety in the proteins. These studies concluded that presence of low sequence complexity and amino acid compositional bias, with low content of hydrophobic, bulky amino acids (Val, Leu, Ile, Met, Phe, Trp and Tyr), and a high proportion of charged amino acids (Gln, Ser, Pro, Glu, Lys, Gly and Ala) were indicative of a probable intrinsic disorder [14]. It has also been found that IPDs generally are: resistant to boiling temperature; insensitive to denaturing chemicals; and susceptible to proteolytic cleavage [15]. Due to the intrinsic disorders
these proteins are also difficult to crystallize in X-ray crystallographic studies. Thus protein sequence comparison has become a powerful tool to characterize the protein sequences and possibly their functional analysis. The amino acid sequence is therefore the most important feature of a protein for its localization, structure and function. In these respects, the spore coat or exosporium proteins have peculiar amino acid sequences such as the collagen-like repeat sequences, histidine-rich or serine-rich C-termini of proteins, Proline-Glycine-Tyrosine rich protein sequences etc.

The Uniprot protein database (http://www.uniprot.org/) illustrates the relative contribution of each of the 20 natural amino acids to the entire protein compliment, across all the prokaryotic and eukaryotic organisms studied till date. In comparison with this relative contribution of amino acids, it was observed that the proteomes of individual species - *B. subtilis* 168, *B. cereus* ATCC 14579 and *C. difficile* 630 follow a similar trend and barring certain amino acids such as glutamine, threonine etc. the general amino acid spectrum of spore outer layer proteins strongly resembles that of the vegetative cell proteins. In general, the spore surface proteins appear to be rich in aliphatic hydrophobic and acidic amino acids suggesting a possibility that spore surfaces are hydrophobic and the charged residues could be involved in cross-linking amongst different proteins. Figure 1 shows the relative amino acid composition of the spore surface proteomes from *B. subtilis* 168 [16, 17], *B. cereus* ATCC 14579 [18] and *C. difficile* 630 [18]. Clearly, the anaerobic *C. difficile* vegetative cell proteome appears to diverge from the aerobic Bacilli in case of isoleucine, asparagine, glutamine and lysine distributions. Interestingly, though these proteins have higher lysine contents the arginine contents are considerably low. Higher amounts of lysine, glutamine, and tyrosine in *B. subtilis* are in agreement with the predictions of ε-γ-glutamyl-lysine[19] and di-tyrosine crosslinks [20] in the coat proteins. The highest histidine content in *B. subtilis* spores might indicate their role in protein stability[21], protein interactions[22], or in protein function (the histidine residues in CotA are involved in stabilization of the copper moieties thereby establishing catalytic function of the *B. subtilis* laccase [1]) or as a cleavage site for proteases to convert precursor proteins into their mature forms [23]. On the other hand the *B. cereus* ATCC 14579 coat and exosporium proteome appear to be rich in threonine, glycine, proline and asparagine. The Bcl-family proteins identified from the exosporium of *B. cereus* are well represented by the higher glycine, proline (GXX repeats) and threonine (potential glycosylation site) contribution. Earlier work performed by Aronson and Fitz-James [24] indicated that except in *B. licheniformis* the spores from other *Bacillus* spp. are rich in glycine and lysine and our observations for amino acid distributions for the identified proteins illustrate the same trend. The *C. difficile* 630 proteomes for vegetative cells and spore surfaces emerged to be highly similar except for isoleucine and leucine contents. This higher leucine content in the vegetative cells could result from the cell surface proteins that contain leucine rich repeat (LRR) motives. These proteins may play a role in adhesion of cells to surfaces as well as in host pathogen interactions [25]. The spore surface proteins identified from *C. difficile* 630 in our study are extremely acidic but are also rich in cysteine, tyrosine, asparagine and lysine again confirming the previous observations [26]. A combination of charged residues of the surface proteins also suggest a possibility of salt-bridge formation, as in case of viruses [27], amongst proteins thus
Figure 1. Amino acid contributions to proteomes of *B. subtilis* 168, *B. cereus* ATCC 14579 & *C. difficile* 630 as compared to the averaged proteome predicted from Uniprot database. The x-axis shows the one-letter symbols of the amino acids. The continuous lines represent the vegetative cell proteomes and the dotted lines represent the spore surface proteomes.
Molecular properties of spore surface proteins

117

... adding to the stability of the coat structure. Intriguingly, all the amino acid residues mentioned above have high propensities [28] to be present either on the exposed protein surface or in the protein-protein binding site.

Additionally, some amino acid biases were identified in the spore surface proteins in the form of single amino acid repeats or internal tandem repeat sequences (Table 1) that have escaped the focus since past. The repeat sequences, in general, are abundant in proteins and vital for protein structure, function and evolution. Single amino acid repeats are mainly responsible for unstructured regions of proteins. Such repeats have been shown to be the cause of many diseases such as Huntington’s disease in humans [29]. It is hypothesized that such single amino acid repeats can formulate the structure of that particular region of the protein and may also be involved in protein aggregation [30]. The later property is beneficial for rendering spore coat a proteinaceous structure. The tandem repeat sequences on the other hand indicate strong biological relevance wherein the phylogenetically conserved repeats among the orthologous proteins should have a conserved function. These conserved properties are generally responsible for maintaining the essential functions whereas the varying regions may arise due to the effects on environment or due to ecological pressures. Using computational tools previously, tandem repeats have been identified from the archaeabacterial cell surface proteins [31, 32]. The authors identified 6 different tandem repeat sequences from the cell surface proteins of *Methanosarcina acetivorans*. Progressing further it was also shown by the same group that these repeats could be involved in formation of the β-propeller fold structure. Proteins that can form the β-propeller fold are universally found in nature from viruses to bacteria, prokaryotes to eukaryotes, invertebrates to mammals [32]. Also in protozoan parasites proteins with tandem repeats have been found to act as the target of B-lymphocyte response. Thus in another study, efforts were taken to identify the tandem repeat protein sequences from protozoan *Trypanosoma cruzi* [33]. These examples indicate the significance of mining the protein sequences in search of unique sequence regions. We analyzed the proteins identified in our studies for the presence of the repeat sequences and we found that both single amino acid as well as tandem repeat sequences exist in spore surface proteins. We used the recently published algorithm T-REKS [34] for this purpose. Some representative proteins from the three species are shown in Table 1. In addition, the *B. subtilis*, proteins CotT and YmaG are rich in proline, glycine and tyrosine. Their sequences show homology with the PGY-rich protein from rice (*Oryza sativa*). OsPGYRP contains GYPPX-repeat at N-terminus and a cysteine rich region at the C-terminus of the protein [35]. On the contrary, CotT has PYYPRPYYPF and GYGG repeats and YmaG has GRPFGE repeats. The OsPGYRP is expressed in response to cold, salt or osmotic stress but overexpression of this gene was reported to enhance cold stress in *E. coli* [35]. The role of tandem repeats in CotT, YmaG and the other spore surface proteins could provide the knowledge about their role in maintaining the structural integrity of spores.

Apart from the tandem repeat sequences some spore surface proteins also have single amino acid repeats. For instance the inner coat protein YeeK from *B. subtilis* spores contains a region with 11 histidine (H) residues towards its C-terminal end (in the
Table 1. Internal tandem repeat sequences found in spore surface proteins in *B. subtilis* 168, *B. cereus* ATCC 14579 & *C. difficile* 630. $P_{\text{sim}}$ similarity coefficient for the repeats.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
<th>Repeat length</th>
<th>No. of repeats</th>
<th>Residues From</th>
<th>To</th>
<th>$P_{\text{sim}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>YirY (SbcC)</td>
<td>EQAEKVL</td>
<td>7</td>
<td>2</td>
<td>216</td>
<td>229</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>EQAEARL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CotT</td>
<td>PTYPFRPPPFFYT</td>
<td>11</td>
<td>2</td>
<td>21</td>
<td>44</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>PTYPFRPPPFFYTF--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CotU(YnaH)</td>
<td>KYDHDKHKYDCC--</td>
<td>12</td>
<td>3</td>
<td>21</td>
<td>59</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>KYDHDKHKYD--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CotC</td>
<td>DYYKDYD</td>
<td>7</td>
<td>3</td>
<td>33</td>
<td>53</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>DYYKDY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CotB</td>
<td>--SSRSLKSSSVYQS</td>
<td>14</td>
<td>2</td>
<td>317</td>
<td>342</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>SKSSR--SSRSVYQ--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>B. cereus</em> ATCC 14579</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC_1559 (YppG)</td>
<td>NQAQQFFQQQQQPYV</td>
<td>12</td>
<td>3</td>
<td>54</td>
<td>105</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>NQAQQFFQQQQ--PYV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC_1560 (CotD)</td>
<td>SPFPPGFNY</td>
<td>9</td>
<td>3</td>
<td>100</td>
<td>125</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>SPFGPGPNV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC_2030 (CotG)</td>
<td>--VKKWTFVTKCTRV</td>
<td>13</td>
<td>3</td>
<td>60</td>
<td>97</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>RVQKWTFVTKVTR--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC_2874 (CotX)</td>
<td>EPKKRFVF</td>
<td>7</td>
<td>4</td>
<td>21</td>
<td>48</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>EPKKKEDC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC_3582</td>
<td>SCGFGGCGG--GFS</td>
<td>10</td>
<td>3</td>
<td>77</td>
<td>107</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>SC--CCGCGGCG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SC--CCGCGGGGG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC_4420 (SafA)</td>
<td>QEVEVQ--KFP</td>
<td>9</td>
<td>6</td>
<td>109</td>
<td>157</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>QKEMQV--KF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>QKEMQV--KF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>QKEVQV--KF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>QKEVQKOP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PHI-MPI-MDNNQF--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PHI-MPI-MDNNQP--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PHI-MPI-MDNNQF--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PHI-MPI-MDNNQP--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PHI-MPI-MDNNQM--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PHI-MPI-MDNNQM--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PHI-MPI-MDNNQP--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PHI-MPI-MDNNQP--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*C. difficile* 630

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
<th>Repeat length</th>
<th>No. of repeats</th>
<th>Residues From</th>
<th>To</th>
<th>$P_{\text{sim}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD1067 (CdeC)</td>
<td>CNPCCKKFP</td>
<td>8</td>
<td>2</td>
<td>64</td>
<td>79</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>CNFCCKFP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD3664</td>
<td>PEGLVFTH</td>
<td>8</td>
<td>2</td>
<td>313</td>
<td>328</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>PEGGLFTH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
sequence region 110-125) and CotB protein has a stretch of serine residues (potential glycosylation sites) towards its C-terminal. The His-rich region of YeeK may be necessary for the overall protein stability [36] but in addition there are also other repeats in the N-terminal as well as the middle region of the protein and YeeK is speculated to play a role in survival in certain environmental conditions [37]. In *B. cereus* ATCC 14579, proteins BC_3582 and BC_3992, BC_4420 (SafA) contain glycine, proline and methionine repeat regions respectively while proteins BC_2149 and BC_5391 contain glutamine repeats. Also protein CD3522 from *C. difficile* is rich in Glutamic acid (13 residues) in the sequence region 226-238 of the protein. Repeats with complex patterns known as domain repeats are also observed in spore surface proteins. The best examples representing the domain repeats are YXY isodityrosine motifs [38] in proteins CotC and CotU from *B. subtilis* and the GXX repeats from the Bcl-family of exosporium proteins identified from *B. cereus* spores. Isodityrosine, though not yet identified in spores, has been identified from the plant cell wall glycoprotein and is said to play a role in protein cross-linking [39]. All these findings point towards the unconventional proteins sequence characteristics and thereby the structures of spore surface proteins.

(b) Localization signals & Trans-membrane Helices (TMHs)

Amino acid sequences of proteins are also important for the protein sorting assemblies in cells. In prokaryotes and eukaryotes, exported proteins are usually synthesized as precursors with an amino-terminal signal peptide, which is recognized by a cellular sorting and translocation machinery and which guides a protein to its destination. Once the protein is delivered to its destination these short stretches of signal peptides are cleaved off by special signal peptidases (SPases) resulting in the release of mature protein from the protein translocation machinery. The signal peptides generally contain three distinct domains. The N-domain (amino terminal domain) of signal peptides, suggested to interact with the translocation machinery and the phospholipids in the lipid bilayer of the cell membrane, contains at least one arginine or lysine residue. The hydrophobic domain (H-domain), following the N-domain, is a stretch of hydrophobic residues that may form an α-helix in the membrane [40]. Glycine or proline residues that may act as helix-breakers are frequently present in the middle of this hydrophobic core. The following residues (C-domain) might allow the signal peptide to take up a hairpin-like structure in order to insert into the membrane. Signal peptides can be cleaved of by SPase type I or II, the latter being active in pre-lipoproteins that contain the lipobox 4 amino acid consensus sequence [(L/V/I)-(A/S/T/V/I)-(G/A/S)-C] [41]. The SPase-I cleaved pre-proteins can be transported by the regular Sec pathway of protein transport. Some of such SPaseI cleaved proteins may contain a so called twin arginine motif (RR-motif), and thus are transported via the Tat pathway. During sporulation the communication between the mother cell and the forespore requires protein transport. Previously two proteins SpoIIIR and SpoIIIG have been shown to be transported across the forespore to interact with the mother cell membrane [42, 43]. Since the coat and exosporium layers are synthesized inside the mother cell cytoplasm, these layers are assembled with most of the proteins being independent of a signaling peptide for their localization. However certain σ^G and/or
σF regulated proteins from *B. subtilis* 168 require these signals for their incorporation in the spore coat as mentioned in Chapter 2. *B. cereus* ATCC 14579 and *C. difficile* 630 spore surface proteins also have some candidates that have signal sequences, as predicted by SignalP v. 4.1 tool and confirmed with Uniprot database and these are collectively shown in Table 2(A). As seen only a small fraction of proteins appear to depend on their transport by translocation pathways. However, it is also evident from the work of van Ooij and colleagues [44] that certain coat proteins firstly form rings that then encircle the entire spore. Also the work of Stelma (Jr.) and co-workers [45] has shown that coat formation, development of spore body and deposition of coat onto the spores could possibly be independent events. Therefore the details of the pathways of the spore coat assembly and the role of signal-peptides for the assembly of certain proteins (for instance those mentioned in Table 2(A)) into the coat still remain elusive.

**Table 2.** (A) Proteins containing signal sequences. *Proteins with signal peptides containing a Twin-arginine motif (R-R-X-#-#). (B) Proteins containing transmembrane helices (TMHs).**

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Uniprot ID</th>
<th>Residues in signal peptide</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B. subtilis 168</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YbfO</td>
<td>P96619</td>
<td>1-28</td>
</tr>
<tr>
<td>YckD</td>
<td>P42402</td>
<td>1-23</td>
</tr>
<tr>
<td>TeyA (YckK)</td>
<td>P42199</td>
<td>1-19</td>
</tr>
<tr>
<td>YhcN</td>
<td>P54598</td>
<td>1-20</td>
</tr>
<tr>
<td>OppA</td>
<td>P24141</td>
<td>1-20</td>
</tr>
<tr>
<td>SleB</td>
<td>P50739</td>
<td>1-29</td>
</tr>
<tr>
<td>DacF</td>
<td>P38422</td>
<td>1-23</td>
</tr>
<tr>
<td>CoxA</td>
<td>P94446</td>
<td>1-21</td>
</tr>
<tr>
<td><strong>B. cereus ATCC 14579</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC_0212</td>
<td>Q81HY1</td>
<td>1-24</td>
</tr>
<tr>
<td>BC_0263*</td>
<td>Q81IT6</td>
<td>1-22</td>
</tr>
<tr>
<td>BC_1279</td>
<td>Q81GC8</td>
<td>1-29</td>
</tr>
<tr>
<td>BC_1281</td>
<td>Q81GC6</td>
<td>1-16</td>
</tr>
<tr>
<td>BC_1284 (InA)</td>
<td>Q81GC3</td>
<td>1-23</td>
</tr>
<tr>
<td>BC_1334</td>
<td>Q81U34</td>
<td>1-26</td>
</tr>
<tr>
<td>BC_2026</td>
<td>Q81F3</td>
<td>1-28</td>
</tr>
<tr>
<td>BC_2375*</td>
<td>Q81D0</td>
<td>1-24</td>
</tr>
<tr>
<td>BC_2481</td>
<td>Q81D93</td>
<td>1-33</td>
</tr>
<tr>
<td>BC_2753 (SleB)</td>
<td>P0A3V0</td>
<td>1-34</td>
</tr>
<tr>
<td>BC_3586</td>
<td>Q81A2</td>
<td>1-31</td>
</tr>
<tr>
<td>BC_4075 (DacF)</td>
<td>Q81B1</td>
<td>1-23</td>
</tr>
<tr>
<td>BC_4419</td>
<td>Q817V9</td>
<td>1-20</td>
</tr>
<tr>
<td>(YhcN)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC_5056</td>
<td>Q81S6</td>
<td>1-28</td>
</tr>
<tr>
<td><strong>C. difficile 630</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD0855 (OppA)</td>
<td>Q18A51</td>
<td>1-31</td>
</tr>
<tr>
<td>CD1291 (DacF)</td>
<td>Q18BF4</td>
<td>1-23</td>
</tr>
<tr>
<td>CD1622</td>
<td>Q18H8</td>
<td>1-20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Uniprot ID</th>
<th>Protein Length (a.a.)</th>
<th>No. of TMHs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B. subtilis 168</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YdcC</td>
<td>P96619</td>
<td>338</td>
<td>1</td>
</tr>
<tr>
<td>AtcL (YloB)</td>
<td>Q34431</td>
<td>890</td>
<td>10</td>
</tr>
<tr>
<td>YodL</td>
<td>Q34654</td>
<td>83</td>
<td>1</td>
</tr>
<tr>
<td>YpeB</td>
<td>P38490</td>
<td>450</td>
<td>1</td>
</tr>
<tr>
<td>Yqdx</td>
<td>P54481</td>
<td>129</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Uniprot ID</th>
<th>Protein Length (a.a.)</th>
<th>No. of TMHs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B. cereus ATCC 14579</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC_0337</td>
<td>Q81A56</td>
<td>125</td>
<td>2</td>
</tr>
<tr>
<td>BC_1391</td>
<td>Q81G20</td>
<td>124</td>
<td>2</td>
</tr>
<tr>
<td>BC_1456</td>
<td>Q81F5</td>
<td>142</td>
<td>1</td>
</tr>
<tr>
<td>BC_2099</td>
<td>Q81E9</td>
<td>112</td>
<td>2</td>
</tr>
<tr>
<td>BC_2382</td>
<td>Q81D4</td>
<td>401</td>
<td>5</td>
</tr>
<tr>
<td>BC_2481</td>
<td>Q81D93</td>
<td>107</td>
<td>1</td>
</tr>
<tr>
<td>BC_2752</td>
<td>Q81315</td>
<td>446</td>
<td>1</td>
</tr>
<tr>
<td>(YpeB)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC_3582</td>
<td>Q81A6</td>
<td>131</td>
<td>1</td>
</tr>
<tr>
<td>BC_3992</td>
<td>Q81H6</td>
<td>113</td>
<td>1</td>
</tr>
<tr>
<td>BC_4410</td>
<td>Q817W7</td>
<td>86</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Uniprot ID</th>
<th>Protein Length (a.a.)</th>
<th>No. of TMHs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C. difficile 630</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD1063</td>
<td>Q18AR2</td>
<td>68</td>
<td>2</td>
</tr>
<tr>
<td>CD1463</td>
<td>Q18BY4</td>
<td>148</td>
<td>1</td>
</tr>
<tr>
<td>CD2434</td>
<td>Q182C1</td>
<td>200</td>
<td>1</td>
</tr>
<tr>
<td>CD2598</td>
<td>Q182T7</td>
<td>259</td>
<td>1</td>
</tr>
<tr>
<td>CD2808</td>
<td>Q183P2</td>
<td>209</td>
<td>1</td>
</tr>
<tr>
<td>CD3457</td>
<td>Q180V4</td>
<td>138</td>
<td>1</td>
</tr>
</tbody>
</table>
Along with the signalling sequences the transmembrane segments are also important to be considered for identification of possible transmembrane anchored proteins (Table 2 (B)). The transmembrane helical segments in membrane proteins are usually composed of hydrophobic amino acids [46]. In contrast, water soluble proteins are composed of both polar and apolar amino acids. In the trans-membrane helical (TMH) domains, there is a large variety of conserved sequences. Therefore, there are a significant number of transmembrane segments that potentially interact with the lipid bilayer and in many different modes [47]. Though there are very less transmembrane proteins identified from the spore integument region (region of the boundary between outer forespore membrane and the inner spore coat) some proteins like hydrolase, Ca$^{2+}$ transporters and ATPases do contain trans-membrane domains. Protein SpoVM, mostly identified from the soluble coat protein fraction is also identified from the integument region and has been shown to responsible for anchoring the coat protein SpoIVA to the spore outer membrane [48]. Thus the proteins discussed in this table could either be important for germination or speculatively may serve as anchors for subsequent proteins layers.

2. Adhesive properties of spores

Bacterial adhesion to surfaces has been reasoned as a possible virulence factor for many pathogenic microorganisms that are important in the medical, pharmaceutical, and food industries. Many cell surface proteins from bacteria are known to be involved in adhesion processes [49]. Physical properties such as surface energy, texture, and relative charge distribution are the basis of adhesion of proteins to surfaces. Larger proteins are more likely to adhere and remain attached to a surface due to the higher number of contact points between amino acids and the surface. In addition, the role of hydrophobic interactions in bacterial cell adhesion to surfaces has previously been examined by using several methods, based on the precipitation of cells by salts, hydrophobic interaction chromatography (HIC [50]), and adherence to various liquid hydrocarbons including hexadecane (BATH test[51]). Separation using HIC is carried out due to the reversible interaction between a protein and the hydrophobic cells or ligands bound to the chromatography matrix whereas in the BATH test, the hydrophobic cells associate more to the hydrocarbon or the oil layer when cells are mixed into a hydrocarbon/ oil : water system. In both cases the estimate, of cells that are more hydrophobic than the rest of the cell population, can be interpreted in % values. For BATH assays, the % hydrophobicity is defined as the average % decrease in the $A_{440}$ of the aqueous phases after partitioning and for HIC, it is defined as the average % decrease in the $A_{440}$ of the spore suspensions eluded from duplicate Sepharose columns (Figure 2). Each of these methods has some limitations. The two-phase separation technique, id est the BATH test, may alter surface hydrophobicity [52] of B. subtilis and B. thuringiensis spores. For HIC, spore aggregation can be a problem as the clumps may be trapped in the matrix thereby interfering with the hydrophobicity measurements. Compared to the vegetative cells, spores are more adhesive in nature however their surface hydrophobicity has not been studied in good detail. Koshikawa and colleagues established a relationship between spore surface hydrophobicity and the presence of an exosporium layer in their study using the BATH
assay where more than 70% of the spores from species that showed presence of an exosporium were found to associate with the hexadecane layer [53]. Later Wiencek and his group [54] showed that spores possessed significantly higher hydrophobic surfaces compared to the vegetative cells. The authors suggested that amongst the species considered for their study, spore form *B. subtilis* strain A possessed the least surface hydrophobicity when measured by the BATH assay (Figure 2).

The chemical and morphological characteristics of spores from different species may relate to the large variations in the hydrophobicity behaviours of spores. Large differences in the chemical composition of spore coats have been found in spores of *Bacillus* species [55]. We analysed the identified spore surface proteins for their hydrophobicity characteristics (hydrophilicity or hydrophobicity) and molecular masses. The GRAVY index indicates the solubility of the proteins. Therefore the identified protein set from the spore surface layers from three species was subjected to GRAVY (Grand Average of hydropathy) index analysis using the ProtParam tool from the ExPaSy proteomics server (http://web.expasy.org/protparam/). The results are compiled in Figure 3. For the three organisms studied, all the proteins were < 100kDa in size, except BC_2639 (~520kDa). The mean size of identified pore surface layer proteins increases from 29 kDa for *B. subtilis* 168, 33 kDa for *C. difficile* 630 to 36 kDa for *B. cereus* ATCC 14579. Most of the identified proteins were relatively hydrophilic (negative GRAVY indices). The mean GRAVY-index was -0.28 for *B. cereus* ATCC 14579, -0.37 for *C. difficile* 630, and -0.61 for *B. subtilis* 168. However previous studies indicated that spore surfaces are hydrophobic and also the proteins were found to contain high number of hydrophobic amino acid residues (Figure 1) in their structure. It is plausible that the spore surface proteins form a cross-linked matrix in the coat and exosporium and along with the other reported components such as lipids, phosphorous, sugar moieties [56, 57] these layers become hydrophobic. To address the question of spore adhesion further, it is desirable to select a group of coat and exosporium proteins and study their role in spore adhesion e.g. by mutational studies. The small size of proteins suggests importance of quantitative studies in order to get the knowledge about the abundance and distribution of these proteins in the surface layers.

The maximum adhesiveness of spores due to surface hydrophobicity has been proposed to be at pH 3 [58]. Husmark & Rönner found that pH 3 correlated well with the isoelectric point of the spore surface and thus the differences in the number of adhered spores to surfaces was attributed to the differing charges of the spore surface. The differing charge will change the electrostatic attractions between the spores and the external surface as well as the steric stability of the adhesion will be altered. To gain insights into the role of spore surface proteins in the pH dependent adhesive properties we analyzed the proteins for their respective pI values. The hydrophilic proteins were spread over both the acidic and basic pH ranges. The mean pI of the identified *B. cereus* spore coat & exosporium proteome was 6.46, while the mean pI for the identified spore coat proteomes from *C. difficile* 630 and *B. subtilis* 168 were respectively 5.79 & 6.99. In general, the pI distribution of identified proteins was found to be unimodal for *C. difficile* 630, and bimodal for both the *Bacillus* spp. Based on a study performed by Schwartz and colleagues [59] in vegetative bacteria as well as eukaryotic cells, we speculate that this
modality behavior may reflect the localization of proteins in the spore layers. Interestingly in all three organisms, the pI values for most of the outer coat, crust and exosporium proteins were in the range of 3 - 5 while some abundant proteins (CotG, CotB, ExsK, and BcIC etc.) peculiarly possessed pI values in the range of 6 - 11. If these proteins are important for spore adhesion then according to the study performed by Husmark & Rönner, at pH 3 all these proteins will be charged and therefore the adhesiveness of spores at this pH will be decreased due to electrostatic repulsions between the charged spore surface and the charged stainless steel surface.

Spore adhesion is a major problem in dairy industries. In a typical dairy industry the storage tanks, centrifuges, pasteurizers, heat exchangers, packaging machines and many other equipments used are made of principally two varieties of stainless steel - 304 and/or 316. According to a study performed by Gispert and others [60], the surface charge of 316 variety of stainless steel is near 0 at pH 6.0. Therefore in their study the protein bovine serum albumin (BSA) did not adhere to the stainless steel surface as much as it adhered to the alumina surface which is more positively charged. The authors speculated that the Ca\(^{2+}\) and Mg\(^{2+}\) ions present in the solvent, used in this study, may have bridged the negatively charged BSA surface and the stainless steel surface. Ca\(^{2+}\) and Mg\(^{2+}\) ions along with many other proteins are also present in the milk (pH ~7). At this pH, in case of B. subtilis many of the spore coat proteins will carry a negative charge leading to
electrostatic repulsion with the stainless steel surface making the spores less adhesive. However, hydrophobic surfaces might still induce adhesion. In case of *B. cereus* spores, the components such as lipids and phospholipids impart hydrophobicity to the exosporium; sugars from glycoproteins tend to make exosporium sticky and bulky and proteins from the hair-like nap (e.g. BclC in *B. cereus*) will potentially carry a positive charge at pH 7 thus overcoming the electrostatic repulsion forces as predicted by Busscher & Weerkamp [61]. Thus *B. cereus* spores may adhere more strongly to the stainless steel surfaces compared to the *B. subtilis* spores. Therefore it is strongly recommended that the stainless steel surfaces should be replaced or treated to obtain more hydrophilic surfaces thereby minimizing the risk of spore adhesion.

### 3. Therapeutic applications of spores surface proteins

Since past the conventional means for identifying bioactive or antimicrobial peptides have been tedious and time consuming. Also from past few years emergence of antibiotic resistance amongst the bacterial domain has been a major problem and it still continues to threaten the medical field. Thus pharmaceutical industries are now in search of quick methods to identify and synthesize potent antimicrobial agents to control the bacterial infections. The field of bioinformatics has emerged as a supportive tool to solve this problem. Past few years have produced a lot of bioinformatics based prediction data for antibacterial compounds as well as for potential immunogenic behavior of proteins and peptides [62-66]. However all such prediction based methods strongly demand clinical trials and studies in animal models for their further use. As discussed in Chapter 1, spores and spore surface proteins can serve beneficial in many other ways. Spores have been used as drug vehicles, in probiotic treatments whereas spore coat proteins from *B. subtilis* have been used in surface display systems. For the identified set of proteins from the spore surface of three spore formers we analysed the potential peptides that can be used as antimicrobial agents as well as are immunogenic in behavior.

**Anti-microbial peptides (AMPs)**

Antimicrobial peptides (AMPs) are relatively short polypeptides (12-100 amino acids) that are positively charged (net charge of +2 to +9, most commonly +4 to +6) and amphiphilic. Such peptides may play a role in innate and adaptive immunity, including e.g. immuno-modulation, chemotaxis, inflammatory response, and wound repair [67] and therefore such peptides could have applications as therapeutics against biologically harmful agents. Usually the curative measures taken against bacterial or viral infections include vaccination, passive antibody therapy and antibiotic treatments. Antimicrobial peptides such as Nisin have been shown to affect the spore outgrowth [68] and thus can be used to control spore-mediated infections. In a similar way we ventured the antimicrobial potential of the set of peptides identified from the three spore forming organisms in our study. Using the AMPA automated web tool [69]. A total of 63 peptides from 57 proteins (across three organisms) were identified to have 0-5% probability of being non-antimicrobial peptides *id est* a probability that the predicted stretch of amino
Figure 3. Plot of molecular weight and pI versus GRAVY-score. The pI and molecular mass are plotted for the GRAVY-scores of the identified spore coat & exosporium proteins of B. cereus ATCC 14579, C. difficile 630, and B. subtilis 168 using IBM SPSS statistical tool v. 20. GRAVY-scores were calculated according to the values from Kyte and Doolittle [46].
acid sequence is found by chance in a non-antimicrobial protein. The AMPA algorithm is based on the antimicrobial propensity scale obtained from high-throughput screening results from bactenecin 2A [70]. Of the 63 peptides, 20 peptides had probability of 5%, 14 had probability values of 0% i.e. highest chance of being an antimicrobial peptide stretch (Table 3) and remaining 29 peptides had probabilities in the range 1% - 4%. Notably, the AMPA server predicted 4% probability for the Nisin sequence stretch. Also the predicted average antimicrobial index for Nisin was 0.227 and this value for all the peptides shown in Table 3 ranged from minimum of 0.182 (for CotG) to maximum of 0.215 (for BclC).

Immunogenicity predictions for the identified peptides

We also predicted the immunogenic potential of the identified tryptic peptides using the automated server called POPI v. 2.0 [66]. Considering about 428 human MHC class I binding peptides belonging to four classes of immunogenicity (established by MHCPEP database) as the initial dataset and based on the propensities of amino acids occurring in these peptides with regards to 23 physicochemical properties such as secondary structure, molecular volume, codon diversity, electrostatic charge etc., POPI predicts the overall possibility of a peptide being immunogenic [66]. In Chapter 1 the results for B. subtilis are discussed. We extended the same analysis to the peptides identified from more pathogenic species i.e. B. cereus and C. difficile. Most of the peptides were predicted to have no or little immunogenic potential towards cytotoxic (CTLs) and helper (HTLs) T-cells. From B. cereus a single peptide - LDILGIVAEGNVSR - from inosine-uridine preferring nucleoside hydrolase (IunH/BC_2889) showed a likelihood of being highly immunogenic to both CTLs and HTLs. Both organisms contained similar number of peptides with moderate to high immunogenic potential towards CTLs and HTLs (Table 4).

Concluding remarks

The biological role of spores is well known and discussed in details in this thesis. Their impeccable resistance characters, their unique structure have been studied and described elsewhere in the literature. Protein structure prediction has become an important application for bioinformatics and the knowledge of the structure is important in understanding the function of a protein. In an attempt to characterize the identified spore surface proteins, in situ, for their potential roles in spore integrity a detailed bioinformatics analysis was undertaken. The results showed that spore surface proteins have peculiar sequence properties and therefore structures or surface topologies. There were also certain amino acid biases in the sequences of spore surface proteins. These amino acids could be very important in imparting the spores with such resistant outer layers. Dependence of only a small fraction of proteins on the protein translocation and
Molecular properties of spore surface proteins

Table 3. Prediction of anti-microbial peptides (AMPs) from the identified spore surface proteins.

<table>
<thead>
<tr>
<th>Protein name</th>
<th>Uniprot ID</th>
<th>Residues</th>
<th>Avg. antimi crobial index</th>
<th>Predicted AMP sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>B. subtilis 168</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YabQ</td>
<td>P37559</td>
<td>186-210</td>
<td>0.197</td>
<td>KGAGFLKKKKKLLITIWAITF</td>
</tr>
<tr>
<td>YdhD</td>
<td>O05495</td>
<td>407-419</td>
<td>0.205</td>
<td>GPWLLRKFICTRK</td>
</tr>
<tr>
<td>YfbD</td>
<td>BG13050</td>
<td>44-66</td>
<td>NA</td>
<td>PCEKKKKKKHCFSDKKHHSCCH</td>
</tr>
<tr>
<td>CoxA</td>
<td>P94446</td>
<td>185-197</td>
<td>0.213</td>
<td>DKLGLRKLHKMN</td>
</tr>
<tr>
<td>CotG</td>
<td>P39801</td>
<td>32-166</td>
<td>0.182</td>
<td>SYCYSYKSCSHKKSHKKSG</td>
</tr>
</tbody>
</table>

B. cereus ATCC 14579

<table>
<thead>
<tr>
<th>Protein name</th>
<th>Uniprot ID</th>
<th>Residues</th>
<th>Avg. antimi crobial index</th>
<th>Predicted AMP sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC_2030 (cotG)</td>
<td>Q81EF0</td>
<td>38-125</td>
<td>0.189</td>
<td>HHCTTGCKCTTRGCRPRTRCTR</td>
</tr>
<tr>
<td>BC_2639 (sleB)</td>
<td>Q81CV2</td>
<td>4988-5009</td>
<td>0.212</td>
<td>HRVQWVTRFVTRKPRRKVLRVT</td>
</tr>
<tr>
<td>BC_3547 (beIC)</td>
<td>Q81AL6</td>
<td>923-940</td>
<td>0.206</td>
<td>TATPSKTIWPQIKKGKHIC</td>
</tr>
<tr>
<td>BC_4387</td>
<td>Q817Y7</td>
<td>25-41</td>
<td>0.191</td>
<td>VFSKRMKARKMKSY</td>
</tr>
</tbody>
</table>

C. difficile 630

<table>
<thead>
<tr>
<th>Protein name</th>
<th>Uniprot ID</th>
<th>Residues</th>
<th>Avg. antimi crobial index</th>
<th>Predicted AMP sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD1581</td>
<td>Q186D6</td>
<td>141-159</td>
<td>0.192</td>
<td>GNKNCKCHCKCNCR</td>
</tr>
<tr>
<td>CD2399</td>
<td>Q181Y3</td>
<td>57-68</td>
<td>0.196</td>
<td>LVAKPKKYRGR</td>
</tr>
<tr>
<td>CD3170 (eno)</td>
<td>Q181T5</td>
<td>416-429</td>
<td>0.212</td>
<td>QARYCVLKSYPYNK</td>
</tr>
</tbody>
</table>

Only the peptides with probability value of 0% are shown. Note: here probability is the misclassification probability value, _id est_ the probability that the predicted stretch is found by chance in a non-antimicrobial protein. NA not available.

Table 4. Prediction for the immunogenic potential of identified peptides.

<table>
<thead>
<tr>
<th>Protein name</th>
<th>No. of peptides (in %) showing immunogenic potential towards</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CTLs</td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
</tr>
<tr>
<td>B. subtilis 168</td>
<td>28.0</td>
</tr>
<tr>
<td>B. cereus ATCC 14579</td>
<td>19.7</td>
</tr>
<tr>
<td>C. difficile 630</td>
<td>17.0</td>
</tr>
</tbody>
</table>

Only the no. of peptides predicted to have moderate and high immunogenic potentials are shown. High (PD<sub>50</sub> < 1 nM) and Moderate (PD<sub>50</sub> = 100 nM - 1 nM). PD<sub>50</sub> is the protective dose that protects 50% of the animals challenged.
localization machinery in cells may make spore formation a relatively easy process. The study of the hydrophobicity patterns and pI distributions of the proteins is important to understand the surface character of spores as exemplified by the example of the dairy industry. Lastly, bioinformatic predictions of the bioactivity of peptides could lead to fast and relatively simple drug designing processes thereby aiding in the control of resistant pathogens.

**Bioinformatics tools used in this study**

**Derivation of amino acid composition**

The sequences of identified proteins were subjected to the Amino acid calculator (http://proteome.gs.washington.edu/cgi-bin/aa_calc.pl). The obtained numbers for each amino acid were converted to % manually and averaged over all the proteins from the organism.

**Identification of internal tandem repeats**

The protein sequences were submitted to automated T-REKS server [34] http://bioinfo.montp.cnrs.fr/?r=t-reks. Default parameters were used for repeat identification.

**Localization & signal peptide estimation**

The sequences of proteins were submitted to TMHMM server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) for identification of trans-membrane helices [71]. SignalP v. 4.1 (http://www.cbs.dtu.dk/services/SignalP/) tool[72] was used for prediction of signal sequences.

**Prediction of hydropathy nature and pI**

The hydropathy nature and pI values were predicted by using ProtParam tool http://web.expasy.org/protparam from the ExPASy proteomics server [73].

**Identification of antimicrobial peptides.**

An automated server AMPA (http://tcoffee.crg.cat/apps/ampa/do) [69] was used for prediction of antibacterial character of identified peptides. Default threshold parameters were used.

**Immunogenicity predictions.**

Automated server POPI v. 2.0 (http://iclab.life.nctu.edu.tw/POPI/) [66] was used for immunogenicity predictions of identified peptides.

**References**

Molecular properties of spore surface proteins


42. Hofmeister A. Activation of the proprotein transcription factor pro-sigmaE is associated with its progression through three patterns of subcellular localization during sporulation in *Bacillus subtilis*. Journal of bacteriology. 1998;180(9):2426-33. Epub 1998/05/09.
Molecular properties of spore surface proteins


