A proposed nomenclature for cell wall proteins of *Clostridium difficile*

Robert P. Fagan, Claire Janoir, Anne Collignon, Paola Mastrantonio, Ian R. Poxton and Neil F. Fairweather

1Division of Cell and Molecular Biology, Centre for Molecular Microbiology and Infection, Imperial College London, London SW7 2AZ, UK
2Université de Paris-Sud XI, USC INRA EA 4043, Faculté de Pharmacie, Châtenay-Malabry Cedex, France
3Department of Infectious, Parasitic and Immune-Mediated Diseases, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy
4Medical Microbiology, Centre for Infectious Diseases, University of Edinburgh College of Medicine and Veterinary Medicine, Edinburgh EH16 4SB, UK

Strains of *Clostridium difficile* produce a number of surface-localized proteins, including the S-layer proteins (SLPs) and other proteins that have suspected roles in pathogenesis. During the Third International *C. difficile* Symposium (Bled, Slovenia, September 2010) discussions were held on standardization of nomenclature. Gene designations were proposed for the large family of cell wall proteins that are paralogues of the SLP and contain putative cell wall binding motifs. This paper summarizes the agreed nomenclature, which we hope will be used by research groups currently active in the field.

**INTRODUCTION**

*Clostridium difficile* is now recognized as an important bacterial pathogen, particularly in health-care settings where it can cause serious enteric disease (Bartlett, 2006; Rupnik *et al.*, 2009). While symptoms of disease are mediated largely by two potent cytotoxins, toxin A (TcdA) and toxin B (TcdB) (Kuehne *et al.*, 2010; Lytras *et al.*, 2009), many other factors are thought to be essential for colonization of the enteric system to allow toxin production and eventual sporulation. *C. difficile* produces many surface-localized proteins which may contribute to colonization through, for example, direct adherence to host cells or by competition with the resident microbiota. Early studies showed that diverse strains of *C. difficile* produced two major surface-associated proteins that could be extracted from the cell by treatment with EDTA, urea or low-pH glycine treatments (Cerquetti *et al.*, 1992; Kawata *et al.*, 1984; McCoubrey & Poxton, 2001; Sharp & Poxton, 1988; Takeoka *et al.*, 1991). These proteins, which migrated on SDS-PAGE with apparent molecular masses of between 35 and 55 kDa, displayed the ability to form a two-dimensional array which was visible by electron microscopy (Cerquetti *et al.*, 2000; Kawata & Masuda, 1984), an inherent property of surface-layer (S-layer) proteins (SLPs) from Bacteria and Archaea.

Genetic studies have identified several important points. Firstly, the two dominant proteins observed on the surface of a variety of strains were the result of proteolytic processing of SlpA, the product of a single gene termed *slpA* (Calabi *et al.*, 2001; Karjalainen *et al.*, 2001). These proteins were subsequently termed SLPs. Secondly, the differences in mobility of these proteins between strains, as visualized by SDS-PAGE, were due to differences in the lengths of the *slpA* genes, resulting in antigenically distinct forms of the SLPs. Thirdly, a family of genes (paralogues) related to *slpA* was identified in the genome of strain 630 (Calabi *et al.*, 2001; Karjalainen *et al.*, 2001; Savariou-Lacomme *et al.*, 2003; Sebiahia *et al.*, 2006). Members of this family, which comprises 29 genes, all contain three copies of the Pfam 04122 motif, which is a complex motif annotated as ‘putative cell wall binding repeat 2’. The idea was proposed (Karjalainen *et al.*, 2001) that these Pfam motifs mediate binding of the proteins encoded by these paralogues to the underlying cell wall. In some cases, these proteins carry a second domain which specifies a known or putative function. The best characterized examples are Cwp84 (Janoir *et al.*, 2007), a protease that cleaves the SlpA precursor (Dang *et al.*, 2010; Kirby *et al.*, 2009) and which has degradative activity on a variety of host cell extracellular matrix proteins (Janoir *et al.*, 2007), Cwp66, a putative adhesin (Waligora *et al.*, 2001), and CwpV, a phase-variable protein (Emerson *et al.*, 2009).

Abbreviations: CWPs, cell wall proteins; SLPs, S-layer proteins.
With an increasing number of research groups working on C. difficile and the advent of genome sequences that are uncovering increasing diversity between C. difficile strains, the need for a unified nomenclature of C. difficile S-layer and cell wall proteins is apparent. A similar situation existed a few years ago in regard to the nomenclature of the large clostridial cytotoxins, and a unified nomenclature was proposed (Rupnik et al., 2005) and has been universally adopted.

**PROPOSED NOMENCLATURE**

**SLPs**

All strains of C. difficile examined to date express two predominant proteins on the cell wall, which form the major components of the S-layer. The encoding gene, which appears to have a conserved genomic location between all strains studied to date, has been named slpA. The precursor protein is termed SlpA, and the mature proteins produced by the action of the protease Cwp84 are termed the HMW (high molecular weight) SLP and the LMW (low molecular weight) SLP. TheHMW SLP contains three Pfam 04122 motifs while the LMW SLP shows a high degree of antigenic variation between strains. We suggest that the prefix slp be reserved for those genes whose products have been shown experimentally to form a two-dimensional array, and not used for gene products found within the S-layer, as this could vary depending on the technique used for extraction of SLPs.

**Cell wall proteins**

In C. difficile 630, a family of 29 genes containing three Pfam 04122 motifs was identified. These genes were originally given the prefix ‘orf1’, ‘orf2’, etc. (Calabi et al., 2001; Karjalainen et al., 2001), but due to simultaneous publication, the genes were numbered differently. In addition, the genes cwp66 and cwp84 were individually named according to their molecular masses (Karjalainen et al., 2001). Following this convention, some other genes (Wright et al., 2005) were given a ‘cwp’ prefix (Emerson & Fairweather, 2009; Wright et al., 2008).

Ordinarily, a gene family would be named using only letters, e.g. slpA and cwpV. However, as there are over 26 paralogues, this is not possible. In addition, some names are in common use and we recommend keeping these. We propose that all members of the gene family containing Pfam 04122 motifs be termed CWPs, ‘clostridial wall proteins’ or more aptly ‘cell wall proteins’ and that the genes be given the prefix cwp. The exception is slpA, which, although a member of this gene family, has a clearly established name which relates to its function as an SLP. Where an existing name has been published and has been generally accepted, that name should remain in use. To date these are cwp84, cwp66, cwpV and cwp2. All other cwp genes have been assigned a number as outlined in Table 1. The ORF designation with a CD prefix given upon annotation of strain 630 is also shown. The vast majority of the cwp genes are conserved between 630 and R20291, a recently isolated ribotype 027 strain which has been sequenced at the genome level (Stabler et al., 2009). Genome sequencing of further C. difficile strains will undoubtedly reveal new members of the cwp gene family. We suggest that the naming be continued numerically, i.e. the next cwp gene identified should be named cwp30.

Several of the CWPs contain a passenger domain that may specify a function. The majority of these have not been demonstrated experimentally, but are based on the presence of motifs. The putative functional motifs and the architectures of members of the CWP family are shown in Fig. 1.

**Table 1. Nomenclature of CWPs of C. difficile**

<table>
<thead>
<tr>
<th>Gene name*</th>
<th>Gene number†</th>
<th>Mol. mass‡ (kDa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>slpA</td>
<td>CD2793</td>
<td>76.1</td>
</tr>
<tr>
<td>cwp2</td>
<td>CD2791</td>
<td>66.5</td>
</tr>
<tr>
<td>cwp66</td>
<td>CD2789</td>
<td>66.8</td>
</tr>
<tr>
<td>cwp84</td>
<td>CD2787</td>
<td>87.3</td>
</tr>
<tr>
<td>cwp5</td>
<td>CD2786</td>
<td>57.2</td>
</tr>
<tr>
<td>cwp6</td>
<td>CD2784</td>
<td>73.0</td>
</tr>
<tr>
<td>cwp7</td>
<td>CD2782</td>
<td>39.3</td>
</tr>
<tr>
<td>cwp8</td>
<td>CD2799</td>
<td>68.5</td>
</tr>
<tr>
<td>cwp9</td>
<td>CD2798</td>
<td>52.5</td>
</tr>
<tr>
<td>cwp10</td>
<td>CD2796</td>
<td>67.0</td>
</tr>
<tr>
<td>cwp11</td>
<td>CD2795</td>
<td>58.7</td>
</tr>
<tr>
<td>cwp12</td>
<td>CD2794</td>
<td>57.5</td>
</tr>
<tr>
<td>cwp13</td>
<td>CD1751</td>
<td>87.2</td>
</tr>
<tr>
<td>cwp14</td>
<td>CD2735</td>
<td>51.4</td>
</tr>
<tr>
<td>cwpV</td>
<td>CD0514</td>
<td>166.7</td>
</tr>
<tr>
<td>cwp16</td>
<td>CD1035</td>
<td>74.7</td>
</tr>
<tr>
<td>cwp17</td>
<td>CD1036</td>
<td>74.2</td>
</tr>
<tr>
<td>cwp18</td>
<td>CD1047</td>
<td>37.4</td>
</tr>
<tr>
<td>cwp19</td>
<td>CD2767</td>
<td>77.8</td>
</tr>
<tr>
<td>cwp20</td>
<td>CD1469</td>
<td>111.1</td>
</tr>
<tr>
<td>cwp21</td>
<td>CD3192</td>
<td>61.1</td>
</tr>
<tr>
<td>cwp22</td>
<td>CD2713</td>
<td>72.0</td>
</tr>
<tr>
<td>cwp23</td>
<td>CD1803</td>
<td>63.8</td>
</tr>
<tr>
<td>cwp24</td>
<td>CD2193</td>
<td>51.3</td>
</tr>
<tr>
<td>cwp25</td>
<td>CD0844</td>
<td>33.8</td>
</tr>
<tr>
<td>cwp26</td>
<td>CD1233</td>
<td>52.0</td>
</tr>
<tr>
<td>cwp27</td>
<td>CD0440</td>
<td>41.0</td>
</tr>
<tr>
<td>cwp28</td>
<td>CD1987</td>
<td>51.7</td>
</tr>
<tr>
<td>cwp29</td>
<td>CD2518</td>
<td>50.0</td>
</tr>
</tbody>
</table>

*There is no cwp1 gene. Proteins have identical names, but with a capitalized first letter and not italicized. Gene names in the current literature are: slpA (Calabi et al., 2001), cwp2 and cwpV (Emerson et al., 2009), cwp66 (Waligora et al., 2001) and cwp84 (Karjalainen et al., 2001).
†Gene number according to Sebaihia et al. (2006). All genes except cwp29 have paralogues in R20291.
‡Predicted molecular mass of protein prior to any processing event.
Other surface-localized proteins

*C. difficile* genomes encode a number of putative surface-localized proteins in addition to the CWP family. Some of these have been characterized at the molecular level, for example *fbp68* encoding a fibronectin-binding protein (Hennequin et al., 2003). At least one sortase gene has been identified in 630 and R20291 genomes and a number of putative sortase substrates have also been identified. We suggest that any names assigned to these genes and proteins remain unchanged and that newly characterized genes, for example the sortase substrates, be named by the group who first describe their properties.

**REFERENCES**


