Journal of Medical Microbiology (2006), **55**, 69–83 DOI 10.1099/jmm.0.46204-0

Sequence and phylogenetic analysis of the gene for surface layer protein, *slpA*, from 14 PCR ribotypes of *Clostridium difficile*

Déirdre Ní Eidhin,[1] Anthony W. Ryan,[1] Rachael M. Doyle,[1,2]† J. Bernard Walsh[1,2] and Dermot Kelleher[1]

Correspondence
Déirdre Ní Eidhin
dniidhin@tcd.ie

Department of Clinical Medicine and Dublin Molecular Medicine Centre, Trinity College Dublin, Trinity Centre for Health Sciences[1] and Mercer's Institute for Research in Ageing[2], St James's Hospital, James's St., Dublin 8, Ireland

Clostridium difficile is the commonest cause of antibiotic-associated diarrhoea, with the hospitalized elderly being at particular risk. The organism makes a crystalline surface protein layer (S-layer), encoded by the *slpA* gene, the product of which is cleaved to give two mature peptides which associate to form the layer. The larger peptide (high molecular weight; HMW), derived from the C-terminal portion of the precursor, is relatively conserved, whereas the smaller peptide (low molecular weight; LMW), derived from the N-terminal portion of the precursor, is a dominant antigen which substantially forms the basis for serotyping of isolates. PCR ribotyping is a more discriminatory typing method, based on the intergenic rRNA. We obtained the sequence for *slpA* and some flanking DNA from a collection of *C. difficile* strains of 14 ribotypes isolated from elderly patients. Sequences from different ribotypes were compared with one another and with published sequences. Sequences from *C. difficile* ribotypes 046 and 092 were identical. Sequences from ribotype pairs 005 and 054, 012 and 046/092, 014 and 066 and 031 and 094 differed by 1–3 nt in the *slpA* gene. There were ultimately nine ribotypes or groups of ribotypes with very different *slpA* sequences, particularly in the region encoding the LMW peptide. The sequence from ribotype 002 was very different from previously published sequences. The DNA segment sequenced included the 5′ 315 bp of a *secA* homologue, encoding a putative transport protein required for peptide secretion across the plasma membrane. The amino acid sequences of the predicted HMW peptides were aligned and a neighbour-joining tree was produced using 10 000 bootstrap replicates. The predicted SecA N-terminal region was similarly analysed. For both SlpA and SecA, a strong association was found between ribotypes 012, 046/092, 017, 031 and 094. Ribotypes 001 and 078 formed part of this clade for SlpA but not SecA, indicating independent evolution for *slpA* and *secA*, presumably because they come under different selection pressures.

Received 17 June 2005
Accepted 23 September 2005

INTRODUCTION

Clostridium difficile is now the leading cause of nosocomial diarrhoea among hospitalized patients undergoing antibiotic treatment and is associated with substantial morbidity and mortality. The spectrum of disease ranges from mild diarrhoea to pseudomembranous colitis, which can be fatal (Kelly & LaMont, 1998). The infection routinely requires isolation of affected patients, additional antibiotic therapy and a prolongation of hospital stay, which has implications for patient turnover and health economics (Kyne *et al.*, 2002).

C. difficile makes a crystalline protein surface layer (S-layer), a structural feature of many bacteria. S-layers have been ascribed various roles including nutrient uptake, exclusion of noxious substances, antiphagocytosis and colonization (Sára & Sleytr, 2000). In *C. difficile*, the S-layer is the predominant surface antigen, and a strong serum IgG response has been found among convalescent patients (Doyle, 2004; Pantosti *et al.*, 1989). There are a number of variant types which are serologically distinct, substantially forming the basis for serotyping of *C. difficile* strains (Delmée *et al.*, 1986, 1990).

†Present address: St Columcille's Hospital, Loughlinstown, Co. Dublin, Ireland.

The GenBank/EMBL/DDBJ accession number for the *slpA* and flanking sequences of *C. difficile* isolates are DQ060625–DQ060643.

Compositional data on SlpA from different PCR ribotypes are available as supplementary material in JMM Online.

The S-layer is encoded by the *slpA* gene, the product of which contains a cleavable signal sequence and is further cleaved to give two mature peptides which then associate to form the S-layer (Calabi *et al.*, 2001; Karjalainen *et al.*, 2001). The peptide derived from the N-terminal region of the precursor is smaller and more highly variable, whereas the peptide derived from the C-terminal region is relatively conserved. For convenience, the two products are respectively known as the low molecular weight (LMW) and high molecular weight (HMW) peptides. The LMW peptide appears to be the main serotyping antigen (Poxton *et al.*, 1999). The HMW peptide has sequence similarity to the *N*-acetylmuramoyl-L-alanine amidase from *Bacillus subtilis*, and has been shown to possess amidase activity (Calabi *et al.*, 2001). A number of other genes encoding putative amidases, known as *slpA* paralogues, occur in the vicinity of *slpA* (Calabi & Fairweather, 2002).

A recent study (Doyle, 2004) described the isolation and typing of *C. difficile* from a large number of patients who developed diarrhoea while attending the care of the elderly unit at St James's Hospital, Dublin. A total of 14 types were identified within this population by PCR ribotyping, which is based on polymorphism of the intergenic DNA between the regions encoding 16S and 23S rRNA (O'Neill *et al.*, 1996). Since 116 PCR ribotypes of *C. difficile* have been identified to date (Stubbs *et al.*, 1999), this typing method is considered more discriminatory than serotyping, which only distinguishes 21 types (Delmée *et al.*, 1990). Our primary objective was to sequence the *slpA* gene from the PCR ribotypes identified and to compare the sequences with published data.

Calabi *et al.* (2001) reported a large ORF immediately downstream of *slpA* with strong sequence similarity to the *secA* gene of other bacterial species. The *secA* product is an essential component of the general secretory pathway in the Bacteria. It is a large protein which interacts with nascent proteins and with other components of the export pathway and provides energy for translocation by ATP hydrolysis (Schmidt & Kiser, 1999). This multifunctionality is reflected in a high degree of sequence conservation for *secA* between species. The sequence we obtained from each isolate included 315 bp from the 5′ end of this gene. The consistent presence of a short segment of a conserved housekeeping gene in proximity to *slpA* provided reassurance that we had sequenced the genuine *slpA* allele in each case and not one of its many paralogues. We also constructed phylogenetic trees for segments of the translated *slpA* and *secA* genes and compared their variability between ribotypes.

METHODS

***C. difficile* isolates and culture.** *C. difficile* isolates were selected from a collection of all patient isolates obtained from July 1998 to December 1999 at the care of the elderly unit at St James's Hospital (Doyle, 2004). Isolates were typed by Jon Brazier at the Anaerobe Reference Unit, Public Health Laboratory Service (since renamed the Health Protection Agency), University Hospital of Wales, Cardiff, and identified by toxin production and PCR ribotype according to the scheme of O'Neill *et al.* (1996), based on the variable 16S–23S intergenic spacer region. The strains, identified by their Anaerobe Reference Unit designation, are listed in Table 1. Where possible, more than one isolate was selected, separated by date of isolation or ward or both. Three such well-separated isolates were selected for ribotype 001 (which accounted for approximately half of all isolates) and two each for ribotypes 002, 005 and 012. Single isolates were tested for the remaining ribotypes, either because only one isolate was available or because isolates occurred close together. Cultures were grown in anaerobic jars on Columbia blood agar (Lab M) with 7 % defibrinated horse blood, fastidious anaerobe broth (Lab M) or pre-reduced brain heart infusion broth containing 0·5 % (w/v) thioglycolate (BHI-TG). Stocks were maintained in cooked meat medium (Oxoid), made up in fastidious anaerobe broth.

Preparation of SlpA from *C. difficile*. SlpA was prepared from early stationary phase cultures grown in BHI-TG by extraction with 8 M urea as described by Cerquetti *et al.* (2000). Extraction was done in the presence of Complete protease inhibitor cocktail (Roche Diagnostics). Extracts were dialysed against 50 mM Tris/HCl pH 7·4 and the protein content was measured by Bradford assay. Samples (5 μg total protein) were visualized by SDS-PAGE on 12 % total monomer gels stained with Coomassie brilliant blue R. Relative molecular mass standards (Sigma wide molecular weight range) were included on each gel.

DNA isolation, amplification and sequencing. DNA was isolated from overnight BHI-TG-grown *C. difficile* cultures using the Gentra Puregene DNA isolation kit for yeast and Gram-positive bacteria with an additional proteinase K step (400 μg ml^{-1} added to the lytic buffer, incubation for 1 h at 55 °C followed by 10 min at 80 °C) and omission of RNase treatment. DNA was amplified by PCR using HotStar *Taq* polymerase (Qiagen), with an initial DNA denaturation step of 15 min at 95 °C followed by 30 cycles of denaturation for 1 min at 94 °C, annealing at 37 °C for 30 s and extension at 68 °C for 2 min 45 s. A final extension was done at 72 °C for 10 min. The primers used were (forward) −107F (5′-ATGGATTATTATAGAGATGTGAG-3′) or −27F (5′-AATATAATGTTGGGAGG-3′) and (reverse) +562R (5′-ACCTTCACCAGTTTTCAT-3′). Primer −27F was used to amplify DNA from ribotypes 002, 010, 014 and 066. Product purity, size and yield were checked on 0·8 % agarose gels using lambda DNA cut with *Eco*RI and *Bam*HI or with *Eco*RI and *Hin*dIII as standard. DNA products were cloned in the pBAD ThioTOPO vector and transformed into competent *Escherichia coli* Top10, supplied as One Shot competent cells [genotype F^- *mcrA* Δ(*mrr–hsdRMS–mcrBC*) *ϕ*80*lacZ*ΔM15 Δ*lac*X74 *recA1 deoR araD139* Δ(*ara–leu*)7697 *galU galK rpsL* (Str^R) *endA1 nupG*], as recommended by the manufacturer (Invitrogen). Recombinants were selected on LB agar supplemented with ampicillin and checked by restriction digestion and plasmid DNA was isolated for sequencing. Sequencing was also carried out on PCR products directly. DNA was sequenced commercially at the Biochemistry Department, University of Cambridge. We designed the custom primers required to complete sequencing from both strands.

Analysis of *slpA* sequences. Signal sequences and their cleavage sites were predicted by the SignalP tools (http://www.cbs.dtu.dk/services/SignalP), revised most recently by Bendtsen *et al.* (2004), which combine two predictors based on neural network and hidden Markov model algorithms. The positions of both cleavage sites on the SlpA precursor protein were also predicted by comparison of deduced sequences with experimentally determined N-terminal amino acid sequences of the mature LMW and HMW peptides (Calabi *et al.*, 2001; Cerquetti *et al.*, 2000). Multiple sequence alignment was done using the CLUSTAL W tool (Thompson *et al.*, 1994) with the Blosum matrix, a gap opening penalty of 10 and a gap extension penalty of

Table 1. *C. difficile* strains used in this study

Isolates were obtained from St James's Hospital, Dublin (1998–2000) (Doyle, 2004). Serogroups are as assigned by Stubbs *et al.* (1999) and Brazier (2001).

Isolate	PCR ribotype	Associated serogroup	Toxin A/B production	GenBank accession no. for sequenced region
R12879	001	G	+/+	DQ060625
R13537	001	G	+/+	DQ060626
R14637	001	G	+/+	DQ060627
R13541	002	A_2	+/+	DQ060628
R13549	002	A_2	+/+	DQ060629
R12884	005	Unknown	+/+	DQ060630
R14640	005	Unknown	+/+	DQ060631
R13700	010	D	−/−	DQ060633
R13550	012	C	+/+	DQ060634
R12882	012	C	+/+	DQ060635
R12885	014	H	+/+	DQ060638
R13702	017	F	−/+	DQ060640
R13711	031	K	−/−	DQ060641
R12883	046	Unknown	+/+	DQ060636
R13708	054	A_1	+/+	DQ060632
R13699	066	A_9	−/−	DQ060639
R13540	078	Unknown	+/+	DQ060643
R12871	092	Unknown	+/+	DQ060637
R12865	094	Unknown	+/+	DQ060642

0·2. Secondary structure prediction was based on the consensus from the SOPM (Geourjon & Deléage, 1994), HNN (Guermeur, 1997), DPM (Deléage & Roux, 1987), DSC (King & Sternberg, 1996), GOR IV (Garnier *et al.*, 1996), PHD (Rost & Sander, 1994), PREDATOR (Frishman & Argos, 1996) and SIMPA96 (Levin, 1997) tools, using the NPS interface (Combet *et al.*, 2000). Internal peptide repeats were detected using the RADAR tool (Heger & Holm, 2000). Molecular mass and pI of predicted mature peptides were calculated by the Compute pI/Mw tool of the ExPASy proteomics server of the Swiss Institute of Bioinformatics (http://us.expasy.org/tools/pi_tool.html; Gasteiger *et al.*, 2003). Codon usage was calculated by the CODONFREQUENCY tool from the Wisconsin Package (Accelrys Inc.) and values for relative synonymous codon usage were calculated. Relative synonymous codon usage is defined as the observed occurrence of a given codon divided by the expected occurrence. Values close to 1 are indicative of a lack of bias. Rho-independent terminators were detected by the TERMINATOR tool from the Wisconsin package (Brendel & Trifonov, 1984). To study evolutionary relationships between ribotypes, amino acid sequences were aligned using CLUSTAL W as before. Pairwise Poisson correction distances, which correct for multiple substitutions at the same site, were calculated from the resulting alignment and unrooted neighbour-joining trees were drawn from the resulting distance matrix using MEGA2 software (Kumar *et al.*, 2001). Bootstrap analyses of the phylogeny were performed using 10 000 bootstrap replications.

RESULTS

Amplification and sequencing of *slpA* gene and flanking DNA

The *slpA* gene and flanking DNA was sequenced from strains of all 14 ribotypes isolated from patients at St James's Hospital over a 16 month period (Table 1). Forward primer −107F, based on non-coding sequence starting at position −107 from the *slpA* gene from strain 630 (ribotype 012), was not successful in amplifying sequences from all ribotypes. Forward primer −27F, used to amplify DNA from ribotypes 002, 010, 014 and 066, was based on a 17 nt stretch beginning 26–27 nt upstream of the *slpA* gene which was conserved among the remaining ribotypes. A single reverse primer, +562R, based on nt 298–315 of the *secA* homologue from strain 630 (Calabi *et al.*, 2001), was used throughout. A single product was obtained from each isolate under the PCR conditions chosen. Products varied from 2496 to 2927 bp depending on ribotype (Fig. 1). Each fragment contained a complete ORF of 1830 to 2301 bp, presumed to be *slpA*, and terminated in a fragment containing 315 bp from the 5′ end of another ORF, predicted to be a *secA* gene based on similarity to the allele in the genome sequence. The intergenic DNA between the two ORFs varied from 202 to 268 bp. The upstream DNA segment was 106–107 bp for products generated from primer −107F and 26–27 bp for products generated from primer −27F.

Comparison of sequences of *slpA* and flanking DNA from Dublin isolates

Sequences obtained from different isolates of the same ribotype were all identical. The fragments from ribotypes 046 and 092 were also identical, and their sequences were treated as one in subsequent analyses. The sequences from some ribotypes were almost identical, with 1–3 nt differences

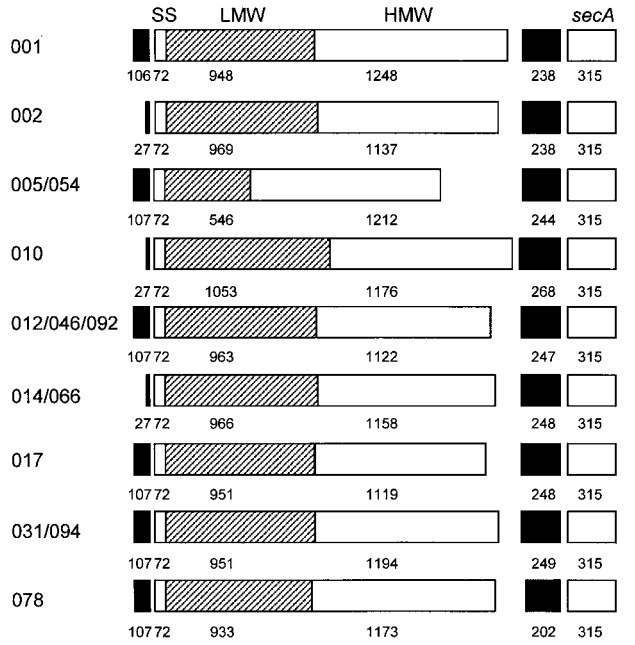


Fig. 1. Layout of DNA fragments containing *slpA* and flanking DNA for different PCR ribotypes of *C. difficile*. Numbers refer to length in nucleotides. Ribotypes with identical layouts are grouped together. Intergenic regions are shown as filled bars, and the coding regions for the signal sequence (SS), LMW and HMW peptides are shown for *slpA*. All fragments contain the 5′ 315 bp from the *secA* gene.

which always occurred in *slpA* (Table 2). These differences always translated to amino acid differences, often radical. Fig. 2(a) shows a CLUSTAL W alignment of the translated *slpA* ORFs from the different ribotypes, with residue differences between pairs of nearly identical peptides underlined. Thus ribotype pairs 005 and 054 differed by 3 nt and 2 amino acid residues and pairs 012 and 046/092, 014 and 066 and 031 and 094 each differed by 1 nt and 1 amino acid residue. From the 14 ribotypes, we thus identified nine classes which varied substantially in sequence.

As noted by others, the sequence of the predicted HMW peptide was far more strongly conserved than that of the LMW peptide, with the alignment showing substantial blocks of similar or identical sequence interspersed with stretches which varied in length and in sequence between ribotypes. The pattern was quite different for the available 105 residues of the SecA homologue sequence (Fig. 2b), which showed complete alignment and strong conservation (over 62 % identical residues).

Comparison with other published *slpA* sequences

Sequence differences between *slpA* genes from the ribotype groups identified among the Dublin isolates are summarized in Table 2 and comparisons are made with published sequences. Many of the strains from which the *slpA* gene has been sequenced are known by serogroup only and, since there are more ribotypes than serogroups, a serogroup designation alone is often ambiguous. Moreover, not all ribotypes have been assigned to serogroups. Nonetheless, all published *slpA* sequences from ribotype 012 (or serogroup C) and 017 (or serogroup F) strains are identical to the corresponding sequences from our collection. Indeed, strong similarity was generally found between sequences from strains of a given serogroup. Serogroup A was an exception. This grouping is based primarily on a shared flagellar antigen, and subgroups are based on cross-reactivity of bacteria from which flagella have been removed mechanically (Delmée *et al.*, 1990). It is therefore to be expected that *slpA* sequences from strains of different serogroup A subgroups may vary. Surprisingly, the sequence from a serogroup A10 strain showed strong similarity to the sequence from serogroup A1 strains. The strain in question, the only A10 strain for which any *slpA* sequence is available, is peculiar among serogroup A strains in that it has no flagella and does not react with anti-flagellin serum (Delmée *et al.*, 1990). Also surprisingly, we found only 1 nt difference between *slpA* from ribotype 014 (serogroup H) and ribotype 066 (serogroup A9). We infer that limitations of the serogrouping method may result in discrepancies such as these.

We noted a close similarity between sequences from ribotypes 005, 016 and 054. Ribotype 005 has not been assigned to a serogroup, but its *slpA* sequence closely resembles that of a serogroup A1 strain (Karjalainen *et al.*, 2001) and that of ribotype 054, which has been assigned to serogroup A1. No sequence strongly similar to the *slpA* from ribotype 002 was found among published *slpA* sequences.

Prediction of post-translational cleavage sites

The predicted N-terminal amino acid sequence was well conserved for SlpA and showed the hallmarks of a Gram-positive cleavable sequence (Fig. 2a; van Wely *et al.*, 2001). For most ribotypes, both the neural network and hidden Markov models predicted the cleavage site just C-terminal of A24. This prediction concurs with experimental prediction based on N-terminal sequencing of the native SlpA LMW peptide for ribotypes 001, 012 and 017 and possibly 010, 014 and 066 (Calabi *et al.*, 2001; Cerquetti *et al.*, 2000). For ribotypes 002, 005 and 054, there were slight discrepancies between the predictors, with only the neural network predicting the cleavage site after position A24 for ribotype 002 and only the hidden Markov model making a similar prediction for ribotypes 005 and 054. This ambiguity may have been caused by the three consecutive alanine residues at positions 24–26 in these ribotypes (J. D. Bendtsen, personal communication), since alanine is strongly favoured in the position preceding the cleavage site. Position 24 occurs at the end of a highly conserved stretch of 7 amino acid residues with the consensus sequence SAAPVFA, which seems unlikely to be coincidental.

The position of the secondary cleavage site, to cleave the precursor SlpA into the LMW and HMW peptides, was

Table 2. Comparison of similar *slpA* sequences from Dublin (SJH) isolates and comparison with published *slpA* sequences

All Dublin isolates of a given ribotype had identical sequences; numbers of strains are indicated in parentheses. Serogroups and ribotypes are listed according to known correlations between ribotype and serogroup from Stubbs *et al.* (1999) and Brazier (2001) unless indicated. Since there are more serogroups than ribotypes and not all ribotypes have been assigned to a serogroup, it is more tenuous to use knowledge of serogroup to predict ribotype.

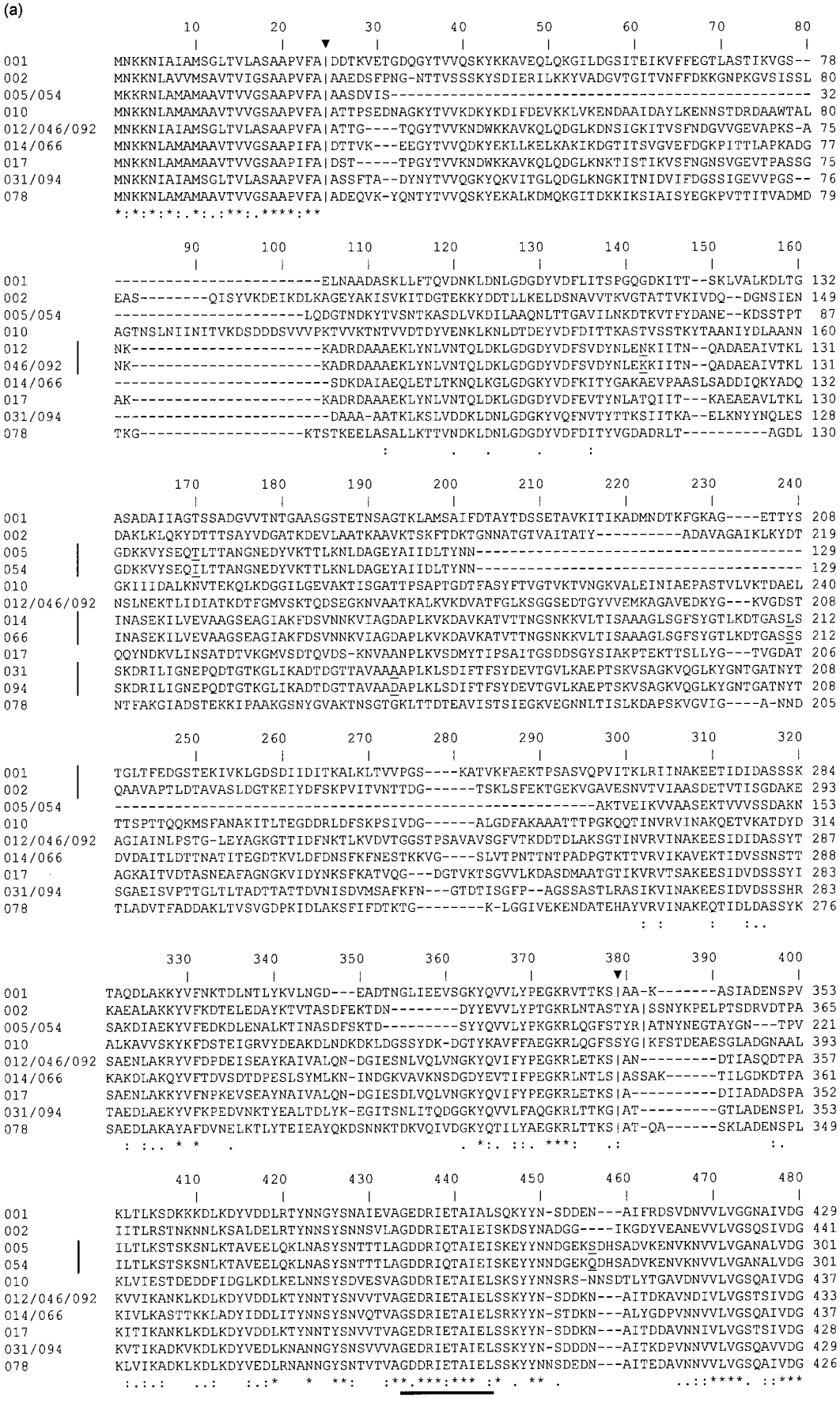
SJH ribotype (*n*)	Associated serogroup	Compared strain	Origin	Ribotype	Serogroup	Sequence differences		Reference*
						nt	aa	
001 (3)	G	R8366	UK	1†	G	0	0	AJ300676 (Calabi *et al.*, 2001)
		ATCC 43599	Belgium	001, 115	G†	8	1	*AF448128* (Karjalainen *et al.*, 2002)
		96-392	France	001, 115	G†	8	1	*AF448129* (Karjalainen *et al.*, 2002)
005 (2)	Unknown	ATCC 43594	Belgium	021, 054, 075	A_1†	4 (includes 3 extra nt)	4	*AF458877* (Karjalainen *et al.*, 2002)
		R13708	Ireland	054†	A_1	3	2	This study
054 (1)	A_1	167	USA	016‡	Unknown	4	3	AF478570 (Calabi & Fairweather, 2002)
		TO005	Canada	039, 067	A_{10}†	12 (includes gaps and extra nt)	19 (due to frame-shift)	*AF458878* (Karjalainen *et al.*, 2002)
010 (1)	D	ATCC 43597	Belgium	010	D†	0	0	*AF458880* (Karjalainen *et al.*, 2002)
		90-111	France	010	D†	0	0	*AF458881* (Karjalainen *et al.*, 2002)
		93-136	France	010	D†	0	0	*AF458882* (Karjalainen *et al.*, 2002)
		Y	UK	010‡	D	4	3	AF478571 (Calabi & Fairweather, 2002)
012 (2)	C	630	Switzerland	012†	C	0	0	Sanger sequencing project
		C253	Italy	012	C†	0	0	AJ291709 (Karjalainen *et al.*, 2001)
		ATCC 43596	Belgium	012	C†	0	0	*AF448123* (Karjalainen *et al.*, 2002)
		R12883	Ireland	046†	Unknown	1	1	This study
		R12871	Ireland	092†	Unknown	1	1	This study
046 (1)	Unknown	R12871	Ireland	092†	Unknown	0	0	This study
066 (1)	A_9	R12885	Ireland	014†	H	1	1	This study
		ATCC 43600	Belgium	014, 020	H†	2	1	*AF448365* (Karjalainen *et al.*, 2002)
		89-638	France	014, 020	H†	2	1	*AF448366* (Karjalainen *et al.*, 2002)
		90-204	France	014, 020	H†	2	1	*AF448367* (Karjalainen *et al.*, 2002)
017 (1)	F	R7404	UK	017†	F	0	0	AJ300677 (Calabi *et al.*, 2001)
		ATCC 43598	Belgium	017	F†	0	0	*AF448125* (Karjalainen *et al.*, 2002)
		GAI 95600	Japan	017	F†	0	0	*AF448126* (Karjalainen *et al.*, 2002)
		GAI 95601	Japan	017	F†	0	0	*AF448127* (Karjalainen *et al.*, 2002)
094 (1)	Unknown	R13711	Ireland	031†	K	1	1	This study
		ATCC 43602	Belgium	031, 053, 057	K†	3 (includes 1 extra nt, 1 nt gap)	18 (due to frame-shift)	*AF448368* (Karjalainen *et al.*, 2002)
		94-416	France	031, 053, 057	K†	As above	As above	*AF448369* (Karjalainen *et al.*, 2002)
		48-515	Belgium	031, 053, 057	K†	As above	As above	*AF448370* (Karjalainen *et al.*, 2002)
078 (1)	Unknown	9354	France	Unknown	A† (unknown subgroup)	3	2	*AF448120* (Karjalainen *et al.*, 2002)

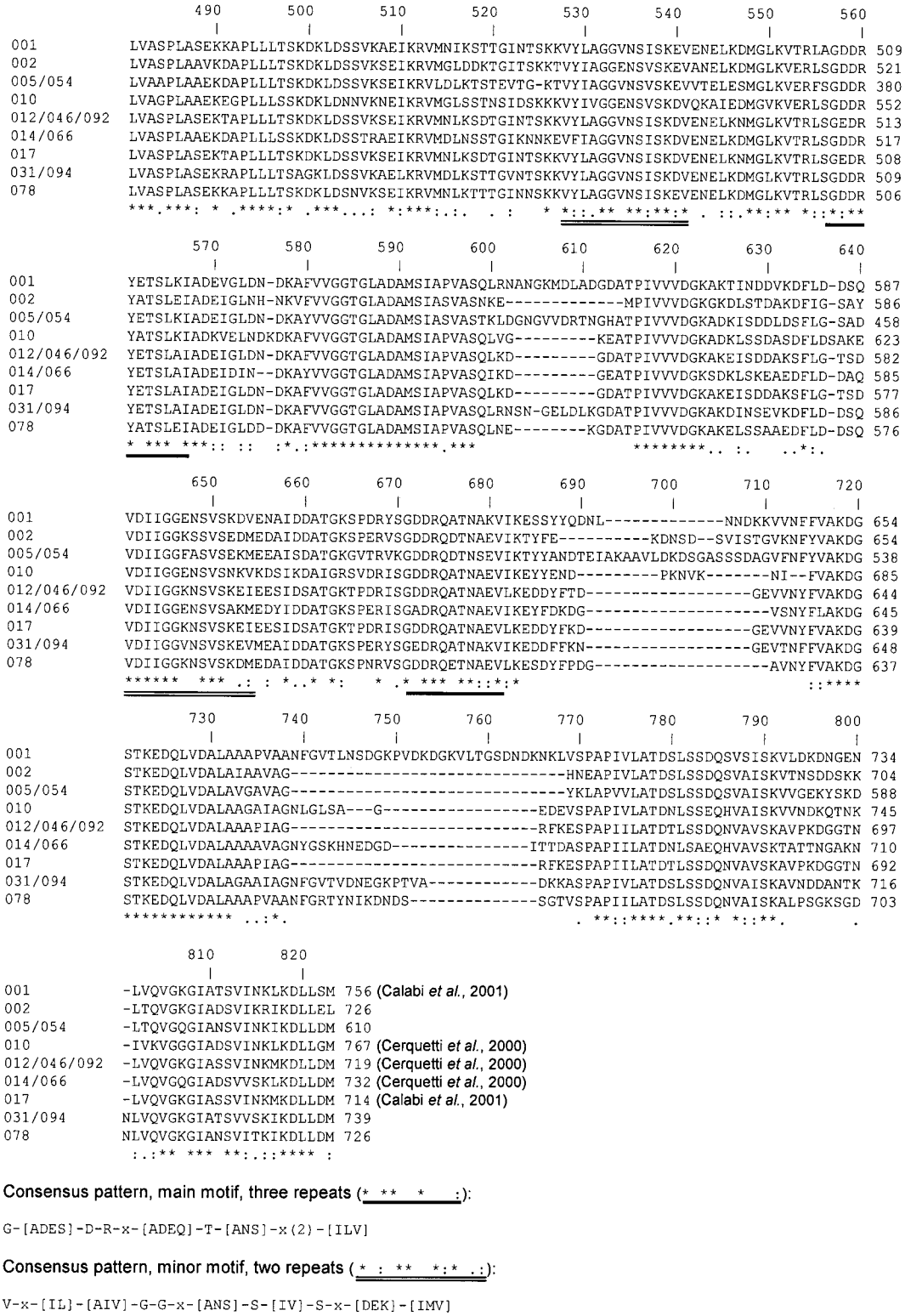
*Accession numbers in italics represent partial sequences (1017–1185 nt from 5′ region of *slpA*).

†From listed reference.

‡Personal communication from Neil Fairweather.

(a)


10 20 30 40 50 60 70 80
001 MNKKNIAIAMSGLTVLASAAPVFA|DDTKVETGDQGYTVVQSKYKKAVEQLQKGILDGSITEIKVFFEGTLASTIKVGS-- 78
002 MNKKNLAVVMSAVTVIGSAAPVFA|AAEDSFPNG-NTTVSSSKYSDIERILKKYVADGVTGITVNFFDKKGNPKGVSISSL 80
005/054 MKKRNLAMAMAAVTVVGSAAPVFA|AASDVIS---------- 32
010 MNKKNLAMAMAAVTVVGSAAPVFA|ATTPSEDNAGKYTVVKDKYKDIFDEVKKLVKENDAAIDAYLKENNSTDRDAAWTAL 80
012/046/092 MNKKNIAIAMSGLTVLASAAPVFA|ATTG----TQGYTVVKNDWKKAVKQLQDGLKDNSIGKITVSFNDGVVGEVAPKS-A 75
014/066 MNKKNLAMAMAAVTVVGSAAPIFA|DTTVK---EEGYTVVQDKYEKLLKELKAKIKDGTITSVGVEFDGKPITTLAPKADG 77
017 MNKKNLAMAMAAVTVVGSAAPIFA|DST-----TPGYTVVKNDWKKAVKQLQDGLKNKTISTIKVSFNGNSVGEVTPASSG 75
031/094 MNKKNIAIAMSGLTVLASAAPVFA|ASSFTA--DYNYTVVQGKYQKVITGLQDGLKNGKITNIDVIFDGSSIGEVVPGS-- 76
078 MNKKNLAMAMAAVTVVGSAAPVFA|ADEQVK-YQNTYTVVQSKYEKALKDMQKGITDKKIKSIAISYEGKPVTTITVADMD 79
90 100 110 120 130 140 150 160
001 ----------ELNAADASKLLFTQVDNKLDNLGDGDYVDFLITSPGQGDKITT--SKLVALKDLTG 132
002 EAS--------QISYVKDEIKDLKAGEYAKISVKITDGTEKKYDDTLLKELDSNAVVTKVGTATTVKIVDQ--DGNSIEN 149
005/054 ----------LQDGTNDKYTVSNTKASDLVKDILAAQNLTTGAVILNKDTKVTFYDANE--KDSSTPT 87
010 AGTNSLNIINITVKDSDDDSVVVPKTVVKTNTVVDTDYVENKLKNLDTDEYVDFDITTKASTVSSTKYTAANIYDLAANN 160
012 NK----------KADRDAAAEKLYNLVNTQLDKLGDGDYVDFSVDYNLENKIITN--QADAEAIVTKL 131
046/092 NK----------KADRDAAAEKLYNLVNTQLDKLGDGDYVDFSVDYNLEKKIITN--QADAEAIVTKL 131
014/066 ----------SDKDAIAEQLETLTKNQLKGLGDGKYVDFKITYGAKAEVPAASLSADDIQKYADQ 132
017 AK----------KADRDAAAEKLYNLVNTQLDKLGDGDYVDFEVTYNLATQIIT---KAEAEAVLTKL 130
031/094 ----------DAAA-AATKLKSLVDDKLDNLGDGKYVQFNVTYTTKSIITKA--ELKNYYNQLES 128
078 TKG----------KTSTKEELASALLKTTVNDKLDNLGDGDYVDFDITYVGDADRLT----------AGDL 130
170 180 190 200 210 220 230 240
001 ASADAIIAGTSSADGVVTNTGAASGSTETNSAGTKLAMSAIFDTAYTDSSETAVKITIKADMNDTKFGKAG----ETTYS 208
002 DAKLKLQKYDTTTSAYVDGATKDEVLAATKAAVKTSKFTDKTGNNATGTVAITATY----------ADAVAGAIKLKYDT 219
005 GDKKVYSEQTLTTANGNEDYVKTTLKNLDAGEYAIIDLTYNN---------- 129
054 GDKKVYSEQILTTANGNEDYVKTTLKNLDAGEYAIIDLTYNN---------- 129
010 GKIIIDALKNVTEKQLKDGGILGEVAKTISGATTPSAPTGDTFASYFTVGTVKTVNGKVALEINIAEPASTVLVKTDAEL 240
012/046/092 NSLNEKTLIDIATKDTFGMVSKTQDSEGKNVAATKALKVKDVATFGLKSGGSEDTGYVVEMKAGAVEDKYG---KVGDST 208
014 INASEKILVEVAAGSEAGIAKFDSVNNKVIAGDAPLKVKDAVKATVTTNGSNKKVLTISAAAGLSGFSYGTLKDTGASLS 212
066 INASEKILVEVAAGSEAGIAKFDSVNNKVIAGDAPLKVKDAVKATVTTNGSNKKVLTISAAAGLSGFSYGTLKDTGASSS 212
017 QQYNDKVLINSATDTVKGMVSDTQVDS-KNVAANPLKVSDMYTIPSAITGSDDSGYSIAKPTEKTTSLLYG---TVGDAT 206
031 SKDRILIGNEPQDTGTKGLIKADTDGTTAVAAAAPLKLSDIFTFSYDEVTGVLKAEPTSKVSAGKVQGLKYGNTGATNYT 208
094 SKDRILIGNEPQDTGTKGLIKADTDGTTAVAADAPLKLSDIFTFSYDEVTGVLKAEPTSKVSAGKVQGLKYGNTGATNYT 208
078 NTFAKGIADSTEKKIPAAKGSNYGVAKTNSGTGKLTTDTEAVISTSIEGKVEGNNLTISLKDAPSKVGVIG----A-NND 205
250 260 270 280 290 300 310 320
001 TGLTFEDGSTEKIVKLGDSDIIDITKALKLTVVPGS----KATVKFAEKTPSASVQPVITKLRIINAKEETIDIDASSSK 284
002 QAAVAPTLDTAVASLDGTKEIYDFSKPVITVNTTDG------TSKLSFEKTGEKVGAVESNVTVIAASDETVTISGDAKE 293
005/054 ----------AKTVEIKVVAASEKTVVVSSDAKN 153
010 TTSPTTQQKMSFANAKITLTEGDDRLDFSKPSIVDG------ALGDFAKAAATTTPGKQQTINVRVINAKQETVKATDYD 314
012/046/092 AGIAINLPSTG-LEYAGKGTTIDFNKTLKVDVTGGSTPSAVAVSGFVTKDDTDLAKSGTINVRVINAKEESIDIDASSYT 287
014/066 DVDAITLDTTNATITEGDTKVLDFDNSFKFNESTKKVG----SLVTPNTTNTPADPGTKTTVRVIKAVEKTIDVSSNSTT 288
017 AGKAITVDTASNEAFAGNGKVIDYNKSFKATVQG---DGTVKTSGVVLKDASDMAATGTIKVRVTSAKEESIDVDSSSYI 283
031/094 SGAEISVPTTGLTLTADTTATTDVNISDVMSAFKFN---GTDTISGFP--AGSSASTLRASIKVINAKEESIDVDSSSHR 283
078 TLADVTFADDAKLTVSVGDPKIDLAKSFIFDTKTG--------K-LGGIVEKENDATEHAYVRVINAKEQTIDLDASSYK 276
330 340 350 360 370 380 390 400
001 TAQDLAKKYVFNKTDLNTLYKVLNGD---EADTNGLIEEVSGKYQVVLYPEGKRVTTKS|AA-K-------ASIADENSPV 353
002 KAEALAKKYVFKDTELEDAYKTVTASDFEKTDN--------DYYEVVLYPTGKRLNTASTYA|SSNYKPELPTSDRVDTPA 365
005/054 SAKDIAEKYVFEDKDLENALKTINASDFSKTD---------SYYQVVLYPKGKRLQGFSTYR|ATNYNEGTAYGN---TPV 221
010 ALKAVVSKYKFDSTEIGRVYDEAKDLNDKDKLDGSSYDK-DGTYKAVFFAEGKRLQGFSSYG|KFSTDEAESGLADGNAAL 393
012/046/092 SAENLAKRYVFDPDEISEAYKAIVALQN-DGIESNLVQLVNGKYQVIFYPEGKRLETKS|AN---------DTIASQDTPA 357
014/066 KAKDLAKQYVFTDVSDTDPESLSYMLKN-INDGKVAVKNSDGDYEVTIFPEGKRLNTLS|ASSAK------TILGDKDTPA 361
017 SAENLAKKYVFNPKEVSEAYNAIVALQN-DGIESDLVQLVNGKYQVIFYPEGKRLETKS|A----------DIIADADSPA 352
031/094 TAEDLAEKYVFKPEDVNKTYEALTDLYK-EGITSNLITQDGGKYQVVLFAQGKRLTTKG|AT---------GTLADENSPL 353
078 SAEDLAKAYAFDVNELKTLYTEIEAYQKDSNNKTDKVQIVDGKYQTILYAEGKRLTTKS|AT-QA------SKLADENSPL 349
410 420 430 440 450 460 470 480
001 KLTLKSDKKDLKDYVDDLRTYNNGYSNAIEVAGEDRIETAIALSQKYYN-SDDEN---AIFRDSVDNVVLVGGNAIVDG 429
002 IITLRSTNKNNLKSALDELRTYNNSYSNNSVLAGDDRIETAIEISKDSYNADGG----IKGDYVEANEVVLVGSQSIVDG 441
005 ILTLKSTSKSNLKTAVEELQKLNASYSNTTTLAGDDRIQTAIEISKEYYNNDGEKSDHSADVKENVKNVVLVGANALVDG 301
054 ILTLKSTSKSNLKTAVEELQKLNASYSNTTTLAGDDRIQTAIEISKEYYNNDGEKQDHSADVKENVKNVVLVGANALVDG 301
010 KLVIESTDEDDFIDGLKDLKELNNSYSDVESVAGDDRIETAIELSKSYYNNSRS-NNSDTLYTGAVDNVVLVGSQAIVDG 437
012/046/092 KVVIKANKLKDLKDYVDDLKTYNNTYSNVVTVAGEDRIETAIELSSKYYN-SDDKN---AITDKAVNDIVLVGSTSIVDG 433
014/066 KIVLKASTTKKLADYIDDLITYNNSYSNVQTVAGSDRIETAIELSRKYYN-STDKN---ALYGDPVNNVVLVGSQAIVDG 437
017 KITIKANKLKDLKDYVDDLKTYNNTYSNVVTVAGEDRIETAIELSSKYYN-SDDKN---AITDDAVNNIVLVGSTSIVDG 428
031/094 KVTIKADKVKDLKDYVEDLKNANNGYSNSVVVAGEDRIETAIELSSKYYN-SDDDN---AITKDPVNNVVLVGSQAVVDG 429
078 KLVIKADKLKDLKDYVEDLRNANNGYSNTVTVAGDDRIETAIELSSKYYNNSDEDN---AITEDAVNNVVLVGSQAIVDG 426



Consensus pattern, main motif, three repeats (* ** * :):

```
G-[ADES]-D-R-x-[ADEQ]-T-[ANS]-x(2)-[ILV]
```

Consensus pattern, minor motif, two repeats (* : ** *:* .:):

```
V-x-[IL]-[AIV]-G-G-x-[ANS]-S-[IV]-S-x-[DEK]-[IMV]
```

predicted from comparison with published N-terminal amino acid sequence of the mature HMW peptide (Calabi *et al.*, 2001; Cerquetti *et al.*, 2000). For previously unpublished sequences, prediction was based on comparison with published sequences (Fig. 2a). Cleavage is generally predicted to occur N-terminal to an alanine or serine residue and C-terminal to a consensus motif TKS or TYX. Cleavage might actually occur some way upstream of this site, with some residues lost from the N terminus of the peptide during maturation. An absolutely conserved GKR motif

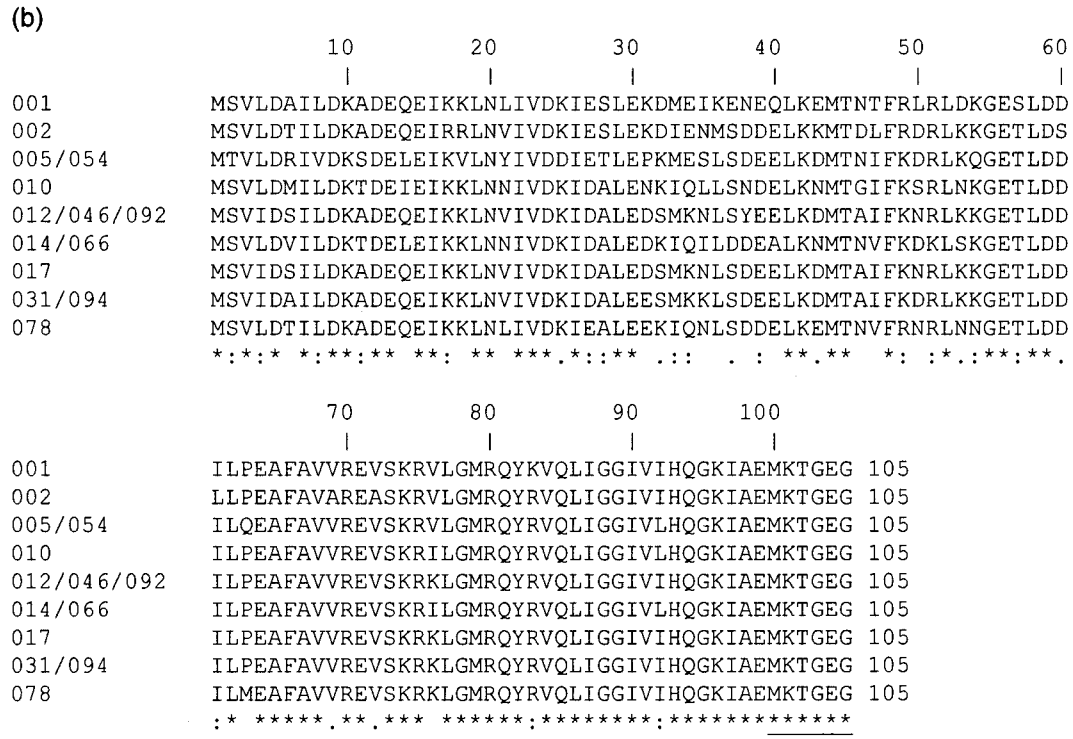


Fig. 2. CLUSTAL alignments of the translated ORF for *slpA* (a) and the translated partial ORF for *secA* homologue (b) from different ribotypes. Ribotype numbers are given on the left, and residue numbers on the right. Sequences for ribotypes 046 and 092 are identical and ribotype pairs 005 and 054, 012 and 046/092, 014 and 066 and 031 and 094 are only slightly different. Residue differences between respective pairs are underlined and the sequence is duplicated only where differences occur along a line (pairs are indicated by vertical bars to the left of alignments). Arrowheads and vertical bars indicate predicted cleavage sites (signal sequence C-terminal of residue 24, cleavage into LMW and HMW peptides at approximately 380) and references are given where cleavage sites have been determined experimentally from N-terminal amino acid sequencing for an identical or nearly identical sequence. Asterisks indicate identical residues, colons indicate strongly similar residues and dots indicate weakly similar residues. Heavy underlining indicates the main (threefold) repeated motif and double underlining indicates the minor (twofold) repeated motif of SlpA (see text for details). The C-terminal six residues of SecA, encoded by the cloning primer, are underlined.

occurs close to the predicted C termini of all the LMW peptides, which are otherwise generally dissimilar (Fig. 2a). The GKR motif is generally followed by a leucine residue, or occasionally valine, and is preceded by a conserved tyrosine at position −8, an aromatic residue at −3 and hydrophilic and hydrophobic residues at positions −7 and −4, respectively. The pattern appears in all other published sequences, although, presumably due to the heterogeneity of the LMW peptide, computer-generated alignments do not always recognize it (Calabi & Fairweather, 2002; Karjalainen *et al.*, 2002).

Internal repeats in the HMW peptide

The RADAR program identified an 11 residue motif in the HMW peptide, beginning 40–50 residues from the predicted N terminus and repeated twice at approximately 100–120 residue intervals (Fig. 2a). Four positions were absolutely conserved in all three repeats and a fifth position contained strongly similar residues. Another motif, repeated once, was found N-terminal of the second and third repeats. Both types of repeat occurred in regions of the sequence which were strongly conserved between ribotypes, so that a given repeat showed more similarity between ribotypes than to other repeats within a ribotype. For example, the unvarying DR of the principal motif was followed by the dissimilar residues isoleucine, tyrosine and glutamine in repeats 1, 2 and 3, respectively.

Molecular mass and isoelectric point prediction for the mature peptides

Fig. 3 shows an SDS-PAGE gel of crude SlpA preparations from all ribotypes. SlpA is seen as two dominant bands for all but ribotypes 005 and 054, where a single strong band was observed at approximately 45 000. Calabi & Fairweather (2002) report a similar absence of the second band for strain 167, which is ribotype 016 (Neil Fairweather, personal communication) and has an almost identical *slpA* sequence to that of ribotype 054. As expected, migration profiles of SlpA

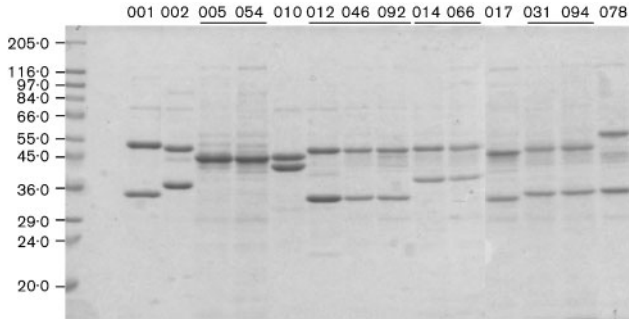


Fig. 3. SDS-PAGE profiles of crude SlpA preparations from different ribotypes of *C. difficile*. Preparations from ribotypes with similar or identical *slpA* sequences are in adjacent lanes (underlined). Relative molecular mass markers and their positions ($\times 10^3$) are shown on the left.

peptides were identical for ribotypes with identical or nearly identical *slpA* sequences. Relative molecular mass (exclusive of glycosylation) and pI were calculated using ExPASy software (Supplementary Table S1 available in JMM Online). The SlpA peptides tended to migrate more slowly than predicted, probably due to the high content of acidic amino acid residues or to post-translational modification. The pI range for the LMW peptides was within the typical range (4–6) for bacterial surface layer proteins (Sleytr *et al.*, 1999) and was consistently more acidic for the HMW peptide (4·46–4·69) than for the LMW peptide (4·83–5·09).

Amino acid composition

There was broad similarity when amino acid composition of the mature SlpA peptides was compared between ribotypes. Supplementary Fig. S1(a) (available in JMM Online) shows a comparison of the average composition of the predicted mature peptides, the translated *slpA* ORF and the global average of a compilation of 148 *C. difficile* coding sequences (http://www.kazusa.or.jp/codon/). Typical features of bacterial surface layer proteins were present (Pum *et al.*, 2000; Sára & Sleytr, 2000; Sleytr *et al.*, 1999), i.e. high content of acidic amino acids, no cysteine and very little methionine, low arginine content and little or no histidine. Lysine, alanine, aspartic acid and valine were the most abundant amino acids, and the last three were present at strikingly high levels compared with the global average. The aromatic amino acid content was generally low. Among functionally similar groups of amino acids, there was a general preference for the same residue in both peptides, e.g. both peptides showed a strong preference for aspartate over glutamate and for tyrosine over phenylalanine, in contrast to the global average. A striking exception was threonine, which was much more abundant in the LMW peptide. Supplementary Fig. S1(b) shows a comparison for groups of functionally similar amino acids. The HMW peptide had a higher content of acidic and hydrophobic amino acids and a lower content of aromatic amino acids. The LMW peptide from ribotypes 005 and 054, although much smaller than that from the other ribotypes (180 residues compared to $\geqslant$311 residues), contained numbers of asparagines close to the mean for all ribotypes, possibly indicating that a minimum number of these residues is essential to maintain the structural integrity of the protein.

Codon usage

Codon usage was analysed for the *slpA* genes of all ribotypes and for the *slpA* segments encoding predicted mature peptides. Codon usage patterns were compared with those of the same compilation of 148 sequences used for comparison of amino acid composition (Supplementary Fig. S2a). The species bias in favour of codons containing minimal G and C was generally stronger for the *slpA* gene, which is predicted to be highly expressed. Where codons permitted a choice of A or T in the wobble position, there was an occasional variation in preference. However, the combined frequencies of synonymous codons with a higher AT content generally well exceeded those of codons with a higher GC content.

A comparison of codon usage between the 5′ and 3′ regions of *slpA*, encoding the LMW and HMW peptides, respectively (Supplementary Fig. S2b), showed some differences in usage patterns for several amino acids, while maintaining the general bias in favour of A or T in the wobble position. Thus GCA was strongly favoured over GCT for alanine in the LMW peptide, with the reverse being true of the HMW peptide; and GTA strongly favoured over GTT for valine in the LMW peptide, with the reverse being true for the HMW peptide. Analogous differences occurred for serine and leucine, which have six codons each, while maintaining the AT-rich bias. Phenylalanine was an exception, with TTC preferred to TTT for phenylalanine in the HMW peptide, though not the LMW peptide.

Flanking DNA

In strain 630, *slpA* is flanked upstream by 213 nt of non-coding DNA preceded by an *slpA* paralogue and downstream by 247 nt of non-coding DNA followed by the *secA* homologue (Calabi *et al.*, 2001). In our sequences, 106–107 nt of upstream sequence is available for most ribotypes and was found to be identical for ribotypes 012, 046, 092, 017 and 078, which differed by a single nucleotide from ribotypes 031 and 094 and showed greater divergence from ribotypes 001 and 005/054 (Fig. 4a). These ribotypes shared a 22 nt stretch a short distance upstream of the *slpA* start codon, from which it was possible to design a primer to amplify the relevant sequence from the remaining ribotypes. This shared sequence included a polypurine stretch at positions −9 to −16 (−8 to −15 for ribotypes 005 and 054), containing the motif GGGAGG, strongly suggestive of a ribosome-binding site (Shine–Dalgarno box) in composition and location. No rho-independent terminators were identified in any of the upstream sequences, either the 107 nt available from the fragments sequenced from 10 ribotypes or the 213 nt of

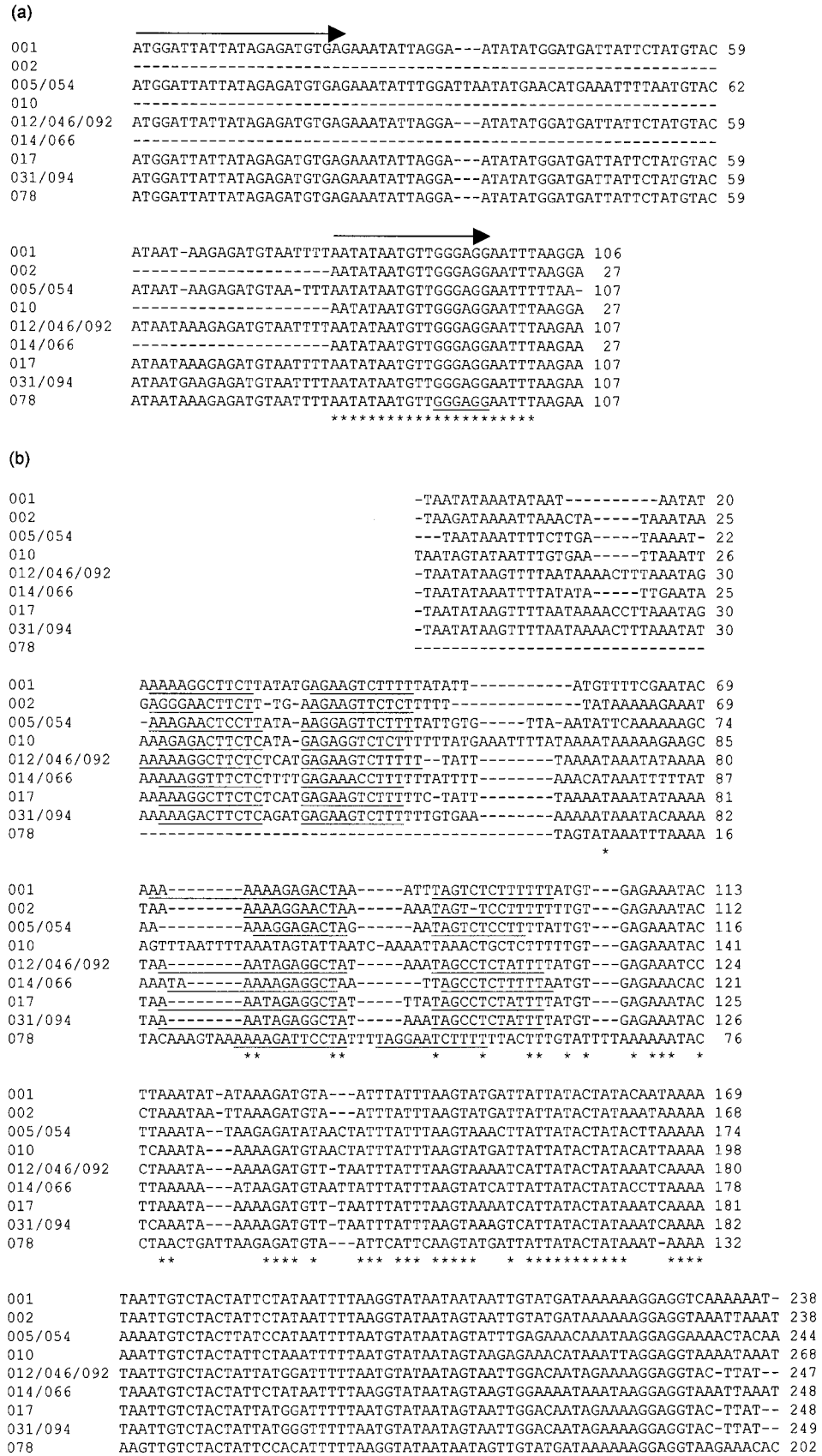

(a)
001 ATGGATTATTATAGAGATGTGAGAAATATTAGGA---ATATATGGATGATTATTCTATGTAC 59
002 ---------- 59
005/054 ATGGATTATTATAGAGATGTGAGAAATATTTGGATTAATATGAACATGAAATTTTAATGTAC 62
010 ---------- 62
012/046/092 ATGGATTATTATAGAGATGTGAGAAATATTAGGA---ATATATGGATGATTATTCTATGTAC 59
014/066 ---------- 59
017 ATGGATTATTATAGAGATGTGAGAAATATTAGGA---ATATATGGATGATTATTCTATGTAC 59
031/094 ATGGATTATTATAGAGATGTGAGAAATATTAGGA---ATATATGGATGATTATTCTATGTAC 59
078 ATGGATTATTATAGAGATGTGAGAAATATTAGGA---ATATATGGATGATTATTCTATGTAC 59
001 ATAAT-AAGAGATGTAATTTTAATATAATGTTGGGAGGAATTTAAGGA 106
002 ---------- 27
005/054 ATAAT-AAGAGATGTAA-TTTAATATAATGTTGGGAGGAATTTTTAA- 107
010 ---------- 27
012/046/092 ATAATAAAGAGATGTAATTTTAATATAATGTTGGGAGGAATTTAAGAA 107
014/066 ---------- 27
017 ATAATAAAGAGATGTAATTTTAATATAATGTTGGGAGGAATTTAAGAA 107
031/094 ATAATGAAGAGATGTAATTTTAATATAATGTTGGGAGGAATTTAAGAA 107
078 ATAATAAAGAGATGTAATTTTAATATAATGTTGGGAGGAATTTAAGAA 107
(b)
001 -TAATATAAATATAAT----------AATAT 20
002 -TAAGATAAAATTAAACTA-----TAAATAA 25
005/054 ---TAATAAATTTTCTTGA-----TAAAAT- 22
010 TAATAGTATAATTTGTGAA-----TTAAATT 26
012/046/092 -TAATATAAGTTTTAATAAAACTTTAAATAG 30
014/066 -TAATATAAATTTTATATA-----TTGAATA 25
017 -TAATATAAGTTTTAATAAAACCTTAAATAG 30
031/094 -TAATATAAGTTTTAATAAAACTTTAAATAT 30
078 ---------- 30
001 AAAAAGGCTTCTTATATGAGAAGTCTTTTTATATT----------ATGTTTTCGAATAC 69
002 GAGGGAACTTCTT-TG-AAGAAGTTCTCTTTTT----------TATAAAAAGAAAT 69
005/054 -AAAGAACTCCTTATA-AAGGAGTTCTTTTATTGTG-----TTA-AATATTCAAAAAAGC 74
010 AAAGAGACTTCTCATA-GAGAGGTCTCTTTTTTATGAAATTTTATAAAATAAAAAGAAGC 85
012/046/092 AAAAAGGCTTCTCTCATGAGAAGTCTTTTT--TATT--------TAAAATAAATATAAAA 80
014/066 AAAAAGGTTTCTCTTTTGAGAAACCTTTTTTATTTT--------AAACATAAATTTTTAT 87
017 AAAAAGGCTTCTCTCATGAGAAGTCTTTTTC-TATT--------TAAAATAAATATAAAA 81
031/094 AAAAAGACTTCTCAGATGAGAAGTCTTTTTTGTGAA--------AAAAATAAATACAAAA 82
078 ----------TAGTATAAATTTAAAA 16
001 AAA--------AAAAGAGACTAA-----ATTTAGTCTCTTTTTTATGT---GAGAAATAC 113
002 TAA--------AAAAGGAACTAA-----AAATAGT-TCCTTTTTTTGT---GAGAAATAC 112
005/054 AA---------AAAGGAGACTAG------AATAGTCTCCTTTTATTGT---GAGAAATAC 116
010 AGTTTAATTTTAAATAGTATTAATC-AAAATTAAACTGCTCTTTTTGT---GAGAAATAC 141
012/046/092 TAA--------AATAGAGGCTAT-----AAATAGCCTCTATTTTATGT---GAGAAATCC 124
014/066 AAATA------AAAAGAGGCTAA-------TTAGCCTCTTTTTAATGT---GAGAAACAC 121
017 TAA--------AATAGAGGCTAT-----TTATAGCCTCTATTTTATGT---GAGAAATAC 125
031/094 TAA--------AATAGAGGCTAT-----AAATAGCCTCTATTTTATGT---GAGAAATAC 126
078 TACAAAGTAAAAAAGATTCCTATTTTAGGAATCTTTTTTACTTTGTATTTTAAAAAATAC 76
001 TTAAATAT-ATAAAGATGTA---ATTTATTTAAGTATGATTATTATACTATACAATAAAA 169
002 CTAAATAA-TTAAAGATGTA---ATTTATTTAAGTATGATTATTATACTATAAATAAAAA 168
005/054 TTAAATA--TAAGAGATATAACTATTTATTTAAGTAAACTTATTATACTATACTTAAAAA 174
010 TCAAATA---AAAAGATGTAACTATTTATTTAAGTATGATTATTATACTATACATTAAAA 198
012/046/092 CTAAATA---AAAAGATGTT-TAATTTATTTAAGTAAAATCATTATACTATAAATCAAAA 180
014/066 TTAAAAA---ATAAGATGTAATTATTTATTTAAGTATCATTATTATACTATACCTTAAAA 178
017 TTAAATA---AAAAGATGTT-TAATTTATTTAAGTAAAATCATTATACTATAAATCAAAA 181
031/094 TCAAATA---AAAAGATGTT-TAATTTATTTAAGTAAAGTCATTATACTATAAATCAAAA 182
078 CTAACTGATTAAGAGATGTA---ATTCATTCAAGTATGATTATTATACTATAAAT-AAAA 132
001 TAATTGTCTACTATTCTATAATTTTAAGGTATAATAATAATTGTATGATAAAAAAGGAGGTCAAAAAAT- 238
002 TAATTGTCTACTATTCTATAATTTTAAGGTATAATAGTAATTGTATGATAAAAAAGGAGGTAAATTAAAT 238
005/054 AAAATGTCTACTTATCCATAATTTTAATGTATAATAGTATTTGAGAAACAAATAAGGAGGAAAACTACAA 244
010 AAATTGTCTACTATTCTAAATTTTTAATGTATAATAGTAAGAGAAACATAAATTAGGAGGTAAAATAAAT 268
012/046/092 TAATTGTCTACTATTATGGATTTTTAATGTATAATAGTAATTGGACAATAGAAAAGGAGGTAC-TTAT-- 247
014/066 TAAATGTCTACTATTCTATAATTTTAAGGTATAATAGTAAGTGGAAAATAAATAAGGAGGTAAATTAAAT 248
017 TAATTGTCTACTATTATGGATTTTTAATGTATAATAGTAATTGGACAATAGAAAAGGAGGTAC-TTAT-- 248
031/094 TAATTGTCTACTATTATGGGTTTTTAATGTATAATAGTAATTGGACAATAGAAAAGGAGGTAC-TTAT-- 249
078 AAGTTGTCTACTATTCCACATTTTTAAGGTATAATAATAGTTGTATGATAAAAAAGGAGGTAAGAAACAC 202

Fig. 4. (a) Sequence immediately upstream from *slpA*. The left-hand primers used to amplify the DNA are shown by arrows, to indicate where the sequence information may not be completely accurate, and the predicted Shine–Dalgarno box for *slpA* is underlined. (b) Intergenic region between *slpA* and *secA* genes. The positions of possible rho-independent terminators for the *slpA* gene are underlined, along with the predicted Shine–Dalgarno box for the *secA* gene.

intergenic DNA from strain 630 (ribotype 012) known from the Sanger sequencing project. It is tentatively inferred that transcription of the upstream *slpA* paralogue terminates by a rho-mediated process. Rho-dependent terminators are not readily identifiable in sequences (Henkin, 1996).

The distance between the *slpA* and *secA* genes varied from 202 nt (for ribotype 078) to 268 nt (for ribotype 010). Fig. 4(b) shows an alignment of the sequences. The *slpA* gene terminated in TAA for all ribotypes except 078, where it ended in TAG. The first half of the sequence was relatively unconserved, especially for ribotypes 010 and 078, and was considerably shorter in the latter. Sequences encoding potential mRNA stem and loop structures, reminiscent of rho-independent terminators, were found in this region. Rho-independent terminators show similarity between widely distributed bacterial genera (Vermat *et al.*, 2002). Potential regions of dyad symmetry were 10–13 bp, with an unpaired loop of 3–4 nt, and typically contained three or four GC pairs, which would stabilize the structure of the transcript. Two such structures were found in all ribotypes except 010 and 078, which had one each. In rho-independent terminators, the region of dyad symmetry commonly gives way to a 3′ run of non-pairing Ts, which is predicted to facilitate the release of the transcript, and the consensus sequence TCTG (Brendel & Trifonov, 1984). The terminators predicted for *slpA* were generally found to contain either the T-trail, the consensus sequence TATG/TGTG or occasionally both. A region of quite strong conservation was identified (59 % identity) from approximately nt −110 to nt −8 to −11 upstream of *secA*, ending in a very likely Shine–Dalgarno box for *secA* (AGGAGG).

Phylogenetic analysis

Evolutionary relationships between ribotypes based on the *slpA* genes were examined by aligning the HMW peptides using CLUSTAL W and constructing neighbour-joining trees (Fig. 5a). Alignment gaps were omitted from the phylogenetic analysis. The LMW peptide was not included in the analysis, as there is very little overall sequence conservation between ribotypes in this region and two of the ribotypes contain a much shorter peptide than the others. The same analysis was carried out with the available N-terminal 105 residues of sequence from the SecA homologue (Fig. 5b), so that variation in SlpA, presumed to be an antigen under evolutionary pressure to diversify, could be compared with variation in SecA, an essential protein which is not surface exposed.

The trees suggest different evolutionary histories for *slpA* and *secA*, consistent with either recombination and/or positive selection. Recombination would produce new antigenic variants that could enable a strain to evade host defences, which could undergo positive selection. In particular, both trees show a robust clade (indicated by an arrow) containing ribotypes 012, 046/092, 017, 031 and 094. However, ribotypes 001 and 078 (indicated by asterisks) are also contained within this clade for the SlpA tree. This is consistent with recombination between lineages on the *slpA* gene. There are other differences between the trees, but these are less well supported statistically. Furthermore, the *secA* clade containing ribotypes 012, 046/092, 017, 031 and 094 exhibits considerably shorter branch lengths than the corresponding *slpA* clade, consistent with the housekeeping role of SecA.

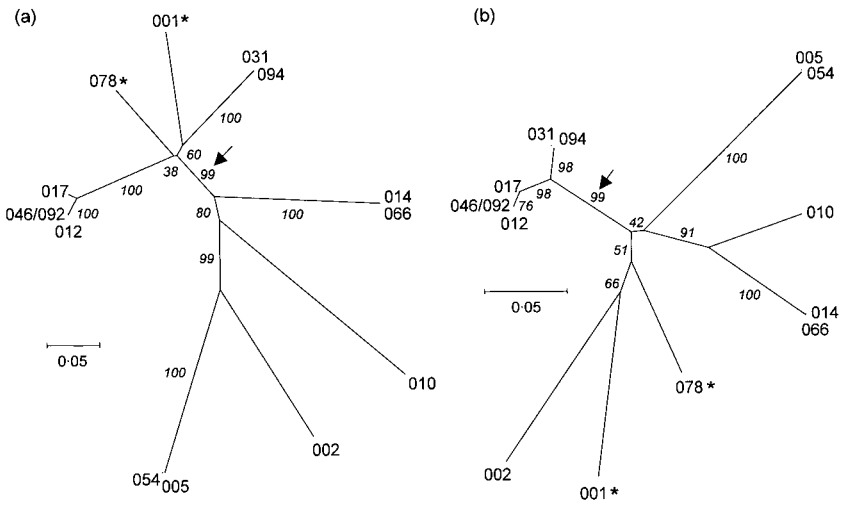


Fig. 5. Neighbour-joining trees for SlpA (a) (aligned sequence from HMW peptide) and SecA (b) (N-terminal 105 residues from SecA2 homologue). Ribotypes 031, 094, 017, 012 and 046/092 are on a robust clade for each tree (left of arrows). Note that ribotypes 001 and 078 (asterisks) are included in this clade for SlpA but not for SecA. Numbers along branches represent bootstrap support (10 000 replicates) for each branch.

DISCUSSION

We have sequenced the *slpA* gene and flanking DNA from *C. difficile* ribotypes isolated from patients in St James's Hospital, Dublin over a 16-month period. The most frequently occurring ribotypes were 001, 012 and 017, and *slpA* had already been sequenced from these. However, for the first time, we report complete DNA sequences for *slpA* from strains formally assigned to ribotypes 002, 005, 010, 014, 031, 046, 054, 066, 078, 092 and 094. DNA sequence obtained from two or three isolates of each of the more common ribotypes was always identical, indicative of clonal spread. We also provide information on flanking DNA, including the 5′ 315 nt from a putative *secA* gene. The availability of sequence from an adjacent conserved gene provides reassurance that a single amplicon has been sequenced in all strains. Given the existence of numerous paralogues of *slpA* in the genome, this is not a trivial consideration.

We show a strong relatedness between the *slpA* sequences from ribotypes 005, 016 and 054, between ribotypes 012, 046 and 092, between ribotypes 014 and 066 and between ribotypes 031 and 094. It is useful to have this sequence information on *slpA*, which is strongly related to serogroup designation and complements ribotyping, which is based on non-coding DNA that is subject to different evolutionary pressure. Since there are more ribotypes than serogroups, it is not surprising that different ribotypes should have similar serogrouping antigens. However, the sequence data we have acquired indicate how very alike some *slpA* sequences from different ribotypes are, at both the nucleotide and amino acid sequence levels. This near-identity occurs alongside important strain differences, such as toxin production. Thus ribotype 031, which is deficient in both A and B toxins, has an *slpA* with one nucleotide difference compared with ribotype 094, which produces both toxins. Similarly, *slpA*s from ribotypes 066 (non-toxigenic) and 014 (toxin A^+ and B^+) differ by a single nucleotide. The sequence from ribotype 002 strains (equivalent serogroup A2 according to Stubbs *et al.*, 1999) does not appear to resemble closely that published for any other strain.

An international effort is being made to coordinate typing methods in order to correlate the dominant types in outbreaks of *C. difficile* disease around the world. Serogrouping, which is largely based on differences between SlpA variants, is complicated by the existence of flagellated strains, since flagellar antigens may cross-react and production of flagella can vary with culture age and conditions. Serogroup A strains share serologically cross-reactive flagellins and subgroups were originally distinguished by SDS-PAGE profiles of major peptides, probably SlpA, in whole-cell preparations (Delmée *et al.*, 1986). The role of flagellar antigens was appreciated later (Delmée *et al.*, 1990) and it became possible to distinguish the A subgroups by slide-agglutination on removal of the flagella by sonicating the bacteria. However, it was also noted that strains of some other serogroups were flagellated and that cross-reactive flagellins occurred widely. Although it was claimed that non-serogroup-A strains had fewer flagella, and did not cross-react with group A strains in slide-agglutination, it is conceivable that mistakes could occur with heavily flagellated non-group-A strains.

Although some of the present work highlights the diversity of SlpA in different ribotypes and serogroups, some features appear consistently, notably the conserved stretch of sequence in the leader peptide, the GKRL/V motif and adjacent conserved residues in the vicinity of the secondary cleavage site and the five blocks of conserved sequence in the HMW peptide, interspersed with stretches of variable length and sequence. We also detected the DR-containing repeat motif, originally identified by Calabi & Fairweather (2002) in a comparison of SlpA molecules from six strains, in all of our sequences.

The GKRL/V motif is intriguing, given that GKR represents the longest stretch of absolutely conserved sequence in the LMW peptide and occurs near the precursor cleavage site. A considerable number of proteins from eukaryotes and viruses are cleaved post-translationally C-terminal to a pair of basic residues. These include mammalian growth factors and receptor proteins and glycoproteins from HIV and varicella zoster virus (reviewed by Seidah & Chrétien, 1999). Some bacterial toxins are activated by cleavage at pairs of basic residues by host proteases, but cleavage of bacterial proteins by bacterial proteases at such sites has not been widely reported. It is possible that the conservation of the site we have identified has more to do with binding of the relevant protease or transient association with the plasma membrane to facilitate cleavage at the distal sites already predicted experimentally.

There are several reports (Calabi *et al.*, 2001; Cerquetti *et al.*, 2000) that SlpA is glycosylated, probably more extensively on the HMW peptide, but the nature and extent of glycosylation are unknown. The widespread occurrence of glycosylation among bacterial proteins, including S-layers, has only recently been accepted and its potential role in virulence appreciated (reviewed by Schäffer & Messner, 2001; Schmidt *et al.*, 2003; Spiro, 2002). Clostridial S-layers were among the first bacterial glycoproteins to be described (Sleytr & Thorne, 1976). In bacteria, glycopeptide linkages may be *N*-glycosyl on asparagine residues or, more often, *O*-glycosyl on serine, threonine and tyrosine residues. The sequence and structural requirements for glycosylation sites seem less well defined in prokaryotes than in eukaryotes, making their prediction from primary sequence virtually impossible, at least for the moment (Schäffer & Messner, 2001). Although it is known that the toxins mediate their activity by glycosylation of a critical threonine residue on the host Rho family of GTPases (reviewed by Spiro, 2002), at least two toxin-negative strains are reported to have a glycosylated SlpA (Cerquetti *et al.*, 2000), implying the existence of a separate mechanism for SlpA glycosylation.

Some striking differences were noticed in amino acid composition between SlpA and the compilation published for

C. difficile proteins. Some of these, e.g. the unusually high content of acidic amino acids and the absence of cysteine, are features of S-layers in general. There is a certain bias in favour of amino acids with smaller side-groups where these are functionally similar, i.e. aspartate is favoured over glutamate, asparagine over glutamine and alanine and valine over other hydrophobic amino acids, possibly because of the need for economy or possibly because of constraints imposed by the crystalline structure. Conversely, some amino acids, although present at a low level, may be critically important. Thus arginine is rare in SlpA, but where it occurs it is often in a relatively conserved region. Of the seven to ten arginines found in the HMW peptide, four positions are absolutely conserved and three of these are associated with the repeated DR motif. In the LMW peptide, which has two to five arginines, one is always found in the GKR motif.

S-layers composed of two peptides have been described for a few other organisms, but the SlpA from *C. difficile* is the only reported incidence of a two-component S-layer derived from a single precursor peptide. The reason for this phenomenon is not clear. Although it is known that the LMW and HMW peptides have different lattice structures (Cerquetti *et al.*, 2000), detailed tertiary structure comparisons are not possible at this time. Secondary structure analyses (not shown) indicated that both peptides were composed predominantly of random coil (44–49 %) interspersed with alpha helical regions (28–34 %) and extended strand (19–22 %) and were not generally informative. We therefore compared the amino acid composition of the LMW and HMW peptides. This generally reflected the overall composition of the translated ORF, with the greatest difference noted for threonine, which was quite abundant in the LMW peptide at an average of just over 11 %, compared with 5·8 and 5·7 %, respectively, for the HMW peptide and the global average. A survey of published eubacterial S-layer sequences shows that a high threonine content (>9 %) is quite common, with the highest content (18·2 %) for *Caulobacter crescentus*, for reasons which have not been explained. The LMW peptide also had a higher content of aromatic amino acids, although SlpA had a rather low content of aromatic amino acids overall. We noted that the LMW peptide had two absolutely conserved tyrosines in close proximity (Fig. 2a), one absolutely conserved phenylalanine and one position which always contained either residue. It is tempting to suppose that the aromatic amino acids might be involved in binding of the LMW to a host ligand via a carbohydrate receptor or that they play an essential role in stabilizing the crystalline lattice structure.

Given the high rate of protein synthesis needed to maintain the integrity of the S-layer during growth and division (Sleytr & Messner, 1983), it is not surprising that codon usage in *slpA* reflects the strong AT-rich bias of *C. difficile*. The differences in codon usage between the LMW and HMW peptides (albeit within this bias) may reflect different evolutionary origins or perhaps an influence of DNA secondary structure.

A short stretch of presumably non-coding DNA upstream of *slpA* was fairly well conserved among 10 ribotypes, but differed sufficiently to prevent the use of the original left-hand primer in ribotypes 002, 010, 014 and 066. The intergenic region from *slpA* to *secA* varied substantially in sequence, particularly the moiety closest to the *slpA* gene. The fragment from ribotype 078, which had an upstream sequence similar or identical to those of nine other ribotypes, seems truncated in the region immediately downstream of *slpA*, lacking a predicted terminator and even employing a different termination codon. A recombination event in this region might explain why ribotype 078 is found on different clades in the bootstrap analyses of SlpA and SecA (Fig. 5). In ribotype 010, for which we have very little sequence upstream of *slpA*, the *slpA–secA* intergenic region, which is relatively long, also appears to encode a single terminator. This ribotype is found on a fairly strong clade with ribotypes 014 and 066 for SecA in the bootstrap analyses, but is not strongly associated with any other ribotype for SlpA. There is little variation in the DNA sequence either upstream or downstream from *slpA* in ribotypes 012, 046/092, 017, 031 and 094, and all of these are quite tightly linked on both trees. However, ribotype 001, which shows small to moderate differences in upstream and downstream flanking sequences, is part of this clade for SlpA, but not for SecA. A strain that acquires a novel surface layer protein by recombination, thereby generating a new strain, presumably has greater ability to evade a host's immune response. Toxin production, resulting in profuse diarrhoea, is an effective means of dispersing spores and spreading the disease. This may be the case with ribotype 001, which appears to have acquired a novel *slpA* gene by recombination with another clade and is responsible for approximately half of *C. difficile* infections in the UK (Brazier, 1998) and in St James's Hospital (Doyle, 2004).

A sequence similarity search of the published genome for *C. difficile* 630 has revealed the presence of a second *secA* allele located approximately 1 Mb from the gene flanking *slpA* and in the opposite orientation. Although uncommon, two *secA* alleles have been found in a number of genera, including *Mycobacterium* and *Listeria*, where the second SecA (SecA2) appears to have a specific role in the export of virulence-related proteins (Pallen *et al.*, 2003). The product of the gene flanking *slpA* in *C. difficile* 630 (genome sequence) more closely resembles the SecA2 of other species, being the smaller of the two proteins and lacking a region predicted to interact with the chaperone SecB (Fekkes *et al.*, 1997). Although it may play a subsidiary role, the predicted SecA2 does show strong similarity to its counterpart in other species, notably *Listeria monocytogenes* and *Streptococcus parasanguinis* (approx. 40 % identical residues), indicating a high degree of conservation between genera. Since *C. difficile* is not known to grow outside a human or mammalian host in nature, if *secA2* is required for colonization, it is de facto an essential gene. The location of a gene for a major transport protein immediately downstream from *slpA* is unlikely to be coincidental, and it is known that S-layer

expression in *Aeromonas salmonicida* depends on an ATP-dependent transport protein encoded by a downstream gene (Chu & Trust, 1993). Interestingly, in *Listeria monocytogenes*, SecA2 is believed to be responsible for the secretion of two autolysins, NamA and p60, and deletion of either *secA2* or either of the structural genes leads to accelerated clearance of the organism from spleens and livers of infected mice (Lenz *et al.*, 2003). The HMW peptide of *C. difficile* SlpA has *N*-acetylmuramoyl amidase activity, and *slpA* is located among a cluster of genes encoding similar domains (Calabi *et al.*, 2001). Although *C. difficile* autolysins are predicted to be involved in wall remodelling during cell growth and division (Calabi *et al.*, 2001), it is possible that they may have additional roles in virulence.

ACKNOWLEDGEMENTS

This work was supported by grants from Enterprise Ireland (ATRP-01/165) and from the Higher Education Authority. The authors wish to thank Ms Patricia O'Brien and the staff at the Microbiology Laboratory, St James's Hospital, for advice on *C. difficile* culture. Thanks are also due to Dr Denis Shields, Royal College of Surgeons in Ireland, for help with interpreting neighbour-joining trees and to Neil Whitehead, DNA sequencing laboratory, Biochemistry Dept, University of Cambridge, for providing excellent support and service with DNA sequencing. Thanks to Dr Jon Brazier of the Anaerobe Reference Laboratory at the HPA in Cardiff for PCR ribotyping of isolates. We are grateful to Drs Neil Fairweather and Emanuela Calabi, Imperial College, London, for sharing unpublished data on ribotyping. Thanks to Dr Julian Parkhill and colleagues of the *C. difficile* (Strain 630) Sequencing Group at the Sanger Institute (http://www.sanger.ac.uk/Projects/C_difficile/) for providing open access to the genome sequence and for permission to publish our *in silico* findings with respect to *secA* homologues.

REFERENCES

Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783–795.

Brazier, J. S. (1998). The epidemiology and typing of *Clostridium difficile*. *J Antimicrob Chemother* **41** (Suppl. C), 47–57.

Brazier, J. S. (2001). Typing of *Clostridium difficile*. *Clin Microbiol Infect* **7**, 428–431.

Brendel, V. & Trifonov, E. N. (1984). A computer algorithm for testing potential prokaryotic terminators. *Nucleic Acids Res* **12**, 4411–4427.

Calabi, E. & Fairweather, N. (2002). Patterns of sequence conservation in the S-layer proteins and related sequences in *Clostridium difficile*. *J Bacteriol* **184**, 3886–3897.

Calabi, E., Ward, S., Wren, B., Paxton, T., Panico, M., Morris, H., Dell, A., Dougan, G. & Fairweather, N. (2001). Molecular characterization of the surface layer proteins from *Clostridium difficile*. *Mol Microbiol* **40**, 1187–1199.

Cerquetti, M., Molinari, A., Sebastianelli, A., Diociaiuti, M., Petruzzelli, R., Capo, C. & Mastrantonio, P. (2000). Characterization of surface layer proteins from different *Clostridium difficile* clinical isolates. *Microb Pathog* **28**, 363–372.

Chu, S. & Trust, T. J. (1993). An *Aeromonas salmonicida* gene which influences A-protein expression in *Escherichia coli* encodes a protein containing an ATP-binding cassette and maps beside the surface array protein gene. *J Bacteriol* **175**, 3105–3114.

Combet, C., Blanchet, C., Geourjon, C. & Deleage, G. (2000). NPS@: network protein sequence analysis. *Trends Biochem Sci* **25**, 147–150.

Deléage, G. & Roux, B. (1987). An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng* **1**, 289–294.

Delmée, M., Laroche, Y., Avesani, V. & Cornelis, G. (1986). Comparison of serogrouping and polyacrylamide gel electrophoresis for typing *Clostridium difficile*. *J Clin Microbiol* **24**, 991–994.

Delmée, M., Avesani, V., Delferriere, N. & Burtonboy, G. (1990). Characterization of flagella of *Clostridium difficile* and their role in serogrouping reactions. *J Clin Microbiol* **28**, 2210–2214.

Doyle, R. M. (2004). *Humoral immune response to Clostridium difficile associated disease*. MD thesis, Trinity College Dublin, Ireland.

Fekkes, P., van der Does, C. & Driessen, A. J. (1997). The molecular chaperone SecB is released from the carboxy-terminus of SecA during initiation of precursor protein translocation. *EMBO J* **16**, 6105–6113.

Frishman, D. & Argos, P. (1996). Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng* **9**, 133–142.

Garnier, J., Gibrat, J. F. & Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol* **266**, 540–553.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. & Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* **31**, 3784–3788.

Geourjon, C. & Deléage, G. (1994). SOPM: a self-optimized method for protein secondary structure prediction. *Protein Eng* **7**, 157–164.

Guermeur, Y. (1997). *Combinaison de classifieurs statistiques, application à la prédiction de structure sécondaire des protéines*. PhD thesis, Université Paris 6, France (in French).

Heger, A. & Holm, L. (2000). Rapid automatic detection and alignment of repeats in protein sequences. *Proteins* **41**, 224–237.

Henkin, T. M. (1996). Control of transcription termination in prokaryotes. *Annu Rev Genet* **30**, 35–57.

Karjalainen, T., Waligora-Dupriet, A. J., Cerquetti, M., Spigaglia, P., Maggioni, A., Mauri, P. & Mastrantonio, P. (2001). Molecular and genomic analysis of genes encoding surface-anchored proteins from *Clostridium difficile*. *Infect Immun* **69**, 3442–3446.

Karjalainen, T., Saumier, N., Barc, M. C., Delmée, M. & Collignon, A. (2002). *Clostridium difficile* genotyping based on *slpA* variable region in S-layer gene sequence: an alternative to serotyping. *J Clin Microbiol* **40**, 2452–2458.

Kelly, C. P. & LaMont, J. T. (1998). *Clostridium difficile* infection. *Annu Rev Med* **49**, 375–390.

King, R. D. & Sternberg, M. J. E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci* **5**, 2298–2310.

Kumar, S., Tamura, K., Jakobsen, I.-B. & Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244–1245.

Kyne, L., Hamel, M. B., Polavaram, R. & Kelly, C. P. (2002). Health care costs and mortality associated with nosocomial diarrhea due to *Clostridium difficile*. *Clin Infect Dis* **34**, 346–353.

Lenz, L. L., Mohammadi, S., Geissler, A. & Portnoy, D. A. (2003). SecA2-dependent secretion of autolytic enzymes promotes *Listeria monocytogenes* pathogenesis. *Proc Natl Acad Sci U S A* **100**, 12432–12437.

Levin, J. M. (1997). Exploring the limits of nearest neighbour secondary structure prediction. *Protein Eng* **10**, 771–776.

O'Neill, G. L., Ogunsola, F. T., Brazier, J. S. & Duerden, B. I. (1996). Modification of a PCR ribotyping method for application as a routine typing scheme for *Clostridium difficile*. *Anaerobe* **2**, 205–209.

Pallen, M. J., Chaudhuri, R. R. & Henderson, I. R. (2003). Genomic analysis of secretion systems. *Curr Opin Microbiol* **6**, 519–527.

Pantosti, A., Cerquetti, M., Viti, F., Ortisi, G. & Mastrantonio, P. (1989). Immunoblot analysis of serum immunoglobulin G response to surface proteins of *Clostridium difficile* in patients with antibiotic-associated diarrhea. *J Clin Microbiol* **27**, 2594–2597.

Poxton, I. R., Higgins, P. G., Currie, C. G. & McCoubrey, J. (1999). Variation in the cell surface proteins of *Clostridium difficile*. *Anaerobe* **5**, 213–215.

Pum, D., Neubauer, A., Gyorvary, E., Sára, M. & Sleytr, U. B. (2000). S-layer proteins as basic building blocks in a biomolecular construction kit. *Nanotechnology* **11**, 100–107.

Rost, B. & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19**, 55–72.

Sára, M. & Sleytr, U. B. (2000). S-layer proteins. *J Bacteriol* **182**, 859–868.

Schäffer, C. & Messner, P. (2001). Glycobiology of surface layer proteins. *Biochimie* **83**, 591–599.

Schmidt, M. G. & Kiser, K. B. (1999). SecA: the ubiquitous component of preprotein translocase in prokaryotes. *Microbes Infect* **1**, 993–1004.

Schmidt, M. A., Riley, L. W. & Benz, I. (2003). Sweet new world: glycoproteins in bacterial pathogens. *Trends Microbiol* **11**, 554–561.

Seidah, N. G. & Chrétien, M. (1999). Proprotein and prohormone convertases: a family of subtilases generating diverse bioactive polypeptides. *Brain Res* **848**, 45–62.

Sleytr, U. B. & Messner, P. (1983). Crystalline surface layers on bacteria. *Annu Rev Microbiol* **37**, 311–339.

Sleytr, U. B. & Thorne, K. J. (1976). Chemical characterization of the regularly arranged surface layers of *Clostridium thermosaccharolyticum* and *Clostridium thermohydrosulfuricum*. *J Bacteriol* **126**, 377–383.

Sleytr, U. B., Messner, P., Pum, D. & Sára, M. (1999). Crystalline bacterial cell surface layers (S layers): From supramolecular cell structure to biomimetics and nanotechnology. *Angew Chem Int Ed Engl* **38**, 1035–1054.

Spiro, R. G. (2002). Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology* **12**, 43R–56R.

Stubbs, S. L., Brazier, J. S., O'Neill, G. L. & Duerden, B. I. (1999). PCR targeted to the 16S-23S rRNA gene intergenic spacer region of *Clostridium difficile* and construction of a library consisting of 116 different PCR ribotypes. *J Clin Microbiol* **37**, 461–463.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680.

van Wely, K. H., Swaving, J., Freudl, R. & Driessen, A. J. (2001). Translocation of proteins across the cell envelope of Gram-positive bacteria. *FEMS Microbiol Rev* **25**, 437–454.

Vermat, T., Vandenbrouck, Y., Viari, A. & d'Aubenton Carafa, Y. (2002). Prediction, distribution and evolution of intrinsic transcription terminators in bacterial genomes. In *JOBIM 2002*, pp. 137–142. Edited by J. Nicolas & C. Thermes. Rennes: IMPG.