Streptococcus pyogenes sclB encodes a putative hypervariable surface protein with a collagen-like repetitive structure

Adrian M. Whatmore

Streptococcus pyogenes is the causative agent in a wide range of diseases of humans of varying severity. During a study scanning the genome sequence of a serotype M1 invasive isolate SF370 for novel surface proteins, an ORF, designated sclB, was identified. The putative protein encoded by sclB contains both a signal peptide and classic Gram-positive wall-associated sequences. Comparison of the sequences of this ORF with those from a number of unrelated isolates demonstrated that sclB encodes a putative surface protein with a variable N-terminal sequence followed by a variable length tract of collagen-like GXY_n repeats. A further feature of sclB is the presence of CAAAA repeat tracts immediately downstream of the putative start codon. The number of these pentameric repeats varies from 4 to 15 between strains and variation in repeat number results in the predicted SclB protein being either in or out of frame relative to the start codon. These observations suggest that expression of this protein may be regulated at the translational level as a result of gain or loss of CAAAA repeats. While the function of SclB remains to be elucidated, an sclB-specific transcript was detected by RT-PCR during in vitro culture. Finally, it is shown that a second gene, sclA, potentially encoding a protein with a similar extensive collagen-like structure and variable N-terminal sequence, is present in all isolates of S. pyogenes tested to date. Thus S. pyogenes harbours a novel family of structurally related and surface-exposed proteins of potential importance in the pathogenic process.

Keywords: bacterial surface protein, genetic variation, short sequence repeats, tandem repeats

INTRODUCTION

The Group A streptococcus (Streptococcus pyogenes) remains an important human pathogen, associated with a range of superficial skin and throat infections as well as a variety of more serious invasive infections and autoimmune sequelae such as acute rheumatic fever and post-streptococcal glomerulonephritis (Cunningham, 2000). As is the case in other Gram-positive bacteria, the cell wall of S. pyogenes is associated with an array of proteins which can share common structural features (Navarre & Schneewind, 1999). Proteins destined for transport across the cytoplasmic membrane often contain a signal (leader peptide) composed of a core of hydrophobic residues and a charged N-terminal end (von Heijne, 1990). Signal peptides of Gram-positive surface proteins, which are located at the N-terminus of the protein, are proteolytically removed by signal peptidases on translocation across the cytoplasmic membrane (Dev & Ray, 1990). Many Gram-positive surface proteins also share striking C-terminal sequence similarities corresponding to cell wall sorting signals. These consist of a highly conserved LPXTG sequence motif followed by a hydrophobic domain and a charged tail of largely positively charged residues (Fischetti et al., 1990). The hydrophobic domain and charged tail are believed to be important in preventing the immediate release of the protein from the secretory pathway into the extracellular milieu. This retention is thought to allow recognition of the LPXTG motif and proteolytic cleavage between the

Abbreviation: SSR, short sequence repeat.
The GenBank accession numbers for the sequences determined in this work are given in the text and Table 1.
threonine and glycine residues, permitting cross-linking of the surface protein with peptidoglycan precursors (Navarre & Schneewind, 1999).

A number of proteins harbouring cell wall sorting signals have now been characterized from *S. pyogenes*. These include C5a peptidase (Chen & Cleary, 1990), the T protein (Schneewind et al., 1990), two distinct fibronectin-binding proteins – SfbI/proteinF (Sela et al., 1993; Talay et al., 1994) and SfbII/OF (Kreikemeyer et al., 1995; Rakonjac et al., 1995) – and a large number of proteins encoded by members of the *emm* gene family (reviewed by Navarre & Schneewind, 1999). As might be expected, many of the surface proteins characterized to date have been postulated to play important roles in virulence either by modulating the host immune response and/or by involvement in adhesion to host surfaces (Jenkinson & Lamont, 1997; Cunningham, 2000). Many, notably members of the *emm* gene family, possess multiple binding activities for an array of host proteins that can include albumin, fibrinogen, IgG, IgA, kininogen, fibronectin, factor H, FHL-1, plasminogen and C4BP (Navarre & Schneewind, 1999).

A further common feature of many surface proteins of Gram-positive bacteria is the presence of tandem repeat domains that can vary in size from just several to several hundred amino acids. Often strain-to-strain variations in the number of repeats are apparent, implying that gain or loss of repeats occurs as a result of intragenic recombination events or slipped-strand mispairing. Perhaps the best examples of such repetitive structures in *S. pyogenes* are members of the *emm* gene family, which can contain up to three distinct sets of tandem repeats (Kehoe, 1994). Analysis of successive isolates from patients demonstrates that variation in the anti-phagocytic *M6* protein occurs as a result of alteration in the number of tandem repeats (Hollingshead et al., 1987) and these changes can alter antigenic determinants, resulting in resistance to phagocytosis (Jones et al., 1988). Variable repetitive structure is also apparent in a number of surface proteins of other streptococci, including the *S. pneumoniae* surface protein PspA (Yother & Briles, 1992) and the alpha C protein encoding gene of *S. agalactiae* (Madoff et al., 1996). A number of possible benefits, which may be of relevance in pathogenesis, of variation in repeat number can be envisaged. These include antigenic variation and subsequent immune escape, alteration in substrate binding properties and the simple physical displacement of parts of proteins further away from the immediate extracellular environment where they may be masked by other surface components. In several Gram-negative bacteria so-called short sequence repeats (SSRs), consisting of 2–20 nucleotide tandem repeats, can be involved in gene regulation either by altering the spacing between promoter domains or by affecting the integrity of ORFs (van Belkum et al., 1998; Henderson et al., 1999). However to date there is no strong evidence for the occurrence of similar repeats implicated in gene regulation in Gram-positive bacteria.

Sequencing of the complete genome of a serotype M1 invasive isolate of *S. pyogenes* (Roe et al., 1999) has facilitated the identification of previously unrecognized surface proteins. Such proteins are clearly of potential interest as many are likely to be involved in interactions with the host and may therefore play important roles in the pathogenic process. Elucidation of the nature of these proteins will facilitate a fuller understanding of the virulence of *S. pyogenes* and may identify novel therapeutic or vaccine targets. In this study, an ORF (*scIB*) identified during a screen of the genome sequence, potentially encoding a surface protein with a highly repetitive collagen-like structure, is described. In addition, the ORF is shown to be highly variable and to contain SSRs immediately downstream of the putative start codon that may regulate the expression of the corresponding ScIB protein product.

**METHODS**

**Chromosomal DNA preparation.** Strains of *S. pyogenes* used in this study are listed in Table 1. Chromosomal DNA was isolated from *S. pyogenes* following overnight incubation at 37 °C on blood agar plates consisting of Todd–Hewitt broth supplemented with 1.5% (w/v) Bactoagar and 5% (v/v) sheep blood (Gibco-BRL). Overnight growth from one or two spread plates was resuspended in 500 µl sterile PBS. To this suspension 10 µl of an enzyme lysis mix was added (30 µg hyaluronidase ml⁻¹, 1 mg mutanolysin ml⁻¹ and 20 mg lysozyme ml⁻¹ in PBS) and incubated for 4 h at 37 °C. Following addition of 50 µl 10% (w/v) SDS the suspension was vortexed briefly and incubated at room temperature for 5–10 min. The resulting lysate was extracted with phenol/chloroform twice followed by chloroform extraction and DNA was precipitated with 2.5 vols ethanol and 10% (v/v) 3 M sodium acetate pH 5.2. Following washing with 70% (v/v) ethanol, DNA was dried, suspended in sterile distilled water and stored at −20 °C until required.

**PCR.** Primer pairs used for initial amplification of RST00068 and RST00596 products were 210up (5'-CCAAGCCTAAGTCTATTGCTCTCTC-3') with 210dn (5'-TTTAACTCTCTTTCTTTCTTTCTTTCTTTCTCTC-3') with 174up (5'-CGTCCAGGAGATTACCAAAAGAT-3') and 174dn (5'-GGTGT-TTCTTTGGCAGTTGT-3') respectively. Primer pair *scIBup* (5'-ATGGTTGCACTCAAAGCAG-3') and *scIBdn* (5'-GTGTGT-TTCTTTGGCAGTTGT-3') was used for amplification of *scIB*. PCR was performed under standard conditions with 32 cycles of 95 °C for 1 min, x °C for 1 min and 72 °C for 1 min where x °C represents an annealing temperature appropriate for the particular primer set.

**Sequencing.** Full-length RST00068 ORFs were sequenced from PCR products amplified using primer 210up, described above, in conjunction with primer Rst69 (5'-TCTTAACGATCTCTCTCGTAC-3') located approximately 200 bp downstream of the RST00068 sequence in the genome sequence. Sequencing was performed using a range of internal primers and dye-labelled dideoxy terminator cycle sequencing kits (Beckman Coulter) with the CEQ2000 automated DNA analysis system.

**DNA sequence analysis.** Preliminary analysis of sequences was performed using the DNASTAR package and multiple sequence alignments were performed using the CLUSTAL package (http://www.ebi.ac.uk).

**Predicted structure of ScIB.** The structure of the putative protein encoded by *scIB* was predicted using the PROTEAN program within the DNA Star package. This package uses the
Chou–Fasman (Chou & Fasman, 1978) and Garnier–Robson (Garnier et al., 1978) algorithms to predict alpha, beta and turn regions, the Garnier–Robson algorithm for predicting coil regions and the Kyte–Doolittle algorithm (Kyte & Doolittle, 1982) for calculating hydrophilicity. Protein secondary structure was confirmed using the psipred program at http://insulin.brunel.ac.uk/psipred (Jones, 1999). The presence and predicted cleavage site of signal peptide was predicted using the SignalP program at http://www.cbs.dtu.dk/services/SignalP (Nielsen et al., 1997).

**RT-PCR.** Total cell RNA was prepared from *S. pyogenes* cultures grown to mid-exponential phase in Todd–Hewitt broth at 37 °C. A 10 ml culture was harvested by centrifugation and the pellet was resuspended in 100 µl nuclease-free water. RNA was isolated using the Hybaid Ribolyser Blue Kit and Ribolyser according to the manufacturer’s instructions and was resuspended in 50 µl nuclease-free water. The RNA was then treated with 5 U RQ1 RNase-free DNase (Promega) for 30 min at 37 °C. Following inactivation of the DNase at 75 °C for 5 min, RNA was stored at −70 °C in the presence of 40 U RNasin (Promega) until required. RT-PCR was performed using the Access RT-PCR system (Promega) according to the manufacturer’s instructions. First-strand cDNA synthesis was performed at 48 °C for 45 min while second-strand cDNA synthesis and PCR amplification were performed using 36 amplification cycles at an annealing temperature of 53 °C. Positive control primers group and grtn, designed to amplify an internal 598 bp fragment of the housekeeping gene glmQ (glutamine-transport ATP-binding protein), were used. Primer pairs 655Exp (5’-GGATCGTGTGAAGATGC-3’), corresponding to the extreme N-terminus of the mature protein, with 210dn (described above) and SigP (5’-TTWGGAGTGYAAGYGCRGTT-3’), corresponding to the signal peptide, with Rst69 (described above) were used to detect an *scpB*-specific transcript.

**emm sequence typing.** This was performed as described by Beall et al. (1996) using primers 1 and 2, which were previously described by Whatmore & Kehoe (1994). The identity of *emm* sequences was confirmed by interrogation of the Centers for Disease Control *emm* sequence database at http://www.ncbi.nlm.nih.gov/ncidod/ biotech/strepinfo.html.

**RESULTS**

**Identification of ORFs potentially encoding collagen-like proteins**

To identify genes encoding putative surface proteins of *S. pyogenes* the genome sequence of *S. pyogenes* SF370, a serotype M1 isolate (http://www.genome.ou.edu/strep.html), was screened for the presence of ORFs potentially encoding proteins with the characteristic wall-associated region seen in many Gram-positive surface proteins (Navarre & Schnewind, 1999). BLAST searches were performed using the nucleotide sequence encoding the wall-associated region of the *emm5* gene from strain NCTC 8193 (accession no. U02480 – bases 1267–1371) described previously by Whatmore & Kehoe (1994). During the course of this investigation two ORFs were identified that potentially encoded proteins with high similarity to collagen on BLAST search as a result of extensive stretches of sequence encoding GXYn repeats (Beck & Brodsky, 1998). These sequences are identified in the WIT database (http://wit.mcs.anl.gov/WIT2/CGI/prot.cgi), which provides annotation and metabolic reconstructions of genome sequences, as RST00068 and RST00596. The unusual structure of these putative proteins and their potential surface location raised the possibility of direct involvement in virulence (e.g. as adhesins or in binding to matrix proteins) or indirect involvement in the autoimmune sequelae of *S. pyogenes* infection. Thus, PCR primers were designed to investigate the presence and distribution of these ORFs in a wider sample of *S. pyogenes*.

In the case of RST00068 the reverse primer was designed immediately downstream of the putative stop codon that could be clearly defined on the basis of the wall-associated sequence seen in many Gram-positive bacteria. However, the start of the gene was more difficult to predict – the WIT database predicted a TTG start codon, but there were no convincing RBS or promoter elements associated with this predicted start codon (see Fig. 1). Because of this uncertainty, the upstream primer (210up) was designed within a conserved 3′ region of a gene upstream of RST00068 which had significant similarity to GTP-binding proteins from other Gram-positive bacteria, such as *Bacillus* and *Mycobacterium*. Initially, 17 randomly selected strains were tested by PCR and a PCR product was amplified from 15 of these – interestingly these products were extremely variable in size. In both negative reactions the integrity of the DNA was confirmed by the amplification of the *emm* gene (Whatmore et al., 1994) from the same preparations. The presence of the sequence was then examined in more extensive studies using strains associated with acute rheumatic fever (6), acute post-streptococcal glomerulonephritis (16), throat infection and carriage (19) and impetigo (9). Forty-five of these 50 isolates (90%) gave a PCR product and these reactions confirmed the extensive size diversity of this product, with bands ranging from some 600 bp to approximately 2 kb. There was no clear relationship between the presence of this gene and the ecological association of the isolates and again *emm* could be amplified from all isolates that were PCR negative for the RST00068 ORF.

Primers to amplify RST00596, the second ORF potentially encoding a protein with collagen-like structure, were designed with the 3′ primer 174dn again just downstream of the putative stop codon and the upstream primer, 174up, approximately 850 bp upstream of this. In contrast to the primer set described above, these primers did not amplify a product in an initial screen of *S. pyogenes* isolates, not even when an M1 isolate was used, although it should be noted that the genome strain (SF370) was not examined in any of these screening studies.

**Sequence of RST00068**

To further characterize the nature of the genetic diversity apparent in ORF RST00068, sequencing of the complete ORF and flanking regions was undertaken from 11...
A.M. WHATMORE

Fig. 1. (a) Comparison of sequences upstream of ORF RST00068 with three distinct isolates showing the location of the predicted start codon and the associated RBS, $-10$ and $-35$ sequences. The TTG start codon predicted in the WIT database is underlined. The $]$ represents the end of the predicted signal peptide. (b) Comparison of sequences downstream of RST00068 showing the stop codon and putative transcription terminator. Isolate 655 lacks the stop codon found in all ten other isolates characterized but may utilize an additional stop codon located slightly farther downstream, but prior to the transcription terminator. Residues conserved across all sequences are shown by an asterisk.

Table 1. Characteristics of strains examined in this study

<table>
<thead>
<tr>
<th>Strain</th>
<th>Place of isolation*</th>
<th>Source/reference†</th>
<th>No. of CAAA repeats</th>
<th>In frame</th>
<th>sclB accession no.</th>
<th>emm sequence‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>B14</td>
<td>Blood isolate, India</td>
<td>T. Menon</td>
<td>4</td>
<td>No</td>
<td>AJ301816§</td>
<td>emm23</td>
</tr>
<tr>
<td>655</td>
<td>Carried isolate, UAE</td>
<td>H. Nsanze</td>
<td>5</td>
<td>Yes</td>
<td>AJ301809</td>
<td>emm3</td>
</tr>
<tr>
<td>M63</td>
<td>NCTC 10881</td>
<td>Whatmore et al. (1994)</td>
<td>5</td>
<td>Yes</td>
<td>AJ301806</td>
<td>emm63 (U11982)</td>
</tr>
<tr>
<td>P180</td>
<td>Throat infection, Bulgaria</td>
<td>A. Detcheva</td>
<td>5</td>
<td>Yes</td>
<td>AJ301807</td>
<td>emm13W</td>
</tr>
<tr>
<td>P27</td>
<td>Pus isolate, India</td>
<td>T. Menon</td>
<td>6</td>
<td>No</td>
<td>AJ301820§</td>
<td>emm49</td>
</tr>
<tr>
<td>652</td>
<td>Carried isolate, UAE</td>
<td>H. Nsanze</td>
<td>6</td>
<td>No</td>
<td>AJ301819§</td>
<td>emm12</td>
</tr>
<tr>
<td>M61</td>
<td>M61-PHLS</td>
<td>Whatmore et al. (1994)</td>
<td>7</td>
<td>No</td>
<td>AJ301822§</td>
<td>emm61 (U11984)</td>
</tr>
<tr>
<td>P64</td>
<td>Blood isolate, Finland</td>
<td>H. Seppala</td>
<td>8</td>
<td>Yes</td>
<td>AJ301817§</td>
<td>emm49</td>
</tr>
<tr>
<td>740</td>
<td>Carried isolate, UAE</td>
<td>H. Nsanze</td>
<td>9</td>
<td>Yes</td>
<td>AJ301821§</td>
<td>emm48</td>
</tr>
<tr>
<td>733</td>
<td>Carried isolate, UAE</td>
<td>H. Nsanze</td>
<td>9</td>
<td>Yes</td>
<td>AJ301813</td>
<td>ND</td>
</tr>
<tr>
<td>684</td>
<td>Carried isolate, UAE</td>
<td>H. Nsanze</td>
<td>10</td>
<td>No</td>
<td>AJ301812</td>
<td>emmSTNS14X.2</td>
</tr>
<tr>
<td>PT4854</td>
<td>4854-PHLS</td>
<td>Whatmore et al. (1994)</td>
<td>12</td>
<td>No</td>
<td>AJ301811</td>
<td>emm43 (U11979)</td>
</tr>
<tr>
<td>P15</td>
<td>Blood isolate, Denmark</td>
<td>J. Henrichson</td>
<td>12</td>
<td>No</td>
<td>AJ301818§</td>
<td>emmSTSA207</td>
</tr>
<tr>
<td>650</td>
<td>Carried isolate, UAE</td>
<td>H. Nsanze</td>
<td>13</td>
<td>No</td>
<td>AJ301805</td>
<td>emm78</td>
</tr>
<tr>
<td>669</td>
<td>Carried isolate, UAE</td>
<td>H. Nsanze</td>
<td>13</td>
<td>No</td>
<td>AJ301814</td>
<td>emm87</td>
</tr>
<tr>
<td>P19</td>
<td>Pus isolate, Denmark</td>
<td>J. Henrichson</td>
<td>14</td>
<td>Yes</td>
<td>AJ301810</td>
<td>emm89</td>
</tr>
<tr>
<td>M60</td>
<td>NCTC 10878</td>
<td>Whatmore et al. (1994)</td>
<td>14</td>
<td>Yes</td>
<td>AJ301808</td>
<td>emm60 (U11985)</td>
</tr>
<tr>
<td>P195</td>
<td>Scarlet fever, Bulgaria</td>
<td>A. Detcheva</td>
<td>15</td>
<td>No</td>
<td>AJ301815</td>
<td>emm87</td>
</tr>
</tbody>
</table>

* UAE, United Arab Emirates; NCTC, National Collection of Type Cultures, Colindale, London, UK; PHLS, Public Health Laboratory Service, Colindale, London, UK.
† T. Menon, University of Madras, India; H. Nsanze, UAE University, Al Ain, United Arab Emirates; A. Detcheva, National Centre for Infectious and Parasitic Diseases, Sofia, Bulgaria; H. Seppala, National Public Health Institute, Turku, Finland; J. Henrichson, Statens Serum Institute, Copenhagen, Denmark.
‡ emm sequences displayed >97% identity to the CDC database reference sequence. ND, Not done.
§ Only the 5’ hypervariable region of sclB was sequenced from these isolates.
Identification of predicted start codon and control elements.

As mentioned above, the rarely used TTG start codon predicted by WIT was not convincing. Fig. 1(a) shows a comparison of the sequences from three isolates, 650, 655 and 733, with the RST00068 sequence. The TTG is marked but no convincing control elements could be identified. On comparing sequences it is also clear that the TTG is located in the middle of a highly conserved region encoding a putative signal peptide as described below. Comparison of sequences at the 5′ of the ORF enabled the identification of much more likely elements. A much more promising GTG start codon was identified farther upstream – GTG is used as an initiation codon at about 5% the frequency of ATG. A purine-rich tract was located immediately upstream of this start codon where a RBS would be expected. Located some 55 bp upstream of this were putative -35 and -10 elements which exactly matched, in both sequence and spacing, the TTTACT A and TATAAT consensus sequences determined from other S. pyogenes genes (unpublished data). Fig. 1(b) shows an alignment downstream of the predicted stop codon. An 11 bp inverted repeat followed by a run of T residues resembling a classical transcription terminator was located downstream of the putative coding sequence.

Identification of pentameric repeats within the ORF. A further interesting feature of this region was the presence of a tract of pentameric CAAAA repeats immediately after the putative start codon. Although only four of these repeats were seen in the SF370 genome sequence, the number of these repeats was found to vary from a minimum of 4 copies up to 15 copies (Table 1). It is clear that variation in repeat number resulted in the ORF being either in or out of frame with respect to the predicted start codon. Although only four of these repeats were seen in the SF370 genome sequence, the number of these repeats was found to vary from a minimum of 4 copies up to 15 copies (Table 1). It is clear that variation in repeat number resulted in the ORF being either in or out of frame with respect to the predicted start codon, as illustrated in Fig. 2. ORFs containing 5, 8, or 14 repeats were in-frame while other numbers of repeats resulted in the generation of stop codons almost immediately downstream of the CAAAA tract. The exceptions were two isolates, each containing nine repeats (733 and 740), that also appear to be in-frame as a result of minor sequence variation at the end of the CAAAA tract. Thus, these repeats may represent a mechanism to modulate expression of this ORF at the translational level.

Predicted protein structure.

An alignment of the predicted amino acid sequences of five of the eleven RST00068 alleles sequenced in full is shown in Fig. 3. These five sequences represent alleles that appear to be in-frame relative to the predicted start codon and were chosen to illustrate the major features of the predicted proteins.
The 5’ region of the ORF was found to be highly conserved and was predicted by SignalP to encode a classical Gram-positive signal (Nielsen et al., 1997). Depending on the particular sequence analysed, SignalP predicted the cleavage point of the mature protein to be either after the alanine residue, indicated by an arrow, or at the alanine four residues earlier. The second alanine appears to be the most likely, given comparison with the signal peptides of other S. pyogenes proteins and the continued conservation prior to this residue. This conservation contrasts strongly with the extensive sequence diversity seen immediately downstream. The putative N-terminal tip of this protein appears to be diverse, with limited sequence similarity between many strains. On moving further downstream the sequence becomes gradually less diverse, with some similarity apparent between sequences. Approximately 60 residues into the protein an extensive region of collagen-like sequence begins. This sequence is characterized by a three-residue periodicity in which the first residue is always glycine, with a high proportion of proline, alanine, lysine, aspartate and glutamate at the other residues. The length of this GXY_3 tract varies substantially, from about 50 aa in the case of strain PT4854 (full sequence not shown) to about 240 aa in the case of P19. The more C-terminal regions of the collagen-like domain are composed of variable numbers of more defined 15 aa repeated domains with the consensus sequence GKDGQN.  

Fig. 3. CLUSTAL alignment of the complete predicted amino acid sequences from five of the RST00068 ORFs sequenced in full in this study. The predicted signal sequence cleavage point is shown by an arrow. Identical residues are shown by an asterisk while colons represent conserved substitutions and full points indicate semi-conserved substitutions. The conserved LPXTG motif is shown in bold.
Secondary structure analysis of all the putative proteins using the PROTEAN package predicted a similar structure for SclB from all strains. A diagrammatic representation of the predicted structure of one SclB protein, that from strain M63, is shown as an example in Fig. 4. Predicted molecular masses for the six isolates containing an in-frame protein range from 31.8 to 47.7 kDa, while the isoelectric points vary from 8.35 to 9.25. All the predicted proteins have a high glycine content (13.7–20.7%) and a mean proportion of 10% each of aspartate, lysine and proline. The proteins are largely hydrophilic molecules with the exception of the known hydrophobic regions constituting the N-terminal signal peptides and the C-terminal wall-associated region. Most of the molecule other than the signal peptide and wall-associated region was predicted to have a coiled structure. Regions with alpha-helical and beta-sheet structure were confined to the C- and N-terminal signals and the N-terminal hypervariable region of the protein. Secondary structure prediction using PSIPRED confirmed the above analysis, predicting an entirely coiled structure from approximately 40 residues into the mature protein as far as the wall-associated region.

**Comparison of the 5’ hypervariable region from multiple isolates.** Comparison of the predicted amino acid sequences encoded by the 5’ variable region of the ORF in all 18 strains examined in this study confirmed that the N-terminal end of the predicted mature protein was variable (data not shown). However, some of the sequences were clearly closely related to each other and appeared to have evolved as a result of the accumulation of nonsynonymous point mutations. In two cases, strains shared identical predicted N-terminal amino acid sequences (P195 with 669 and P15 with PT4854). The full ORFs were sequenced from both P195 and 669, both of which are emm87 isolates and the entire sequence was found to be identical with the exception of a difference in the number of CAAAA repeats. As described above, there is a clear trend towards sequences showing more similarity to each other on moving away from the extreme N terminus of the predicted protein.

**Detection of expression of the RST00068 ORF**

To determine whether a transcript corresponding to RST00068 was present in *S. pyogenes*, RT-PCR was performed using specific primers. Representative results using RNA isolated from strain 655 are shown in Fig. 5. An RNA transcript of approximately the correct size was detected using primer pair 655Exp/210dn located within the predicted RST00068 transcript and predicted from the sequence analysis to give a product of 815 bp. In contrast, when using the primer set SigP/Rst69, where Rst69 is located downstream of the predicted transcription terminator, no RNA transcript was detected in the DNase-treated preparation. In the RNA fraction that was not subjected to DNase treatment, a band of approximately 1 kb was seen, equating to the 1011 bp product which would be predicted assuming the region downstream of RST00068 in strain 655 was of identical length to the equivalent sequence in strain SF370. Primers to the housekeeping gene *glnQ* were used as a positive control and detected a transcript specific for this gene. Similar results to those obtained using the RNA from isolate 655 were also obtained using RNA from three other isolates, P180, P19 and 733, confirming the presence of a transcript corresponding to ORF RST00068 in *S. pyogenes* in culture *in vitro*.

**Confirmation of the presence of a second collagen-like protein**

During the course of this work the sequence of an *S. pyogenes* gene, *scl*, encoding a collagen-like protein was deposited in GenBank (AF252861 – S. Lukomski, K. Nakashima, I. Abdi, V. J. Cypriano, S. D. Reid & J. M. Musser). This gene was isolated from SF370, the genome-sequencing strain and, based on the sequence of the conserved signal and wall-associated regions, clearly corresponded to the second ORF described earlier (RST00596) despite some sequence anomalies. The inability to amplify this gene by PCR using a single primer set based on this sequence in this study (174up and 174dn) most probably reflects genetic variability at

---

> **Fig. 4.** Predicted secondary structure and hydropathy profile of SclB from strain M63. The analysis was performed using the default parameters of the PROTEAN program within the DNA Star package.
this locus and an unfortunate positioning of PCR primers. To confirm the presence of this gene in isolates from which ORF RST00068 had already been characterized new PCR primers were designed based on the scl sequence (sclup and scldn). A product was obtained from all 11 isolates from which RST00068 had been sequenced in full and, as with the RST00068 product, the scl product was of variable size. Sequencing of the 5′ of two of these products from isolates 655 (accession no. AJ301823) and M60 (accession no. AJ301824) confirmed that this gene is distinct from that corresponding to RST00068 (data not shown). This gene appears to share some of the characteristics of RST00068. The limited sequencing performed suggests that, following a conserved signal peptide, the N-terminal sequence encoded by scl is variable between isolates and the size variation of PCR products was likely to reflect a variable length of GXYN repeats. Thus it appears that many isolates of S. pyogenes have the potential to express two collagen-like proteins: scl, already deposited in the database and from hereon referred to as sclA, and a second putative gene which, for the sake of consistency, will be referred to from hereon as sclB.

**DISCUSSION**

This study has characterized an ORF, sclB, with the potential to encode a previously uncharacterized hyper-variable surface protein of S. pyogenes. This ORF appears to be present in a large proportion of isolates although further work using alternative primer sets and/or Southern blotting experiments is necessary to determine whether it is ubiquitous. Several strands of evidence lead to the belief that a real gene has been correctly identified. Firstly, convincing promoter elements are located upstream of the proposed gene with a sequence encoding a signal peptide located immediately downstream of the putative start codon and a sequence encoding classic Gram-positive wall-associated sequences at the 3′ end of the ORF followed by a putative transcription terminator. Furthermore, the pattern of variability is consistent with that seen in some other S. pyogenes surface proteins. Both members of the emm gene family and the sof gene, encoding the fibronectin-binding opacity factor (Whatmore et al., 1994, 1995; Rakonjac et al., 1995; Courtney et al., 1999), encode proteins displaying a similar pattern of C-terminal conservation with extensive diversity in the N-terminal distal part of the protein. This pattern of diversity suggests that sclB is expressed at least transiently and is subject to selective pressures imposed by the host immune response which are likely to be strongest on the more exposed N-terminal tip of the molecule. In agreement with the idea of selective pressure acting against this region, comparison of closely related sequences such as those from P195 and M63 demonstrates an inflation of nonsynonymous change over synonymous. Thus, in the sequence encoding the N-terminal 20 aa there are eight non-synonymous substitutions compared with only three synonymous substitutions – such an inflated ratio is indicative of positive selection for amino acid substitutions (Smith et al., 1995). This evidence that sclB can be expressed in vivo is supported by RT-PCR results demonstrating that a transcript corresponding the ORF defined in this study is produced during in vitro cultivation.

This study also identified a novel repetitive element apparently located within the sclB coding sequence. Repetitive DNA elements, often called SSRs have been increasingly recognized in prokaryotes in recent years. As a result of slipped-strand mispairing altering the number of such repeats they can be involved in gene regulation both when they are positioned upstream of the gene, thus influencing transcription, or within the gene, influencing translation (Henderson et al., 1999). There are now several examples of regulation of genes occurring at the translational level as a result of variation in the number of SSRs at the 5′ of genes, similar to those
seen in sclB, which alter the reading-frame. For example, the adhesins encoded by the opa genes of Neisseria gonorrhoeae and Neisseria meningitidis achieve an on–off phase variation at the translational level by alterations in the reading frame of a semi-variable portion of the leader sequence consisting of a run of CTCTT repeats which can vary as a result of slipped-strand mispairing (Stern et al., 1986). In a similar manner translational control as a result of slipped-strand mispairing within a run of cytosine residues in the ORF has been shown to regulate expression of the Bordetella pertussis bugS gene, a major regulator of virulence genes in this organism (Stibitz et al., 1989).

Similar events have been implicated in the regulation of genes in a variety of other Gram-negative organisms, including Vibrio cholerae, Haemophilus influenzae, Helicobacter pylori, Mycoplasma hyorhinis, Haemophilus somnus, Pasteurella haemolytica and Mycoplasma fermentans (Henderson et al., 1999; van Belkum et al., 1999). Recent analysis of whole genome sequences has led to suggestions that translational slipped-strand mispairing may be a more important and widespread mechanism of regulation than has previously been realized. A substantial number of genes in both H. pylori and N. meningitidis contain such SSRs and are potentially regulated by this mechanism (Saunders et al., 1998, 2000). There is currently no direct evidence that sclB is regulated at the translational level by slipped-strand mispairing of the CAAA repeats. However, circumstantial evidence in the form of the location of these repeats at the 5′ of the signal peptide encoding sequence and immediately downstream of a putative start codon suggests that this is a probable scenario. Although no attempt was made to demonstrate an alteration in repeat number in culture, different numbers of CAAA repeats were found to be associated with sequences sharing otherwise identical sclB sequences (669 and P195), providing additional support for this scenario. To date, translational regulation by slipped-strand mispairing has not been demonstrated in Gram-positive bacteria although slipped-strand mispairing of internal sequence repeats has been implicated in the antigenic variation of some surface proteins, notably the highly repetitive emm genes which encode one of the major virulence proteins of S. pyogenes (Harbaugh et al., 1993; Whatmore et al., 1994).

The striking and extensive collagen-like structure of the putative ScIb protein appears unprecedented amongst the prokaryotic proteins characterized to date. The extensive run of GXY repeats resembles the remarkably regular structure of collagen, where virtually every third residue is glycine. It is thus tempting to speculate that ScIb may exist in the same triple-helix conformation as collagen, composed of three protein units wound around each other to form a stiff cable. The three-residue glycine periodicity is crucial in such a structure as there are three residues per turn of the helix and the only amino acid small enough to fit in an interior position is glycine. Previously only very limited runs of collagen-like sequence have been reported in bacteria, such as in a hydrolase of Klebsiella pneumoniae, which contains up to six tripeptide repeats (Charalambous et al., 1988), and in the Streptococcus sanguis platelet-aggregation-associated protein (Erickson & Herzberg, 1993). Interestingly, a phage-encoded hyaluronidase found in S. pyogenes can contain a run of ten GXY repeats although this region is often missing, implying that it is not important for hyaluronidase activity (Hynes et al., 1995). During the course of this work Lindmark & Guss (1999) reported a novel fibronectin-binding protein, SFS, from Streptococcus equi, which also shows high scores against collagen on similarity searches. The basis of this identity was the high content of glycine, serine and proline present in both proteins and a ten-residue motif (QGERGEAGPP) seen in both collagen and the fibronectin-binding domain of SFS. Although ScIb and SFS share some common characteristics in terms of amino acid composition there are clear differences between the proteins. While the collagen-like GXY repeats of ScIb are virtually perfect and very extensive, SFS has only very limited runs of perfect three-residue periodicity. In addition, although SFS does possess a signal peptide it does not contain any of the sequence motifs known to mediate attachment to the cell wall.

The function of ScIb remains unknown. Attempts to express the protein in Escherichia coli have so far proved unsuccessful, perhaps reflecting the unusual structure of ScIb. It is possible that ScIb is involved in either adhesion to the cell matrix or binding of host proteins from the extracellular milieu. A further intriguing possibility, given the structural similarity to such a ubiquitous host protein as collagen, is the involvement of this molecule in the autoimmune sequelae, particularly acute rheumatic fever, associated with S. pyogenes. Although the sequences of ScIb are rather variable, the M protein precedent demonstrates that S. pyogenes proteins displaying little or no primary sequence similarity can have identical activities (Johnsson et al., 1998). In light of the similarity to SFS described above it is possible that ScIb represents a fibronectin-binding protein, particularly as collagens of types I to V have been shown to bind fibronectin (Engvall et al., 1978). S. pyogenes is already known to harbour at least two fibronectin-binding proteins, SfbI/protein F (Sela et al., 1993; Talay et al., 1994) and SfbII/OF (Kreikemeyer et al., 1995; Rakonjac et al., 1995) although neither are found in all strains (Goodfellow et al., 2000). In addition, some proteins encoded by certain members of the emm gene family are believed to possess fibronectin-binding activity (Frick et al., 1995; Reichardt et al., 1995). However, it is becoming apparent that binding of fibronectin by bacterial surface proteins is a common property and multiple fibronectin-binding proteins have now been reported from a number of other Gram-positive bacteria (Joh et al., 1999).

The confirmation that at least two surface proteins with collagen-like structure are present in at least some isolates of S. pyogenes identifies a new family of variable proteins within this organism. Although limited sequence data are available as yet for ScA, it is likely that...
SclA and SclB share the same highly repetitive structure with a variable-length tract of GXYn repeats following an N terminus displaying extensive sequence diversity. However, SclA appears to lack the pentameric nucleotide repeats seen within the signal-peptide region of SclB. This family appears to share a number of common features with members of the emm gene family. Both families have variable N termini that, at least in part, reflect a positive selection for the accumulation of nonsynonymous point mutations (Whatmore et al., 1994). Both, in common with many Gram-positive surface proteins, also encode extensive tandem amino acid repeats within the structural gene. It seems likely that genetic variation in members of the scl family can be generated by the gain or loss of repeats as a result of intragenic recombination events and/or slipped-strand mispairing as seen in members of the emm gene family (Whatmore et al., 1994). The presence of two genes sharing considerable sequence similarity further raises the possibility that variation could also be generated as a result of intergenic exchange of DNA between sclA and sclB, as has been noted between emm and emm gene representatives in the emm gene family (Whatmore & Kehoe, 1994; Podbielski et al., 1994). It is also possible that one scl gene may serve as a silent cassette acting as a potential reservoir of diversity for antigenic variation.

The results presented in this paper represent an initial characterization of a novel S. pyogenes gene, sclB, potentially encoding a surface protein and belonging to a family of previously uncharacterized proteins. Future work will be aimed at expressing and elucidating the activities of SclB, determining whether post-infection human sera contain antibodies to SclB, confirmation of the role of the pentameric repeats in modulation of expression, and primer extension studies to confirm the location of control elements predicted by sequence comparisons.

ACKNOWLEDGEMENTS

A. M. W. gratefully acknowledges the support of a Wellcome Trust Research Fellowship in Biodiversity (Grant no. 053589) and the assistance of Thangam Menon, Herbert Nsanze, Antoaneta Detcheva, Jorgen Henrichson and Helena Seppala in providing some of the isolates used in this study.

REFERENCES


Received 25 August 2000; revised 10 October 2000; accepted 16 October 2000.