Horizontal transfer of the immunoglobulin A1 protease gene (iga) from Streptococcus to Gemella haemolysans

Nobuko Takenouchi-Ohkubo,1,2 Lotte M. Mortensen,1 Kim R. Drasbek,1 Mogens Kilian1 and Knud Poulsen1

1Institute of Medical Microbiology and Immunology, The Bartholin Building, University of Aarhus, DK-8000 Aarhus C, Denmark
2Department of Pathology, Nihon University School of Dentistry, Tokyo, Japan

Bacterial IgA1 proteases share the ability to cleave human IgA1 at the hinge region. Nature has developed this trait along at least five independent evolutionary lineages. To obtain further insight into the phylogeny and function of IgA1 proteases, the nucleotide sequence of the iga gene that encodes the IgA1 protease was determined from two Streptococcus mitis strains and one Gemella haemolysans strain. Heterologous expression in Escherichia coli confirmed that the genes encode human IgA1-cleaving activity. IgA1 proteases from Streptococcus and G. haemolysans shared structural features, including a motif typical for zinc-dependent metalloproteases of clan MA(E) family M26 and an N-terminal signal sequence followed by an LPXTG cell-wall-anchor motif and two putative membrane-spanning domains. In addition, they all harboured a repeat region preceding the active site of the protease. In the streptococcal IgA1 proteases, a G5 domain, which has been suggested to bind N-acetylgalcosamine, was identified. Conservation of these structures in otherwise diverse proteases suggests that they are essential to the biological function of the enzyme. The phylogenetic distribution of homologous iga genes and conservation of gene order in the iga gene region in different Streptococcus species, combined with the sequence homologies, strongly suggest that the iga gene is more ancient in Streptococcus than in G. haemolysans, and therefore that the IgA1 protease gene was transferred from Streptococcus to G. haemolysans.

INTRODUCTION

Bacterial IgA1 proteases are specific post-proline endopeptidases that cleave human IgA1, including its secretory form, S-IgA1, at the hinge region (reviewed by Plaut, 1983; Kilian et al., 1996). IgA1 is the predominant immunoglobulin subclass in mucosal secretions of the upper respiratory tract and is abundant in serum. Upon cleavage, the resulting Fab fragments retain antigen-binding capacity, whereas Fc-mediated effector functions are abolished. In this way, these bacteria are able to evade the protective functions of this principal mediator of adaptive immunity at mucosal surfaces and may, concomitantly, take advantage of released specific Fab fragments in facilitating adherence and immune disguise (Kilian & Reinholdt, 1987; Kilian et al., 1996; Weiser et al., 2003). The biological significance of these enzymes is indirectly inferred by the facts that nature has developed IgA1-cleaving activity among bacteria colonizing humans along at least five independent evolutionary lineages, and that all three principal bacteria causing meningitis produce an IgA1 protease (Kilian et al., 1996; Kosowska et al., 2002). Metallo-type IgA1 proteases are produced by all members of the species Streptococcus pneumoniae and by some members of the related commensal species Streptococcus sanguinis, Streptococcus oralis, Streptococcus mitis and Streptococcus infantis.

The specificity for human and great ape IgA1 implies that lack of a relevant animal model has precluded definitive proof of the role of the IgA1 proteases in pathogenesis. However, in a mouse model that did not include human IgA1, the pneumococcal IgA1 protease was identified as a virulence factor in colonization as well as in sepsis (Polissi et al., 1998; Chiavolini et al., 2003). This observation suggests that the IgA1 proteases may have features unrelated to the cleavage of human IgA1 that contribute to the pathogenesis of infections caused by IgA1-protease-producing bacteria.

Sequence comparisons clearly show that the S. sanguinis, S. pneumoniae and S. oralis IgA1 proteases are closely related (Gilbert et al., 1991; Poulsen et al., 1996, 1998; Wani et al., 1996; reviewed by Kilian & Reinholdt, 2005). However,
among \textit{S. mitis} strains, the IgA1 protease gene \textit{iga}, in addition to being variably present, shows surprisingly varying degrees of homology to the other streptococcal \textit{iga} genes, as revealed by Southern blot analysis (Poulsen \textit{et al.}, 1998). No sequence data are available to explain this polymorphism.

\textit{Gemella haemolysans} is a Gram-positive, facultatively anaerobic coccos frequently found as part of the normal flora of the upper respiratory tract and the oral cavity of humans (Facklam \& Elliott, 1995), although it has been implicated in the aetiology of periodontitis (Paster \textit{et al.}, 2001) and is an occasional cause of subacute endocarditis and meningitis (Facklam \& Elliott, 1995). \textit{G. haemolysans} has previously been shown to produce human IgA1-cleaving activity, a feature that is unique to this species within the genus \textit{Gemella} (Kilian \textit{et al.}, 1983; Lomholt \& Kilian, 2000). The activity of the \textit{G. haemolysans} \textit{iga} protease is similar to that of \textit{Streptococcus} species, in that it cleaves at the same Pro–Thr bond in the hinge region of human IgA1 and is inhibited by EDTA, indicating that it is a metallo-proteinase (Lomholt \& Kilian, 2000). Otherwise, the \textit{G. haemolysans} \textit{iga} protease has not been examined and its phylogenetic relationships to other \textit{iga} proteases are not known. \textit{G. haemolysans} is not closely related to other bacterial species known to produce \textit{iga} protease activity, and the evolutionary origin of the \textit{iga} protease gene in this species is unknown.

To obtain further insight into the phylogeny and function of \textit{iga} proteases, we sequenced and characterized the \textit{iga} gene encoding the \textit{iga} protease in \textit{G. haemolysans} and in two strains of \textit{S. mitis}.

\section*{METHODS}

\subsection*{Bacterial strains and growth conditions.} \textit{G. haemolysans} strains CCUG 411, CCUG 4815, CCUG 28134, CCUG 33985 and CCUG 37985\textsuperscript{1} were obtained from the Culture Collection of the University of Göteborg, Sweden, and SK940 was a dental plaque isolate from our own collection. Nine strains of \textit{S. mitis}, four strains of \textit{S. oralis} and four strains of \textit{S. infantis}, including the type strains for each of the three species (NCTC 12261, NCTC 11427 and GTC 849, respectively), were from our own collection. \textit{S. pneumoidea} TIGR4 was obtained from The Institute for Genomic Research, Rockville, MD. All strains except for \textit{S. mitis} SK142 (NCTC 12261) and \textit{S. infantis} SK959 (GTC 849) produced human IgA1-cleaving activity. Species affiliation of the streptococci was confirmed by partial sequencing of the \textit{sodA} gene, as described by Hoshino \textit{et al.} (2005). We have previously described the \textit{iga} gene sequence of \textit{S. mitis} strain SK141 (GenBank accession number Y10285) (Poulsen \textit{et al.}, 1998). However, recent re-evaluation of the taxonomy of the \textit{mitis} group of \textit{Streptococcus} reveals that this strain should be assigned to \textit{S. oralis} (unpublished results). Therefore, we also sequenced the \textit{iga} gene from two strains that were unambiguously assigned to \textit{S. mitis}.

IgA1 protease assay. For detection of human IgA1-cleaving activity, the sample and PBS, pH 7.4, up to 50 \mu l, were mixed with 2 \mu l human myeloma IgA1 (8 mg IgA1 ml\textsuperscript{-1}) and incubated overnight at 37 °C, and characteristic cleavage fragments were detected by reducing SDS-PAGE of 10 \mu l reaction mixture followed by staining with Coomassie Brilliant Blue. To differentiate between cell-associated and secreted IgA1 protease activity, an overnight culture was centrifuged to separate the fractions and the bacteria were resuspended in an equal volume of PBS, pH 7.4. The protein synthesis was stopped by adding 20 mg chloramphenicol 1\textsuperscript{-1} prior to adding substrate IgA1 and incubation, as described above. Testing serial twofold dilutions of the fractions, combined with visual inspection of the amount of human IgA1 cleaved in the SDS-PAGE analysis, were used to quantify the IgA1 protease activity. For the activity blot experiment, proteins in the supernatant were concentrated 20 times using Centricon concentrators (Amicon) and separated by 7% SDS-PAGE under reducing conditions. IgA1 protease activity in the gel upon electrophoresis was detected as described previously (Poulsen \textit{et al.}, 1996). In this assay, human IgA1-cleaving activity in the gel was reflected as loss of staining on the membrane.

\subsection*{PCR and DNA sequencing.} For PCR amplification, we used ~1 ng genomic DNA as template and Ready-To-Go PCR beads (Amersham Biosciences). The temperature profile for the PCR was an initial denaturation at 94 °C for 5 min, followed by 30 cycles at 94 °C for 1 min, 60 °C for 2 min, 72 °C for 2 min, and a final extension at 72 °C for 8 min. When used for DNA sequencing, the resulting amplicons were purified using Wizard Minicolumns (Promega). For sequencing, we used the same primers as for the PCR, as well as primers designed from the previously determined sequence. Sequencing was achieved using the Thermo Sequenase dye terminator cycle sequencing kit (Amersham Bioscience) and the resulting products were analysed with an Applied Biosystems PRISM 377 automated sequencer (PerkinElmer/Applied Biosystems). Based on an alignment of 13 published \textit{iga} gene sequences from \textit{Streptococcus} species (Gilbert \textit{et al.}, 1991; Poulsen \textit{et al.}, 1996, 1998; Wani \textit{et al.}, 1996), we designed two degenerate primers, iga-for 5'-GCTAATC-GGGYTWTAACAA-3' and iga-rev 5'-ATGSTCATYTCATGRT-AT-3', located within conserved regions (corresponding to positions 3949–3968 and 4707–4688, respectively, in the ORF of the \textit{S. pneumoniae} strain PK81 \textit{iga} gene, accession number X94909). Inverse PCR was performed as described by Ochman \textit{et al.} (1988) and Kosowska \textit{et al.} (2002). Briefly, restriction enzymes selected on the basis of obtained sequences were used to digest the genomic DNA prior to circularization by self-ligation, and outward-pointing oligonucleotide primers, designed from previous sequences, were used for PCR performed as described above. Direct sequencing was performed on both strands of overlapping PCR amplicons (details of the sequencing strategy are available upon request). The presence and sequences of imperfect tandem repeats were confirmed by the size of PCR products using several pairs of primers flanking the repeat region combined with sequencing using repeat-specific primers.

\subsection*{Heterologous expression in \textit{E. coli}.} The oligonucleotides 5'-ATGGAGAATTTGCAATTATCAGTC-3' and 5'-TCTCCACCA-AATATGATGTGTTCT-3' were used in PCR of genomic DNA from \textit{G. haemolysans} strain CCUG 4815 to amplify a 5541 bp fragment of the ORF encoding the presumed secreted \textit{iga} protease (see Results and Discussion). Similarly, oligonucleotide 5'-GGAAAATAATCTTTTACT-3' was combined with 5'-ATGSGTCATYTCATGRGT-3', located within conserved regions (corresponding to positions 3949–3968 and 4707–4688, respectively, in the ORF of the \textit{S. pneumoniae} strain PK81 \textit{iga} gene, accession number X94909). Inverse PCR was performed as described by Ochman \textit{et al.} (1988) and Kosowska \textit{et al.} (2002). Briefly, restriction enzymes selected on the basis of obtained sequences were used to digest the genomic DNA prior to circularization by self-ligation, and outward-pointing oligonucleotide primers, designed from previous sequences, were used for PCR performed as described above. Direct sequencing was performed on both strands of overlapping PCR amplicons (details of the sequencing strategy are available upon request). The presence and sequences of imperfect tandem repeats were confirmed by the size of PCR products using several pairs of primers flanking the repeat region combined with sequencing using repeat-specific primers.
manufacturer. The PCR products were separated by 1 % agarose gel electrophoresis, visualized by staining with ethidium bromide, and DNA in the band was purified using QIAEXII (Qiagen). Adenosine overhangs were introduced by incubating with Hot Master Mix (Eppendorf) for 8 min at 72 °C. The PCR fragments were cloned using the pBAD-TOPO TA cloning kit and E. coli TOP10F' as host (Invitrogen). The expression vector pBAD-TOPO is intended for cloning PCR products and it adds 14 aa to the N terminus of the recombinant protein and 28 residues, including a V5 epitope and a polyhistidine region, to the C terminus, and expression from the P_{BAD} promoter is controlled by presence of arabinose. A single clone from each of the three transformations with the insert in the right orientation was selected. Expression of the recombinant protein was induced in a fresh broth culture by adding arabinose to a concentration of 0.002 %, as described in the manufacturer’s manual. After incubation for 4 h, the bacteria were disrupted by repeated freeze/thawing.

Sequence analyses. For sequence comparisons and alignments, we used the program PILEUP included in the GCG package (Wisconsin Package Version 10.3; Accelrys).

For searching public sequence databases we used BLAST (available at http://www.ncbi.nlm.nih.gov/blast/). For predicting iga genes, we used the program EasyGene trained on Staphylococcus aureus as host organism (available at http://www.cbs.dtu.dk/services/EasyGene/) (Larsen & Krogh, 2003). For predicting signal peptides, we used the program SignalP (available at http://www.cbs.dtu.dk/services/SignalP) (Nielsen et al., 1997). Transmembrane segments were predicted by the program DAS (available at http://www.sbc.su.se) (Cserzo et al., 1997). Protein domains were predicted by the program Pfam (available at http://pfam.wustl.edu/)(Bateman et al., 2004). For construction of dendrograms based on the neighbour-joining method, we used MEGA3 (available at http://www.megasoftware.net/) (Kumar et al., 2004).

Southern blot analysis. Whole-cell DNA was purified as described by Hohwy & Kilian (1995). Approximately 10 μg DNA was digested with EcoRI and treated with RNase. DNA fragments were separated by 1 % agarose gel electrophoresis and transferred to and fixed onto a nylon membrane, as described by Sambrook et al. (1989). For detection of iga gene sequences on the blot, we used the Gene Images Alkphos Direct Labelling and Detection System (Amersham Bioscience). The six probes used in the Southern blot experiments were equivalent fragments of the 3’ part of the iga genes from G. haemolysans CCUG 4815, S. mitis SK564 and S. pneumoniae TIGR4. The fragments were CCUG 4815-1 and CCUG 4815-2 (positions 4165–4658 and 5577–5906, respectively, in the ORF), SK564-1 and SK564-2 (positions 4354–4844 and 5778–6119, respectively, in the ORF), and TIGR4-1 and TIGR4-2 (positions 4159–4649 and 5577–5915, respectively, in the ORF).

RESULTS AND DISCUSSION

The IgA1 proteases of S. mitis and G. haemolysans

Two strains of S. mitis, SK564 and SK609, and G. haemolysans strain CCUG 4815, all with IgA1 protease activity, were included in this part of the study. Pronounced differences in the amount of cell-associated versus secreted IgA1 protease activity have been demonstrated among different species of bacteria (Reinholdt & Kilian, 1997). In the three strains included in the present study, the greater part (> 80 %) of the IgA1 protease activity was present in the supernatant of an overnight culture (results not shown). In an activity blot experiment performed upon separation of secreted proteins by reducing SDS-PAGE, the IgA1 protease was present in multiple molecular forms (Fig. 1). In all three strains, the largest protein in the supernatant with human IgA1-cleaving activity had an apparent molecular mass of ~ 220 kDa (Fig. 1). This is very similar to what has previously been observed for other streptococcal IgA1 proteases (Poulsen et al., 1996, 1998). For the S. pneumoniae IgA1 protease, we showed that the smaller molecular forms emerged during concentration and prolonged storage, while fresh cultures only showed the largest forms (unpublished results). Autoproteolysis was presumably the cause of this breakdown of the protein into smaller sizes that retained IgA1 protease activity.

Sequencing S. mitis and G. haemolysans iga genes

PCR amplification using the primers iga-for and iga-rev on genomic DNA from S. mitis strains SK564 and SK609 and G. haemolysans strain CCUG 4815 as template resulted in a fragment of ~ 0.7 kb. Sequencing of the resulting amplimers revealed an ORF in all three bacteria with similarity to the known streptococcal iga genes, suggesting that the amplified fragments represented the homologous iga genes. The sequences of the remaining parts of the presumed iga genes and flanking regions were obtained by ‘walking’ in both directions using inverse PCR. For S. mitis strains SK564 and SK609 and for G. haemolysans strain CCUG 4815, continuous sequences of 7054, 7553 and 7453 nt, respectively, were obtained.

Fig. 1. IgA1 protease profiles. Proteins in the supernatant of an overnight culture (results not shown). In an activity blot experiment performed upon separation of secreted proteins by reducing SDS-PAGE, the IgA1 protease was present in multiple molecular forms (Fig. 1). In all three strains, the largest protein in the supernatant with human IgA1-cleaving activity had an apparent molecular mass of ~ 220 kDa (Fig. 1). This is very similar to what has previously been observed for other streptococcal IgA1 proteases (Poulsen et al., 1996, 1998). For the S. pneumoniae IgA1 protease, we showed that the smaller molecular forms emerged during concentration and prolonged storage, while fresh cultures only showed the largest forms (unpublished results). Autoproteolysis was presumably the cause of this breakdown of the protein into smaller sizes that retained IgA1 protease activity.
For *S. mitis* SK564, the sequence revealed an ORF of 6216 nt, starting with ATG, and with the potential of encoding a protein of 2072 aa. The *S. mitis* SK609 sequence contained an ORF of 6093 nt, also starting with ATG and encoding a protein of 2031 aa. The ORFs of both *S. mitis* strains showed significant similarity to known streptococcal IgA1 proteases. In both strains, the ORF was preceded by a typical ribosome-binding site (RBS), AAGGAGGA, 3 nt upstream of the proposed start codon. Just upstream of this, −10 and −35 consensus promoter sequences were present. A possible weak transcription terminator with a stem of 6 nt and a loop of 3 nt was present in both sequences, 65 nt downstream of the stop codon. In both sequences the ORF was predicted to represent a real gene by the program EasyGene. In the same orientation and 209 nt upstream of the ORF, both strains harboured an ORF with homology to the C-terminal part of a gene encoding a conserved hypothetical protein, preceding the *iga* gene in the published genome sequences of *S. pneumoniae* strains TIGR4 and R6. In *S. mitis* strain NCTC 12261T, which lacks IgA1 protease activity, the order of genes was as in the pneumococci, except that the *iga* gene was missing (Fig. 2). In both *S. mitis* SK564 and SK609, the sequence downstream of the ORF was homologous to that downstream of the *iga* gene in *S. pneumoniae* strain PK81 (GenBank accession number X94909), including an ORF with homology to the 5′ end of the paralogous zinc metalloproteinase B gene *zmpD* and *iga*. A similar gene organization, with the *iga* gene followed by an IgA1-protease-like gene termed *zmpD*, has been found in *S. pneumoniae* strain G54 (Polissi et al., 1998), strain PK81 (Poulsen et al., 1996), and approximately half of clinical isolates of *S. pneumoniae* (Camilli et al., 2006), whereas this apparent gene duplication is absent in strains TIGR4 and R6. In summary, the gene order in the region around *iga* was conserved between *S. mitis* and certain *S. pneumoniae* strains (Fig. 2).

For *G. haemolysans* strain CCUG 4815, the combined sequence information revealed an ORF of 6051 nt, starting at the first ATG codon and with significant similarity to the streptococcal *iga* genes. The sequences TTGCAAT and TACAAT, separated by 17 nt and located 18 nt upstream of the ORF, have similarity to the −35 and −10 elements of the consensus promoter sequence (Hawley & McClure, 1983). However, the first ATG in the ORF was not preceded by a possible RBS. We propose that the TTG sequence located 51 nt downstream of this ATG codon serves as a start codon in the *G. haemolysans* *iga* gene, because of the high degree of similarity with known streptococcal IgA1 proteases at the N terminus, combined with the presence of the putative RBS GAAGGA, 7 nt upstream of the TTG triplet. TTG is often used as a start codon in bacteria (Kozak, 1999). In addition, the program EasyGene predicted this ORF to be a real gene. The ORF has the potential of encoding a protein of 2000 aa. Twenty-six nucleotides downstream, the ORF is followed by a possible transcription terminator, consisting of an inverted repeat structure with a stem of 13 nt and a loop of 13 nt. Putative genes flanking the large ORF, and in the same orientation, were identified at the ends of the sequence obtained. At the upstream end, separated from the proposed *iga* gene by 263 nt, 621 nt of an ORF encoding a protein with 70% identity to the C-terminal part of an ABC transporter from *Streptococcus agalactiae* were present. Five nucleotides downstream of this ORF, an inverted repeat structure, which may constitute a transcription terminator, was identified. At the downstream end, separated from the *iga* gene by 166 nt, there was an ORF of 397 nt encoding a
protein with up to 42% amino acid identity to the N terminus of a copper homeostasis protein, CutC, found in a variety of prokaryotes and eukaryotes. This ORF was preceded by a typical ribosome-binding sequence GGAGAA, 6 nt upstream of the ATG start codon. Thus, there was no synteny of the area of the genome containing the iga gene in *G. haemolysans* compared to *S. mitis* and *S. pneumoniae* (Fig. 2).

**Heterologous expression of the *S. mitis* and *G. haemolysans iga* genes in *E. coli***

In *S. pneumoniae*, the iga gene is part of a gene family that includes at least two other paralogous members, zmpB and zmpC. Among these zinc metalloproteinases, only the IgA1 protease cleaves human IgA1 (Oggioni et al., 2003). A higher degree of similarity to iga than zmpB and zmpC suggested that the three sequenced ORFs represented the IgA1 protease in *S. mitis* SK564, *S. mitis* SK609 and *G. haemolysans* CCUG 4815 (see below). In order to verify this, we performed heterologous expression in *E. coli*. For each of the three bacteria, the part of the ORF encoding the presumed secreted version of the IgA1 protease (see below) was amplified by PCR. The PCR products corresponded to the sequences encoding amino acid positions 154–2072, 154–2031 and 155–2000 in *S. mitis* SK564, *S. mitis* SK609 and *G. haemolysans* CCUG 4815, respectively. The amplicons were cloned into the expression vector pBAD using *E. coli* TOP10F’ as host. The resulting recombinant plasmids with the inserts in the right orientation were termed pBAD-SK564-iga, pBAD-SK609-iga and pBAD-gemella-iga. Expression of the recombinant protease proteins was induced by adding arabinose. After incubation the bacteria were disrupted by repeated freezing and thawing. Each of the three resulting lysates possessed human IgA1-cleaving activity, demonstrating that the three sequenced ORFs encode the IgA1 protease (Fig. 3).

**Structural features of the *S. mitis* and *G. haemolysans* IgA1 proteases**

The deduced primary structure of the *S. mitis* and *G. haemolysans* IgA1 proteases shared a number of sequence motifs with known orthologous enzymes of streptococci (Fig. 4). The amino acid sequence HEMTH followed by an E residue 20 aa downstream is typical of bacterial metalloproteinases of the gluclizin type (Vallee & Auld, 1990; Hooper, 1994), and the IgA1 proteases belong to the clan MA(E) family M26 of proteases listed in the MEROPS database (available at http://merops.sanger.ac.uk).

The signal peptides of the *S. mitis* SK564 and SK609 IgA1 proteases each comprised 42 aa, as predicted by the program SignalP. The size of the signal peptide predicted for the IgA1 protease of *G. haemolysans* CCUG 4815 was 44 aa. The presence of a cysteine at position 22 in the *G. haemolysans* IgA1 protease might suggest that it is a lipoprotein, but the sequence upstream of this residue did not match the consensus pattern for signal peptides of lipoproteins in Gram-positive bacteria (Sutcliffe & Harrington, 2002). Otherwise, all three IgA1 proteases were devoid of cysteine residues. The streptococcal and *G. haemolysans* signal peptides included a sequence similar to the motif YSIRK-G/S, which is associated with proteins carrying a Gram-positive anchor domain at the C terminus (Tettelin et al., 2001; Bae & Schneewind, 2003).
In the two *S. mitis* and the *G. haemolysans* IgA1 proteases, as well as in all other known streptococcal IgA1 proteases, a cell-wall-anchor motif LPNTG (LPYTG in *S. mitis* strain SK609) (Navarre & Schneewind, 1999; Pallen *et al.*, 2001) is present 53 aa downstream of the proposed signal peptidase site, and followed by two hydrophobic domains with the potential of spanning the cell membrane (Fig. 4). The region between the two putative transmembrane segments is rich in lysines and, according to the positive-inside rule, this region is predicted to face the cytoplasm (von Heijne, 1989). We have previously hypothesized that the streptococcal protease is cleaved at the LPNTG sortase processing site at its N terminus, traverses the cytoplasmic membrane twice, and is released by proteolysis (Poulsen *et al.*, 1998). N-terminal sequence analysis of the *S. pneumoniae* strain PK81 IgA1 protease showed that it is cleaved at the Ala^{153}–Leu^{154} peptide bond, just after the last hydrophobic region, presumably by the signal peptidase enzyme (unpublished results). Cleavage of the *S. mitis* SK564, *S. mitis* SK609 and *G. haemolysans* CCUG 4815 IgA1 proteases after the corresponding alanine would result in secreted proteases with molecular masses of 213 415 Da with a pI of 5.01, 209 722 Da with a pI of 5.13, and 205 586 Da with a pI of 5.1, respectively. The proposed processing is in agreement with results from the activity blot experiment, which showed that the largest species of secreted proteins with human IgA1-cleaving activity in each of the three strains had an apparent molecular mass of ~220 kDa.

Following the putative cleavage site, the three IgA1 proteases contained 18–23 imperfect repeats, consisting of repeat sequences 16–20 aa in length. Repeat structures are commonly found in surface-exposed and secreted proteins in Gram-positive bacteria, and variation in the number and sequence of repeats may contribute to antigenic variation and thereby to immune evasion (Kehoe, 1994). The repeat structure in the *S. sanguinis* IgA1 protease has been found to be immunodominant and non-essential for enzyme activity (Gilbert *et al.*, 1993).

In contrast to that of *G. haemolysans*, all the streptococcal IgA1 proteases harboured a sequence with similarity to the G5 domain, which has been suggested to be a lectin that binds N-acetylglucosamine (Bateman *et al.*, 2005). The sequences contained four of the five conserved glycine residues in the consensus G5 domain (Bateman *et al.*, 2005). This domain was located within the otherwise highly variable repeat structure, suggesting that it has an essential function. In *S. pneumoniae*, the greater part of the IgA1 protease is non-covalently bound to the cell wall (unpublished results). Further experiments are needed to identify the potential ligand of this domain.

**Southern blot analysis**

We previously found that the *iga* gene sequences of *S. sanguinis* and *S. oralis* are highly conserved within each of the two species, whereas Southern blot analysis has revealed extensive gene polymorphism among *S. pneumoniae* and *S. mitis* strains (Poulsen *et al.*, 1996, 1998). Hakenbeck *et al.* (2001) have also shown variation in the *iga* gene among *S. pneumoniae* strains. Here, we used *iga* gene probes from *G. haemolysans* CCUG 4815, *S. mitis* SK564 and *S. pneumoniae* TIGR4 in Southern blot analyses to assay for the degree of similarity in *iga* genes within and between the species *G. haemolysans*, *S. mitis*, *S. oralis* and *S. infantis*. The two probes from each of the three strains represented equivalent areas flanking the active site (encoding amino acid positions 1388–1552 and 1859–1968 in the *G. haemolysans* CCUG 4815 IgA1 protease; Fig. 4). An example of the Southern blot analysis is shown in Fig. 5 and the results are compiled in Table 1. The *iga* genes, including the areas covered by the probes, are supposed to be homologous. Lack of hybridization for some of the probe/genomic DNA combinations under the stringency conditions used suggests lack of significant similarity in the DNA sequences. For example, the 490 bp SK564-1 probe sequence showed only 74 % similarity to the *G. haemolysans* CCUG 4815 *iga* gene in a 132 bp segment without longer stretches with identity, and the 611 bp SK564-2 probe had no significant similarity to the gemella *iga* gene. A potential complication in interpreting the hybridization results is the presence in *S. pneumoniae* and *S. mitis* genomes, and presumably also in some of the other species, of paralogues of the *iga* gene with varying degrees of homology (Kilian & Reinholdt, 2005). However, the specificity of the hybridization is supported by the fact that *S. mitis* strain SK142 (NCTC 12261), which lacks the *iga* gene but has two paralogous

**Fig. 4.** Schematic structure of the streptococcal and *G. haemolysans* IgA1 proteases. Black boxes indicate hydrophobic regions proposed to serve as transmembrane domains in the signal peptide and in combination with the cell-wall-anchor motif LPXTG. S1 and S2 indicate sites presumably cleaved by the signal peptidase. The G5 domain conserved in the streptococcal IgA1 proteases and location of the amino acids encoded by the two probes used in the Southern blot analysis are indicated.
The intensity of hybridization with the different probes was very similar for the six strains of *G. haemolysans*, indicating that the *iga* gene is highly conserved within this species. The same applied to the four strains of *S. oralis*. The eight strains of *S. mitis* with IgA1 protease activity showed varying degrees of hybridization, with the *iga* gene probes indicating significant sequence diversification of the *iga* gene within this species. This is supported by the high degree of dissimilarity between the two *S. mitis* *iga* genes sequenced in this study (see below). Notably, strong hybridization with the *G. haemolysans* *iga* gene was common among the *S. mitis* strains. The sequence analysis and the Southern blotting show that the *G. haemolysans* *iga* gene is very similar to the *iga* gene found in certain *S. mitis* strains. We also included four strains of the species *S. infantis* (Kawamura et al., 1998), three of which produced IgA1 protease activity. None of the *S. infantis* strains showed hybridization to any of the probes, indicating that the IgA1 protease-positive strains of this species have a more divergent *iga* gene sequence. Though highly unlikely, it is a theoretical possibility that the *S. infantis* *iga* gene may not be evolutionarily related and therefore not homologous to *iga* genes of the other streptococci and *G. haemolysans*.

**Phylogenetic considerations**

The genera *Streptococcus* and *Gemella* both belong to the Gram-positive bacteria with a low G+C content. Phylogeny based on 16S rRNA sequences indicates that *Gemella* is more closely related to the *Bacillus–Staphylococcus* complex than to streptococci and lactobacilli. IgA1 protease with significant sequence similarity to that of *G. haemolysans* described here has been exclusively found among *Streptococcus* species of the mitis group (Kilian et al., 1996; Kilian & Reinholdt, 2005). A likely explanation of the distribution of this finding is horizontal gene transfer, rather than vertical inheritance from a common ancestor combined with subsequent loss of the gene in some species. Deviation in G+C content of a gene compared to the whole genome is often a valuable marker for identifying genes acquired recently by horizontal transfer. The G+C content of the *G. haemolysans* *iga* gene was 34 % which is very similar to the 33–35 % reported for the genome (Reyn, 1986). The streptococci displaying IgA1 protease activity have a genome G+C content of 38–46 % and the content for their *iga* genes varies from 34 to 41 %.

**Table 1. Patterns of hybridization with *iga* gene probes**

<table>
<thead>
<tr>
<th>Species</th>
<th>No. of strains</th>
<th>IgA1 protease activity</th>
<th>Probe used for hybridization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>CCUG4815-1</td>
</tr>
<tr>
<td><em>G. haemolysans</em></td>
<td>6</td>
<td>+</td>
<td>S</td>
</tr>
<tr>
<td><em>S. mitis</em></td>
<td>1</td>
<td>–</td>
<td>N</td>
</tr>
<tr>
<td><em>S. mitis</em></td>
<td>8</td>
<td>+</td>
<td>S/W</td>
</tr>
<tr>
<td><em>S. oralis</em></td>
<td>4</td>
<td>+</td>
<td>N</td>
</tr>
<tr>
<td><em>S. infantis</em></td>
<td>1</td>
<td>–</td>
<td>N</td>
</tr>
<tr>
<td><em>S. infantis</em></td>
<td>3</td>
<td>+</td>
<td>N</td>
</tr>
</tbody>
</table>
However, because *Gemella* and *Streptococcus* have similar low G + C content, it is not possible from these data to draw conclusions on the origin of the iga gene. Several other observations support the hypothesis that the iga gene is transferred from *Streptococcus* to *G. haemolysans*. The iga gene is apparently ancient within the mitis group of the genus *Streptococcus*. This is supported by its presence in several closely related species and by the conserved gene order in the iga gene region of *S. pneumoniae* and *S. mitis*, which is different from that in *G. haemolysans* (Fig. 2). An ancestral streptococcus with an iga gene is also supported by the significant sequence diversification resulting in species-specific clusters of iga gene sequences (Fig. 6). In the genus *Gemella*, only *G. haemolysans* displays an iga gene, and the Southern blot analysis indicates that the gene is highly conserved within this species. This argues strongly for the transfer of the iga gene from a mitis group *Streptococcus* to the ancestor of *G. haemolysans* after its separation from other *Gemella* species.

A subsequent, evolutionarily recent, homologous horizontal gene transfer has apparently taken place between *S. mitis* and *G. haemolysans* since the *S. mitis* SK609 and *G. haemolysans* CCUG 4815 iga genes showed 89 % identity in a 2-8 kb DNA sequence encoding from the end of the repeats down to the active site, whereas the areas flanking this region were more different in the two bacteria. The mosaic structure of the mutual homologies among the IgA1 proteases is shown in Fig. 6(b). This was reflected in the Southern blot hybridizations. The CCUG4815-1 probe located upstream of the active site hybridized strongly to *S. mitis* SK609 genomic DNA, whereas the CCUG4815-2 probe downstream of the active site only showed weak hybridization to SK609 (Fig. 4, Table 1). This pattern of hybridization with the *G. haemolysans* iga gene probes was detected in five of eight IgA1 protease-positive strains of *S. mitis*, suggesting that the SK609 iga gene version is common among members of this species.

**Concluding remarks**

The sequence of the *G. haemolysans* IgA1 protease was homologous to streptococcal IgA1 proteases that belong to a family of zinc metallopeptinases found in the mitis group of *Streptococcus*. The genus *Gemella* is only distantly related to *Streptococcus* and we argue that the iga gene in *G. haemolysans* was acquired from *Streptococcus* by horizontal gene transfer. Structural features conserved among members of this family of IgA1 proteases may delineate sequences essential for enzymic function, including mechanisms of secretion and substrate specificity. Heterologous expression in *E. coli* allows for purification and manipulation and therefore further studies of these intriguing proteases.

**ACKNOWLEDGEMENTS**

Financial support was obtained from the Danish Medical Research Council grant 22-01-0265 ct/mp and Aarhus University Research Foundation. L.M.M. and K. R. D. both received a Novo student
REFERENCES


