Conserved and variable regions in protein Arp, the IgA receptor of *Streptococcus pyogenes*

LARS-OLOF HEDÉN and GUNNAR LINDAHL

Departments of Microbiology and Medical Microbiology, University of Lund, Söhlvegatan 21, S-223 62 Lund, Sweden

(Received 29 March 1993; revised 10 May 1993; accepted 24 May 1993)

The streptococcal M protein family, a number of cell surface molecules that interact with the human immune system, can be divided into two major classes, A and C, characterized by different types of repeats in the central part of the molecule. Class A and class C molecules are known to have a variable N-terminal region and a more conserved C-terminal region, but little is known about the mechanisms that give rise to this structural variation. In this report, we show that two variants of protein Arp, an IgA receptor in class C of the M protein family, have virtually identical signal sequences and C-terminal halves, but unrelated N-terminal sequences. Comparison of the sequences of the two genes and their flanking regions also demonstrates the presence of well-defined variable and conserved regions. Our results strongly suggest that the N-terminal sequence variation between the two variants of protein Arp was generated through an intergenic recombination event, rather than through intragenic recombination or accumulation of mutations.

Introduction

Antigenic variation is an important mechanism by which micro-organisms alter the structure of their surface, thereby escaping the immune system of the infected host (Birkbeck & Penn, 1986; Borst, 1991). Analysis of several different systems has demonstrated that such structural variation takes two different forms: antigenic drift that occurs by the gradual accumulation of mutations, and antigenic shift that occurs by an abrupt gene rearrangement (Borst, 1991). The latter mechanism often involves recombination, either with another sequence in the same genome or with exogenous homologous DNA that is taken up by transformation (Borst, 1991; Meyer, 1990).

An important human pathogen that exhibits antigenic variation is *Streptococcus pyogenes*, the group A Streptococcus. This bacterium expresses an antiphagocytic cell surface molecule, the M protein, that is a target for the immune response and occurs in numerous antigenic variants expressed by different strains (Lancefield, 1962; Fischetti, 1989; Scott, 1990). Sequence analysis of several different M proteins has demonstrated that they are structurally related and are therefore encoded by a family of genes. In particular, all the known M proteins have homologous C-terminal parts, including the centrally located C repeat region (Fischetti, 1989; Scott, 1990). The N-terminal parts of the different M proteins vary in sequence, which explains the antigenic variation (Fischetti, 1989; Miller et al., 1988; Scott, 1990), but the mechanism by which this variation is generated is not clear.

Sequencing of the genes for several Ig-binding cell surface proteins of *S. pyogenes* has shown that these molecules are also members of the M protein family (Fritzh et al., 1989; Heath & Cleary, 1989; Gomi et al., 1990; Bessen & Fischetti, 1992; O'Toole et al., 1992). Some of these Ig-binding proteins have C repeats, like the classical M proteins, while others have a different type of repeat, the A repeat, in the central part of the sequence. The M protein family can therefore be divided into two major classes, A and C, characterized by A repeats and C repeats, respectively (O'Toole et al., 1992). Both class A and class C molecules exhibit N-terminal sequence variation and both classes of molecules comprise proteins that interact with the immune system of the host (O'Toole et al., 1992). However, it is not yet known whether any of the Ig-binding proteins in class A and class C have antiphagocytic function, like the classical M proteins in class C.

One Ig-binding protein that has been extensively

---

* Author for correspondence. Tel. 46 108630; fax 46 157839; e-mail Lars-Olof.Hed@mikrobioi.lu.se.

The nucleotide sequence data reported in this paper have been submitted to GenBank/EMBL and have been assigned the accession number Z22751.
studied is protein Arp, an IgA receptor expressed by some strains of *S. pyogenes* (Frithz et al., 1989). This protein has C repeats, like the antiphagocytic M proteins, and exhibits N-terminal sequence variation, as demonstrated by N-terminal amino acid sequencing of Arp4 and Arp60, two variants of protein Arp expressed by group A streptococcal strains of different serotype (Frithz et al., 1989; Lindahl & Åkerström, 1989). This similarity to the classical M proteins suggested that studies of protein Arp might shed new light on the mechanism of antigenic variation in the M protein family and prompted us to characterize the two variants of protein Arp in greater detail. The sequence of Arp60 has now been determined and compared with the previously reported sequence of Arp4 (Frithz et al., 1989). The results show that the distinction between a variable N-terminal region and a conserved C-terminal region is more clear-cut in protein Arp than in any of the other class C proteins analysed so far, and they strongly suggest that the N-terminal sequence variation between the Arp60 and Arp4 proteins reflects an intergenic recombination event.

**Methods**

*Preparation of DNA, restriction enzyme analysis and cloning.* The λ clone used in this study carries a 15 kbp insert from *S. pyogenes* AW43, an M60 strain (Lindahl & Åkerström, 1989). This clone expresses the arp60 gene, which encodes the IgA-binding protein designated Arp60. Phage stocks were prepared by confluent lysis on *Escherichia coli* LE392 as described by Arber et al. (1983). After polyethylene glycol precipitation and treatment with DNase and RNase, the phage particles were deproteinized using phenol and chloroform. The DNA was ethanol precipitated and purified by cesium chloride gradient centrifugation in the presence of ethidium bromide (Maniatis et al., 1982). Digestion with restriction enzymes and ligations were performed under conditions recommended by the supplier (Boehringer Mannheim). Analysis of restriction enzyme fragments was performed using 0.8% agarose gels in Tris/acetate buffer (Maniatis et al., 1982). Restriction enzyme fragments for subcloning in pUC18, pUC19, M13mp18 and M13mp19 (Yanisch-Perron et al., 1985) were isolated from low gelling temperature agarose (BioRad) as described (Crouse et al., 1983). Preparation of competent cells and transformation of *E. coli* JM83 (Yanisch-Perron et al., 1985) and JM103 (Messing, 1983) were carried out as described by Morrison (1979). Plasmid DNA was isolated using an alkaline lysis procedure (Ish-Horowicz & Burke, 1981). The plasmid DNA was further purified by caesium chloride gradient centrifugation in the presence of ethidium bromide (Maniatis et al., 1982). Southern blot analysis of restriction enzyme fragments was done using a standard protocol (Maniatis et al., 1982) with Biodyne A (BNGG) membranes (PALL Corp.). DNA probes were prepared by nick-repair using [α-32P]dCTP and a kit from Amersham. Oligonucleotide probes were labelled in the 5' end with [γ-32P]ATP and polynucleotide kinase (Maniatis et al., 1982).

*DNA sequencing.* Single stranded DNA from recombinant phages M13mp18 and mp19 was prepared as described by Messing (1983) and the nucleotide sequence of the inserts was determined by the dideoxy chain termination method (Sanger et al., 1977, 1980) using [γ-32P]dATP (deoxyadenosine 5'-[γ-32P]thiotriphosphate; Amersham) and Sequenase (Tabor & Richardson, 1987) (USB) following the protocol provided by the supplier. Part of the sequence was obtained from fragments cloned in pUC18 and pUC19 (Chen & Seeberg, 1985). Autoradiography was performed with Fuji X-ray film.

*Colony immunoblot and protein analysis.* The immunoblotting procedure using 125I-labelled IgA for detection of colonies expressing IgA binding protein has been described (Frithz et al., 1989). For isolation of protein Arp60, strain JM83 carrying the appropriate plasmid was grown in LB broth to early exponential phase and then subjected to osmotic shock (Anraku & Heppel, 1967; Russel & Model, 1982). The protein was purified from the osmotic shock lysate by affinity chromatography on IgA-Sepharose, as described (Frithz et al., 1989).

**Results and Discussion**

*Nucleotide sequence of the arp60 gene.* The structural gene for protein Arp60 was previously shown to be present in a 15 kbp insert of streptococcal DNA in a λ EMBL3 clone (Lindahl & Åkerström, 1989). Digestion of this insert with both SalI and HindIII generated a number of fragments, which were ligated to the plasmid vectors pUC18 and pUC19. After transformation into *E. coli* JM83, bacteria were tested for expression of IgA-binding activity using a colony immunoblot method. One pUC19 clone (pARP104) expressed IgA-binding and was shown to contain an insert composed of a SalI–HindIII fragment of 0.9 kbp and a 1.8 kbp HindIII fragment (Fig. 1).

Hybridization studies indicated that at least part of the arp60 gene was present on the 1.8 kbp HindIII fragment, since this fragment hybridized to a BgII fragment comprising most of the region encoding the previously sequenced arp4 gene (Frithz et al., 1989). Sequence analysis of the HindIII fragment indeed showed the presence of an open reading frame with extensive sequence identity to the C-terminal part of protein Arp4. However, the HindIII fragment did not contain sequences corresponding to a signal sequence and a promoter. This result suggested that the IgA-binding expressed by plasmid pARP104 was due to the expression of an incomplete arp60 gene lacking a segment in the 5' region, including the streptococcal promotor.

A Southern blot analysis of the λ clone used for this work showed that the 5' region of the arp60 gene was present on a 1.4 kbp Psrl fragment. This Psrl fragment was isolated and introduced into pUC18 to generate pARP103 (Fig. 1), which expresses IgA-binding. The sequence of the 1.4 kbp Psrl fragment was determined on both strands (Fig. 1) and was shown to contain a considerable overlap with the sequence in pARP104, as expected (Fig. 2). Combination of the two sequences demonstrated the presence of an open reading frame of 1206 nucleotides, encoding a protein of 402 amino acids with a calculated *M*₂ of 45 522. The identity of this
protein with Arp60 was confirmed by comparison with the previously determined N-terminal amino acid sequence of the processed form of the protein (Lindahl & Åkerström, 1989). The amino acid composition of Arp60 is similar to that of Arp4 and streptococcal M proteins (Fischetti, 1989; Frithz et al., 1989). Possible promoter regions, indicated in Fig. 2, are identical to the promoter regions proposed for several streptococcalemm genes (see Frithz et al., 1989). Downstream (27 bp) of the stop codon there is an inverted repeat with the potential to form a stem-loop structure, which could function as a transcription terminator (Rosenberg & Court, 1979).

Expression of the arp60 gene

The sequence of the arp60 gene described above was deduced from two different inserts of streptococcal DNA, none of which contains the completearp60 gene. To obtain expression of the whole gene inE. coli, aPstI-BglII fragment from pARP103 and a BglII-HindIII fragment from pARP104 (Fig. 1) were ligated to pUC18 digested withPstI and HindIII generating pARP102. The structure of pARP102 was verified by cleavage with the appropriate restriction enzymes. Thearp60 gene was expressed in E. coli JM83 carrying pARP102, and it was also expressed when the fragment of streptococcalDNA
was inserted in the opposite direction in pUC19, which indicates that the arp60 gene is transcribed from an endogenous streptococcal promoter. Strains carrying pARP102 proved to be unstable, possibly due to a copy-number effect, as previously found for another S. pyogenes gene (Kehoe & Timmis, 1984). The arp60 gene was therefore recloned in pBR322 (Bolivar et al., 1977). For this purpose, the 5’ part of arp60 was obtained as an EcoRI–BglII fragment taking advantage of an EcoRI site in the vector, upstream of the cloned fragment in pARP103. This fragment was ligated to the BglII–HindIII fragment from pARP104 and EcoRI/HindIII digested pBR322, generating pARP107. Bacteria carrying pARP107 were not unstable. When the Arp60 protein expressed from pARP107 was isolated from osmotic shock lysates and analysed by Western blotting, using radiolabelled IgA as the probe, several closely grouped bands were seen both on the stained gel and on the blot (data not shown), as previously found for protein Arp4 and for streptococcal M proteins (Frithz et al., 1989; Scott & Fischetti, 1983). The Mr of the largest Arp60 species was estimated to about 41000, which is similar to that of Arp60 isolated from λ lysates (Lindahl & Åkerström, 1989).

**Comparison of proteins Arp60 and Arp4: N-terminal sequence variation**

From the deduced amino acid sequence, protein Arp60 can be divided into different domains as previously described for protein Arp4 (Frithz et al., 1989) and for M proteins (Fischetti, 1989). The Arp60 and Arp4 proteins are compared in Fig. 3. The position of the first residue in the processed forms of the proteins is known from amino acid sequencing work (Frithz et al., 1989; Lindahl & Åkerström, 1989). In both proteins, this residue is preceded by a signal sequence of 41 amino acid residues. These signal sequences are identical, except for a single amino acid residue (number 38). At the DNA level there are three nucleotide changes in this region, one of which is silent. Interestingly, the processing of protein Arp60 occurs in a region where a sequence of seven amino acids (TVKAESS) is tandemly repeated three times (Fig. 2).

The cleavage takes place in the first of these three repeat regions, and the processed form of the protein therefore starts with two identical seven-residue repeats (Lindahl & Åkerström, 1989). No such repeats are found in protein Arp4.

The processed form of protein Arp60 starts with a sequence of 51 amino acids which is different from the corresponding region in protein Arp4. There is no amino acid sequence homology between the Arp60 and Arp4 proteins in this part of the proteins and no obvious sequence similarity at the DNA level. The 69 amino acid residue region following this unique part of Arp60 shows 80% identity to the corresponding part of protein Arp4, and the corresponding regions of the two genes show 94% sequence identity. Of eleven base changes in this region, ten result in an amino acid exchange, which strongly suggests that an evolutionary pressure has selected for variation in this region (Fig. 4). The remaining part of protein Arp60, i.e. the C-terminal half of the protein, is almost identical to the corresponding...
region in protein Arp4, and the corresponding nucleotide sequences are also virtually identical. Except for a 21 nucleotide deletion in the C1 region of the arp4 gene, there are only eleven divergent nucleotide residues (of which eight are silent) in the 3'-terminal half of the two genes, which has a length of 723 nucleotides in arp60. At the protein level, the region of identity in the C-terminal half starts with the three C-repeats (amino acid residues 121–246). The C repeats in Arp60 and Arp4 are compared in Fig. 5, which shows that the C1, C2 and C3 repeats in protein Arp60 show a higher degree of homology to their counterparts in Arp4 than they do internally. This indicates that the C repeat region evolved in a common ancestor to arp4 and arp60.

The striking conservation of the C-terminal region of protein Arp strongly suggests that the N-terminal sequence variation between Arp60 and Arp4 is the result of an intergenic recombination event. The origin of new variable regions and the nature of the postulated recombination event is not known, but it seems possible that recombination takes place after horizontal transfer of DNA between two streptococcal cells. In this context it should be noted that several other streptococcal cell surface proteins (the M49, H and emmL2 proteins) have the same overall structure as protein Arp (Haanes et al., 1989; Gomi & Cleary, 1989; Jeppson et al., 1992). The C-terminal parts of these proteins are homologous to the C-terminal parts of Arp4 and Arp60, but the N-terminal parts of the processed proteins do not show any homology to the Arp molecules. A similar pattern is seen when the sequences of the corresponding genes are compared. It therefore seems possible that the genes encoding these different proteins can recombine with the arp genes after transfer of DNA from one streptococcal cell to another.

The numbers of residues found in the C repeats of proteins Arp60 and Arp4 are multiples of seven (42 amino acid residues except for C1 in Arp4 which is 35 amino acid residues long). This finding supports the conclusion that the distribution of amino acid residues in protein Arp exhibits seven-residue periodicity, as in streptococcal M proteins (Fischetti, 1989; Frithz et al., 1989).

### Conserved sequences upstream and downstream of the arp60 gene

Both the arp4 and the arp60 genes are flanked by open reading frames that have been sequenced or partially sequenced (Fig. 6). In the M4 strain, the arp4 gene is located immediately downstream of the mrp4 gene, which encodes a protein in class A of the M protein family (O’Toole et al., 1992). Downstream of the arp4 gene there is an open reading frame, enn4, from which one can deduce a protein in class C of the M protein family (Jeppson et al., 1992). However, this sequence is probably not expressed at the protein level (Jeppson et al., 1992). A similar arrangement is found in the M60 strain, where the arp60 gene is closely linked to the mrp60 gene (Stenberg et al., 1992) and to the open reading frame enn60, which probably is silent (L.-O. Hedén, unpublished). It has been suggested that the silent enn genes can recombine with the homologous arp genes, thereby giving rise to structural variation (Haanes & Cleary, 1989; Jeppson et al., 1992). However, recombination with the enn gene in the same chromosome cannot explain the N-terminal sequence variation between Arp4 and Arp60, since the N-terminal sequences of the putative Enn proteins are unrelated to the two Arp proteins. However, if an arp gene is horizontally transferred from another cell, it might be incorporated into the chromosome of the recipient by recombination in homologous regions, causing expression of a new variant of protein Arp in the recipient. To analyse the potential for such homologous recombination in the Arp region, it was of interest to characterize the intergenic regions flanking the arp genes.

We determined the sequences of the intergenic regions upstream and downstream of the arp60 gene (Fig. 1) and compared them with the corresponding sequences around the arp4 gene, which are known from previous work (Jeppson et al., 1992; O’Toole et al., 1992). The results (Fig. 6) show that there are only five divergent residues and a deletion of one base in the region upstream of arp60, and that the intergenic region downstream is identical in the two systems. The variable parts of the arp genes are therefore flanked by long sequences that are

---

### Table: Amino acid sequences of the C-repeats in proteins Arp4 and Arp60

<table>
<thead>
<tr>
<th>Protein</th>
<th>Amino acid sequences of the C-repeats</th>
<th>Residue number in the processed protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arp60</td>
<td>LEAEKQLAKEKQI3SDASRQGLSLDNLASRAAKEKLEAEKQK</td>
<td>121–162/C1</td>
</tr>
<tr>
<td>Arp4</td>
<td>T--H--KE--E--KV--DLAA</td>
<td>112–146/C1</td>
</tr>
<tr>
<td></td>
<td>T--H--KE--E--KV--DLAB</td>
<td>147–188/C2</td>
</tr>
<tr>
<td></td>
<td>E--KE--DLAB</td>
<td>189–230/C3</td>
</tr>
</tbody>
</table>

Fig. 5. Comparison of the C repeat sequences in proteins Arp4 and Arp60. The dashes represent identical amino acid residues. The data for protein Arp60 are from Fig. 2 and those for protein Arp4 are from Frithz et al. (1989).
Fig. 6. Nucleotide sequences of the intergenic regions upstream and downstream of the \textit{arp4} and \textit{arp60} genes. The intergenic regions shown here include the postulated stop and initiation codons of the surrounding genes. The dashes represent identical nucleotide residues. The \textit{mrp4} and \textit{mrp60} genes encode proteins in class A of the M protein family (O'Toole et al., 1992). The \textit{mrp4} gene is known to be located immediately upstream of the \textit{arp4} gene (O'Toole et al., 1992), but it has not yet been conclusively shown that the \textit{mrp60} gene is located immediately upstream of the \textit{arp60} gene, as indicated by the dashed line. \textit{enn4} and \textit{enn60} are open reading frames from which one can deduce proteins in class C of the M protein family, but they are probably not expressed (Haanes & Cleary, 1989; Jeppson et al., 1992; L.-O. Heden, unpublished).

**Concluding remarks**

The sequence comparison of Arp60 and Arp4 reported here unequivocally shows that the processed form of protein Arp can be divided into a variable N-terminal region and a conserved C-terminal region. Comparison of the corresponding genes also shows that the \textit{arp} gene has a well-defined variable region, which is surrounded by conserved regions. Upstream, the conserved region corresponds to the sequence between the \textit{mrp} and \textit{arp} genes and to the sequence encoding the signal peptide. Downstream, the conserved region covers most of the \textit{arp} gene and the entire sequence between the \textit{arp} and \textit{enn} genes (Fig. 6).

Jones et al. (1985) first postulated that the streptococcal M proteins have variable N-terminal and conserved C-terminal parts, and suggested that the structural variation in the N-terminal region is the result of a selective pressure that causes this part of the molecule to evolve faster in response to immunological selection by the host. This conclusion, which was made before extensive sequence data were available, was based on analysis of the reactivity of a set of monoclonal antibodies. Sequencing work subsequently led to the suggestion that the variable regions might be inserted by homologous recombination (Miller et al., 1988; Haanes-Fritz et al., 1988). Our results support the hypothesis that recombination between two different genes contributes...
to the generation of structural variability in the M protein family, since it seems very likely that the N-terminal sequence variation between Arp60 and Arp4 is the result of a recombination event. Although it cannot be formally excluded that the N-terminal variation between Arp60 and Arp4 is the result of gradual accumulation of mutations, it would be difficult to reconcile such an explanation with the almost complete sequence conservation in the C-terminal part of the protein (and the corresponding part of the gene). The mechanism of the postulated recombination event is not understood, but it seems possible that new variable regions could be acquired from exogenous sources through transduction, transposition or transformation (Haanes-Fritz et al., 1988; Laibe et al., 1991; Maynard Smith et al., 1991).

Although the data discussed above suggest that intragenic recombination is an important mechanism for the generation of N-terminal sequence variation in the M protein family, it seems likely that accumulation of mutations also contributes to the structural variation (Fig. 4; see also Jones et al., 1985). Furthermore, there is evidence that intragenic recombination between short repeated regions in the 5′ part of the emm6 gene can give rise to antigenically distinct variants of the M6 protein (Jones et al., 1988; Fischetti, 1989). Taken together, these various data indicate that structural variation in the M protein family arises by several different mechanisms.

We are indebted to Miss E. Frithz for valuable contributions during the initial stages of this work. This work was supported by grants from the Swedish Medical Research Council (projects 9096 and 9490), the Medical Faculty of the University of Lund, the Royal Physiographic Society in Lund, the Foundations of Bergvall, Cornell, Crafords, Kock and Osterlund, by Hightech Inc. and by King Gustaf V’s 80-year foundation.

References


