JOURNAL OF BACTERIOLOGY, June 1999, p. 3599–3605 Vol. 181, No. 11
0021-9193/99/$04.00+0

Analysis of the 5′ Portion of the Type 19A Capsule Locus Identifies Two Classes of *cpsC*, *cpsD*, and *cpsE* Genes in *Streptococcus pneumoniae*

JUDY K. MORONA,[1] RENATO MORONA,[2] AND JAMES C. PATON[1]*

Molecular Microbiology Unit, Women's and Children's Hospital, North Adelaide, South Australia 5006,[1] *and Department of Microbiology and Immunology, University of Adelaide, Adelaide, South Australia 5005,*[2] *Australia*

Received 25 January 1999/Accepted 22 March 1999

Analysis of the sequence data obtained from the 5′ portion of the *Streptococcus pneumoniae* type 19A capsular polysaccharide biosynthesis locus (*cps19a*) revealed that the first seven genes are homologous to the first seven genes in the type 19F (*cps19f*) locus. The former genes were designated *cps19aA* to *-G* and were 70 to 90% identical to their *cps19f* counterparts. Southern hybridization analysis of the *cps* loci from various *S. pneumoniae* serotypes with probes specific for the *cps19aC*, *cps19aD*, and *cps19aE* genes indicated a hybridization pattern complementary to that previously reported for *cps19fC*, *cps19fD*, and *cps19fE*. That is, all serotypes tested contained high-stringency homologues of either the *cps19aC* to *-E* genes or the *cps19fC* to *-E* genes, but not both. On this basis *S. pneumoniae cps* loci can be divided into two distinct classes. Long-range PCR was used to amplify the *cps* regions between *cpsB* and *aliA* from a variety of pneumococcal serotypes. Direct sequencing of the 5′ end of these PCR products, and phylogenetic analysis of the sequence data, confirmed the presence of the two distinct classes of *cpsC*. Whereas members within one class are greater than 95% identical to each other, the DNA sequence identity between the two classes is only approximately 70%.

Streptococcus pneumoniae (the pneumococcus) is an important cause of invasive disease in human populations throughout the world, resulting in high morbidity and mortality. Control of pneumococcal disease is being complicated by the increasing prevalence of antibiotic-resistant strains and the suboptimal clinical efficacy of existing vaccines. *S. pneumoniae* produces a polysaccharide capsule which is essential for virulence because it protects the pneumococcus from the nonspecific immune defenses of the host during an infection (2).

There are now 90 recognized serotypes of *S. pneumoniae* (9), each of which produces a structurally distinct capsular polysaccharide (CPS). Classical genetic studies carried out by Austrian et al. (3) demonstrated that the *S. pneumoniae* genes required for biosynthesis and expression of CPS are closely linked on the pneumococcal chromosome. This fact enabled us to clone and sequence the complete capsule locus from *S. pneumoniae* type 19F (designated *cps19f*) (8, 17). Our studies were concentrated on *S. pneumoniae* type 19F because it is one of the commonest causes of invasive disease in children, and the type 19F CPS is one of the poorest immunogens in this group (6). Type 19F belongs to serogroup 19, which also contains the immunologically cross-reactive types 19A, 19B, and 19C. *S. pneumoniae* type 19A is also an important cause of disease, whereas types 19B and 19C are rare causes of disease (25).

We have previously examined the distribution of individual *cps19f* genes among other pneumococcal serotypes, including the other members of serogroup 19, by Southern hybridization analysis (17). Only homologues to *cps19fA* and *-B*, the first two genes in the *cps* locus, were present in all serotypes examined. Cps19fA is a putative transcriptional attenuator, but the function of Cps19fB is unknown. The next two genes in the *cps* locus, *cps19fC* and *-D*, encode proteins which are predicted to be involved in chain length regulation and export of CPS (8, 17). Moreover, Cps19fC and -D are essential for CPS expression in *S. pneumoniae* type 19F, as in-frame deletion mutations in either *cps19fC* or *cps19fD* result in the loss of CPS production (16a). Thus, *cps19fC* and *-D* homologues are probably essential for CPS production in all *S. pneumoniae* serotypes which are synthesized via lipid-linked repeat unit intermediates in a fashion similar to type 19F CPS. To date, this would include all pneumococcal serotypes which have been characterized except type 3, which is synthesized by a processive transferase (1, 5, 21). Surprisingly, however, 10 of the 21 serotypes tested in previous hybridization studies, including type 19A, did not contain high-stringency homologues of *cps19fC* and *-D*.

The structures of the CPS for types 19F and 19A are almost identical, consisting of the same rhamnose→*N*-acetyl mannosamine→glucose trisaccharide repeat units joined by different glycosidic linkages (α[1→2] for 19F and α[1→3] for 19A) (10, 20). Thus the only predicted functional difference between the protein products expressed by the *cps19f* and *cps19a* loci would be that of the polysaccharide polymerase. However, the type 19A *cps* locus appears to be more divergent, with high-stringency homologues of only eight of the *cps19f* genes present, compared to homologues of 13 out of the 15 *cps19f* genes present in types 19B and 19C (17). This study investigates the basis for this apparent diversity.

Bacterial strains. *S. pneumoniae* Rx1-19F has been described previously (8). A clinical isolate of *S. pneumoniae* type 19A, strain 1777/39, was obtained from Jorgen Henrichsen, Statens Seruminstitut, Copenhagen, Denmark. All other *S. pneumoniae* strains were clinical isolates from the Women's and Children's Hospital, North Adelaide, Australia. Pneumococci were routinely grown either in Todd-Hewitt broth (Oxoid Limited, Basingstoke, England) supplemented with 0.5% (wt/vol) yeast extract (Difco Laboratories, Detroit, Mich.) or on blood

* Corresponding author. Mailing address: Molecular Microbiology Unit, Women's and Children's Hospital, North Adelaide, SA 5006, Australia. Phone: 61-8-82046302. Fax: 61-8-82046051. E-mail: patonj@wch.sa.gov.au.

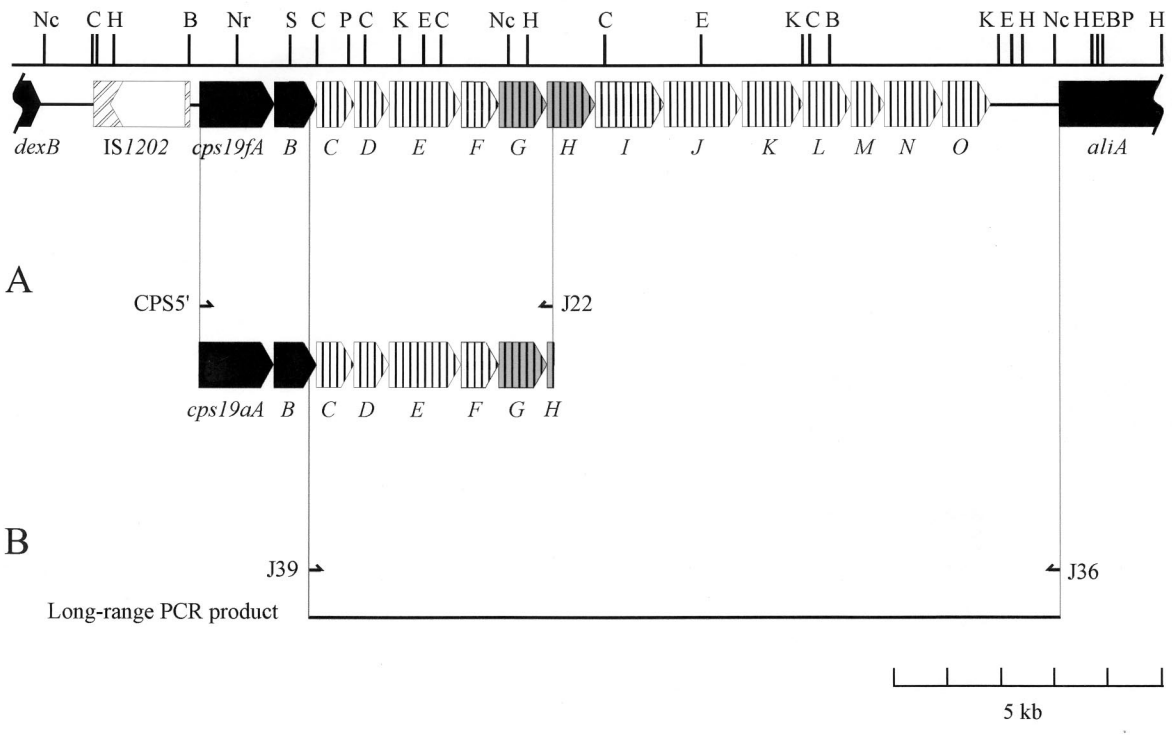

FIG. 1. The *dexB*-to-*aliA* region of the *S. pneumoniae* type 19F chromosome. The regions conserved among all pneumococcal serotypes are indicated in black, the *cps19G* and *-H* genes are shaded grey to indicate a higher degree of identity between *cps19f* and *cps19a* in this region. (A) Arrangement of the *cps19a* locus between the CPS5′ and J22 primers. CPS5′ (5′-TGATGTTCAAGGTATAGGTGTTAATCA) is homologous to nucleotides 146 to 169 of the *cps19f* sequence (17), immediately preceding the *cps19fA* gene, and J22 (5′-AATTGAATTCTTTTATAGATTTAACACAAG) is complementary to nucleotides 6743 to 6772 of the *cps19f* sequence, in the 5′ region of *cps19fH*. (B) The portion of the *cps* locus between *cpsB* and *aliA* from various pneumococcal serotypes was amplified by using the two primers J39 and J36. The positions of the two primers J39 (5′-TAGTTCATGTAGTTGCAAGTGACATGCACAA, homologous to nucleotides 2190 to 2220 of the *cps19f* sequence, in the 3′ region of *cps19fB*) and J36 (5′-CAATAATGTCACGCCCGCAAGGGCAAGT, complementary to nucleotides 16463 to 16490 of the *cps19f* sequence, located just after the start of *aliA*) are indicated with half arrows. Abbreviations for restriction sites are as follows: B, *Bam*HI; C, *Cla*I; E, *Eco*RI; H, *Hin*dIII; K, *Kpn*I; Nc, *Nco*I; Nr, *Nru*I; P, *Pst*I; S, *Sph*I.

agar (Oxoid) and serotyped by the quellung reaction using sera obtained from the Statens Seruminstitut.

Characterization of the 5′ portion of *cps19a*. Genes homologous to *cps19fA*, *-B*, *-G*, and *-H* were predicted to be present in the *cps19a* locus based on previous Southern hybridization data obtained with the *cps19f* genes as probes (17). Thus, the 5′ portion of *cps19a* was amplified by long-range PCR using the Expand Long Template PCR system (Boehringer, Mannheim, Germany), according to the manufacturer's instructions, and was performed in a Hybaid Touchdown Thermal Cycler. The two primers used to amplify this region (CPS5′ and J22) (Fig. 1A) were based on regions of the *cps19f* sequence which are predicted to be homologous to the *cps19a* locus. The resultant PCR product was sequenced by using dye-terminator chemistry with specifically designed primers on an Applied Biosystems model 373A automated DNA sequencer. The sequences were analyzed with DNASIS software (version 7.0; Hitachi Software Engineering, South San Francisco, Calif.). Analysis of the compiled DNA sequence revealed that the *cps19f* and *cps19a* loci are indeed very closely related. There are seven open reading frames (ORFs) in this portion of the *cps19a* locus, designated *cps19aA* to *-G*, which are organized in an order identical to those in *cps19f*, with similarities to the *cps19f* genes ranging from 70.1 to 90.9% identity. The sizes,

TABLE 1. Comparison of *cps19a* and *cps19f* ORFs

cps19a ORF	Predicted protein product		%G+C[a]	*cps19f* ORF	Predicted protein product		%G+C	% Identity[c]	
	Molecular mass (Da)	No. of aa[b]			Molecular mass (Da)	No. of aa		DNA	aa
cps19aA	53,576	481	39.5	*cps19fA*	53,572	481	38.1	90.5	92.3
cps19aB	28,138	243	41.3	*cps19fB*	28,352	243	38	82	85.2
cps19aC	25,473	230	42.1	*cps19fC*	25,497	230	38.2	70.1	71.7
cps19aD	25,155	229	41	*cps19fD*	24,947	227	34.5	73	80.2
cps19aE	51,971	453	37.7	*cps19fE*	52,595	455	33.2	71.2	70.5
cps19aF	28,273	247	34.1	*cps19fF*	28,155	247	33.6	78.9	82.9
cps19aG	31,195	266	37.2	*cps19fG*	31,647	269	36.3	90.9	93.6

[a] Percent guanine plus cytosine content of coding region.
[b] aa, amino acids.
[c] Percent identity between *cps19a* and *cps19f* ORFs.

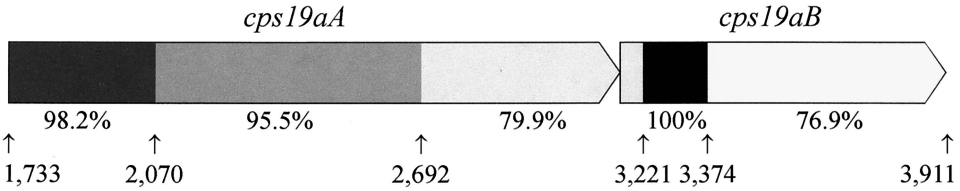

FIG. 2. Diagrammatic representation of the similarity of the *cps19aAB* genes to the *cps19fAB* genes. There are several possible recombination points in this region of the locus. Increasing similarity is represented by progressively darker shades of grey, and the percent identity is shown under the individual shaded regions. The arrows indicate the points of divergence, and the number below each arrow corresponds to the nucleotide number of the *cps19a* sequence.

G+C content, and percent identity of the *cps19a* and *cps19f* protein products are shown in Table 1, and the arrangement of the *cps19aA* to *-G* genes compared to those from *cps19f* is shown in Fig. 1A. The arrangement of the genes in this region of the two loci are remarkably similar; even the intergenic gaps between the *cps19a* genes and the *cps19f* genes are similar in size. The most significant difference between the two loci is the start codon used for *cps19G*. Whereas the start codon for both *cps19aG* and *cps19aH* is TTG, only *cps19fH* has a TTG start codon in the *cps19f* locus.

Interestingly, even though *cps19aA* and *-B* sequences hybridize to *cps19fA*- and *cps19fB*-specific probes (17), the overall identity between the genes is lower than expected (90.5 and 82%, respectively), with no clearly identifiable point from which downstream sequences diverge. Instead, the *cps19aAB* genes present a mosaic pattern with small regions of varying degrees of identity to the *cps19fAB* genes, ranging from 76.6 to 100%, as shown in Fig. 2. This suggests that the *cps19a* locus and the type 19A serotype may be the result of several recombination events between the ancestral *cps* locus and exogenous DNA. Some of these recombination events may have involved small DNA fragments that did not affect the serotype, while others resulted in the exchange of larger regions of the capsule locus, which may have altered the structure and hence serotype of the expressed CPS. A small region of *cps19aB* (nucleotides 3,221 to 3,374) has 100% identity to *cps19fB*. This region presumably accounts for the high-stringency hybridization of the *cps19aB* DNA to a *cps19fB* probe (17), as there is only 76.7% identity between the remainder of the *cps19aB* and *cps19fB* genes. The highly conserved region either may encode a functionally important domain in the *cps19B* gene product or may simply be the result of a recombination event.

Southern hybridization analysis. Previous Southern hybridization data have shown that high stringency homologues of *cps19fA* and *-B* are present in all serotypes tested, whereas *cps19fF* and *-G* are serogroup specific. However, high-stringency *cps19fC*, *-D*, and *-E* homologues were present in some serotypes tested but not others (17). The presence of homologues to the divergent *cps19aC*, *cps19aD*, and *cps19aE* genes in the *cps* loci of various *S. pneumoniae* serotypes was therefore examined by Southern hybridization. Digoxigenin (DIG)-labelled DNA fragments corresponding to the *cps19aC*, *cps19aD*, and *cps19aE* genes were used to probe, at high stringency, *Cla*I-restricted chromosomal DNA from representative pneumococci belonging to serotypes 2, 3, 4, 6A, 6B, 7F, 8, 9N, 9V, 12, 14, 16, 17, 18C, 19F, 19B, 19C, 20, 22, 23F, and 24. The hybridization data for the type 19A *cps19aC*, *cps19aD*, and *cps19aE* gene probes and previous data obtained for the type 19F *cps19fC*, *cps19fD*, and *cps19fE* gene probes (17) are compared in Table 2.

The most remarkable feature seen in Table 2 is that all the serotypes tested contained high-stringency homologues of either *cps19fC* to *-E* or *cps19aC* to *-E*, except types 3 and 4, which do not have a high-stringency homologue of either *cps19fE* or *cps19aE* (the gene encoding the glucose-1-phosphate transferase which catalyzes the addition of glucose-1-phosphate to the lipid carrier, a common first step in biosynthesis of the lipid-linked repeat unit [12, 17]). The absence of a *cpsE* homologue in types 3 and 4 is not surprising, because the type 4 CPS does not contain glucose, and the mode of type 3 CPS biosynthesis is atypical, occurring via a processive transferase (1, 5). Type 4 also contains a hybrid *cpsC* gene, hybridizing to both the *cps19fC* and the *cps19aC* probes, as described below. Thus, these Southern hybridization data suggest that *S. pneumoniae cps* loci can be divided into two distinct classes, designated class I and class II, where class I loci contain high-stringency *cps19fC* to *-E* homologues and class II loci contain high-stringency *cps19aC* to *-E* homologues.

Amplification of capsule loci by long-range PCR. In order to directly characterize the two classes of *cpsC* gene, long-range PCR was used to amplify the portion of the *cps* loci between *cpsB* and *aliA* (Fig. 1B) from several *S. pneumoniae* serotypes so that DNA sequencing could be undertaken. DNAs prepared from serotypes or groups 2, 4, 6A, 6B, 7F, 8, 9N, 9V, 12, 14, 16, 17, 18C, 19F, 19A, 19B, 19C, 20, 22, 23F, and 24 were used as

TABLE 2. Hybridization of *cps19fC* to *-E* and *cps19aC* to *-E* genes with other pneumococcal serotypes[a]

Serotype	DIG-labelled DNA probe					
	cps19fC	*cps19fD*	*cps19fE*	*cps19aC*	*cps19aD*	*cps19aE*
2	−	−	−	+	+	+
3	+	+	−	−	−	−
4	+	+	−	+	−	−
6A	−	−	−	+	+	+
6B	−	−	−	+	+	+
7F	+	+	+	−	−	−
8	−	−	−	+	+	+
9N	+	+	+	−	−	−
9V	−	−	−	+	+	+
12	−	−	−	+	+	+
14	+	+	+	−	−	−
16	+	+	+	−	−	−
17	−	−	−	+	+	+
18C	+	+	+	−	−	−
19F	+	+	+	−	−	−
19A	−	−	−	+	+	+
19B	+	+	+	−	−	−
19C	+	+	+	−	−	−
20	+	+	+	−	−	−
22	−	−	−	+	+	+
23F	−	−	−	+	+	+
24	+	+	+	−	−	−

[a] DNA fragments equivalent to nucleotides 3932 to 4356, 4356 to 4820, and 5594 to 6480 of the *cps19a* sequence were labelled with DIG and used as probes for the *cps19aC*, *-D*, and *-E* genes, respectively. The results for *cps19fC*, *cps19fD*, and *cps19fE* have been published previously (17).

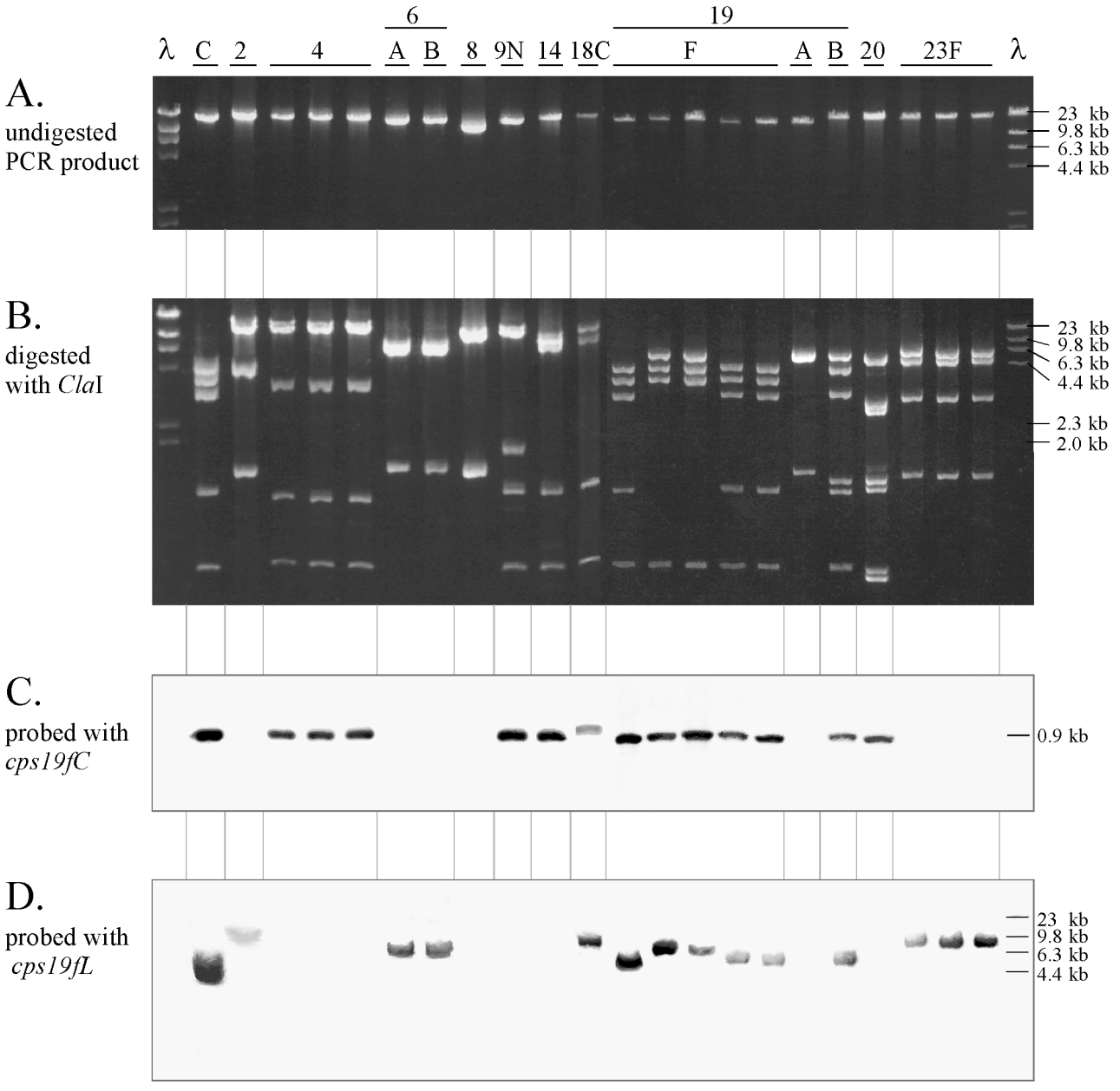

FIG. 3. Long-range PCR products. PCR products, not digested (A) or digested with *Cla*I (B), were electrophoresed on a 1% agarose gel in the presence of ethidium bromide. *Cla*I-restricted PCR product was subjected to Southern hybridization analysis using DIG-labelled probes specific for *cps19fC* (C) or *cps19fL* (D). The probes specific for *cps19fC* and *cps19fL* correspond to nucleotides 2380 to 2998 and 11539 to 12493 of the *cps19f* sequence (17). The molecular size standards are shown on the right-hand side of the figure and correspond to *Hin*dIII-digested λ phage DNA.

templates for long-range PCR. PCR products were obtained from at least one pneumococcal isolate of types 2, 4, 6A, 6B, 8, 9N, 14, 18C, 19F, 19A, 19B, 20, and 23F but not from types 7F, 9V, 12, 16, 17, 19C, 22, and 24. Analysis of the DNA fragments reveals that the PCR products ranged in size from 15 to 20 kb, as shown in Fig. 3A. The PCR products were digested with the restriction endonuclease *Cla*I and electrophoresed on a 1% agarose gel in a Tris-borate-EDTA (TBE) buffer system as described by Maniatis et al. (16) (Fig. 3B). Identical restriction patterns were obtained for three different isolates of serotypes 4 and 23F. However, a restriction site polymorphism was observed in two of the five PCR products from different type 19F strains (Fig. 3B).

Southern hybridization analysis of long-range PCR products. In order to confirm that they contained *cps*-related sequences, the long-range PCR products from the various *S. pneumoniae* serotypes were restricted with *Cla*I and subjected to Southern hybridization analysis using probes specific for two different type 19F gene probes, *cps19fC* (located in the 5′ region of the *cps19f* locus) and *cps19fL* (located in the 3′ region of the *cps19f* locus) (Fig. 3C and D).

The *cps19fC* probe hybridized at high stringency with a 0.9-kb DNA fragment in types 4, 9N, 14, 18C, 19F, 19B, and 20. Both the hybridization pattern and the size of the DNA fragment which hybridized with the *cps19fC* probe are consistent with the Southern hybridization data obtained when probing

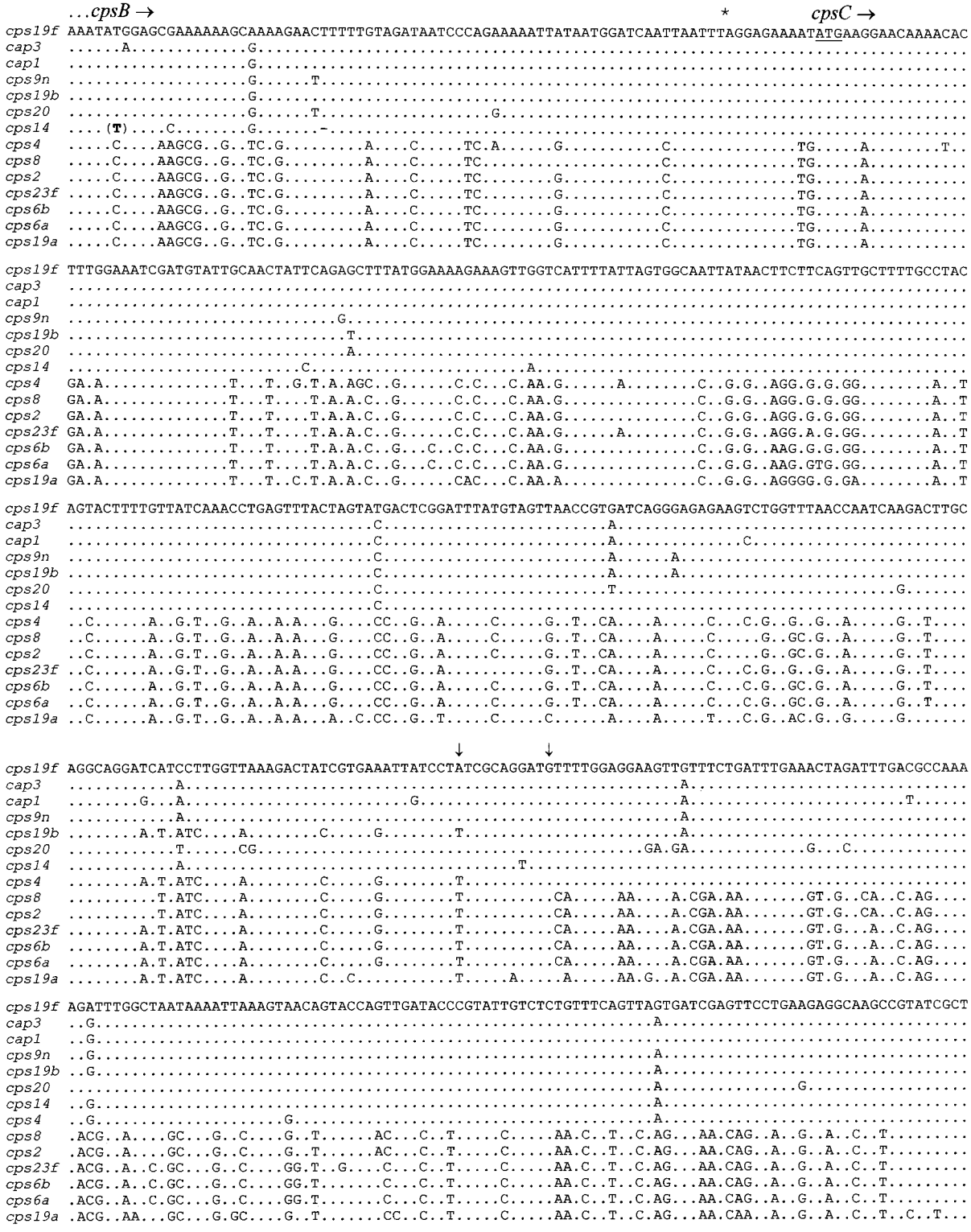

FIG. 4. Comparison of class I and class II *cps* sequences. The first 500 nucleotides of the sequence obtained are shown (100 nucleotides per line). Dots indicate nucleotides which are identical to that for *cps19f*. The stop codon of the *cpsB* gene is indicated with an asterisk. The start codon of *cpsC* is underlined. (**T**) denotes an extra nucleotide, and - denotes the absence of a nucleotide in the *cps14* DNA sequence. The vertical arrows indicate the region where the crossover between class I and class II sequences has occurred in *cps4*.

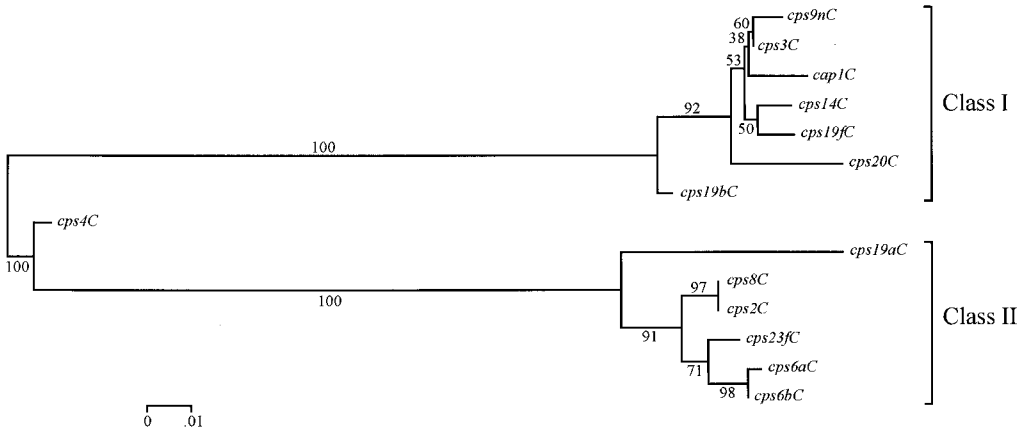

FIG. 5. Phylogenetic tree of *cpsC* sequences. The *cpsC* gene sequences were aligned by using CLUSTAL W (24), and the phylogenetic tree was generated by using MEGA (15), as described in the text. The numbers associated with the branches are bootstrapping confidence limits, resulting from 500 replications, as defined in MEGA. The scale represents the number of nucleotide substitutions per site.

*Cla*I-restricted chromosomal DNA with the *cps19fC* probe (data not shown).

The *cps19fL* probe hybridized with DNA fragments ranging in size from 4 to 10 kb in the *Cla*I-restricted PCR products from types 2, 6A, 6B, 18C, 19F, 19B, and 23F. Hybridization was consistent with that obtained from Southern hybridization with *Cla*I-restricted chromosomal DNA from these isolates, although the sizes of the restriction fragments differ (data not shown). The size of this *Cla*I fragment is affected because there is no *Cla*I site between *cps19fL* and the end of the PCR product in type 19F.

DNA sequencing of the 5′ portion of *cpsC* in the long-range PCR products. The long-range PCR products were subjected to one round of sequence analysis with the J39 primer (located at the 5′ end of the PCR product) in order to determine the presence of a *cpsC* homologue. No sequence data were obtained from the type 18C PCR template, presumably due to the low yield of the PCR product obtained. Analysis of the sequence data obtained from all the other templates and that available for types 1, 3, and 14 (1, 13, 18) showed that, indeed, there were two distinct *cpsC* genes in these loci. Types 1, 3, 9N, 14, 19F, 19B, and 20 have class I *cpsC* genes which exhibit >95% nucleotide sequence identity to *cps19fC*, whereas types 2, 6A, 6B, 8, 19A, and 23F have class II *cpsC* genes which exhibit 72 to 74% identity to *cps19fC* and >95% identity to *cps19aC* (Fig. 4). The sequences obtained from the PCR products also included the last 75 nucleotides of *cpsB*; this region can also be separated into the same two classes as described above (Fig. 4).

An interesting exception is found in type 4, the *cps4C* gene of which is a hybrid consisting of a class II 5′ region and a class I 3′ region, with a distinct crossover point in the vicinity of nucleotide 345 of the *cps4* sequence (Fig. 4). Comparison of the type 4 *cps* sequence data (available from the TIGR microbial database) with the *cps19f* sequence showed another point of divergence within the *cpsB* gene. The *cps4B* gene is almost identical (except for the first 42 nucleotides) to the *cps19aB* gene and shows the same point of sequence divergence from the *cps19fB* gene (nucleotide 3374 in Fig. 2). Thus, in the *cps4* locus a region of 852 nucleotides, including most of *cps4B* and part of *cps4C*, has approximately 74% identity to *cps19f*, whereas the remainder of the *cps4A* to *-D* region exhibits >95% identity to *cps19fA* to *-D*. This may have arisen as a consequence of recombination between a class I *cps* locus and a DNA fragment (approximately 852 nucleotides long) from a class II *cps* locus, resulting in a mosaic *cpsB-cpsC* region. Analysis of the available type 23F sequence data (4, 22) indicated that the class II *cps23f* locus also diverges from the class I *cps19f* locus within the *cpsB* gene, but 98 nucleotides further downstream from the point of divergence for *cps19a* and *cps4*. This suggests that the point of sequence divergence from class I to class II within *cpsB* may vary between different serotypes.

Phylogenetic analysis. To further confirm the presence of two distinct classes of *cpsC* sequences, their phylogenetic relationship was investigated. An alignment of the partial *cpsC* sequences was generated by using CLUSTAL W (24) (data not shown), and this alignment was used to generate a phylogenetic tree by the neighbor-joining method and the distance measure of Tamura and Nei (23), as implemented in the program MEGA (15). The tree in Fig. 5 shows two highly significant clusters of *cpsC* sequences (based on a bootstrapping test with 500 replications) and confirms the observations initially made on the basis of sequence homology that the *cpsC* genes are divided into two classes. The *cps4C* sequence forms a third cluster; as described above, this gene is a hybrid of the two *cpsC* classes and has a recombination crossover point at or near nucleotide 345 (as shown in Fig. 4) within the *cpsC* gene. The *cps19bC* gene is also separated from the other class I *cpsC* sequences; *cps19bC* also appears to have a mosaic structure with a small region of class II sequence (nucleotides 409 to 444 in Fig. 4), which is presumably the result of a recombination event.

Conclusions. The 5′ portion of the *cps* locus from *S. pneumoniae* type 19A is similar to *cps19f*, in that it has the same number of genes arranged in the same order. However, many of these genes demonstrate only 70 to 80% nucleotide sequence identity with their *cps19f* counterpart, suggesting either that the two loci diverged a long time ago or that portions of these loci have separate origins. Some regions within the *cps19aA*, *-B*, and *-G* genes do have >90% identity to those in *cps19f*, which may be a consequence of recombination between the two loci or perhaps is due to a requirement for a higher degree of conservation in regions encoding functionally important domains.

Southern hybridization analysis identified two classes of *cpsC*, *cpsD*, and *cpsE* genes in *S. pneumoniae cps* loci, which were designated as either class I or class II. Class I pneumococcal *cps* loci contain high-stringency *cps19fC* to *-E* homo-

logues, whereas class II loci contain high-stringency *cps19aC* to *-E* homologues. Direct sequencing of the long-range PCR products obtained confirmed the presence of two classes of *cpsC* gene. Phylogenetic analysis of the sequence data also confirmed that the pneumococcal *cpsC* gene is divided into two closely related classes. The presence of the *cpsC* and *cpsD* genes in all *cps* loci examined is consistent with the important role of CpsC and CpsD in pneumococcal CPS production. Both are predicted to be involved in chain length regulation and export of the CPS (8, 17). At this stage, it is not possible to determine whether the differences between class I and class II *cpsC* and *cpsD* gene products are functionally significant. Translation of the genes indicates a similar degree of amino acid sequence divergence between class I and class II CpsC proteins (approximately 70% identity). Interestingly, even small differences between the functionally homologous Rol (Wzz) proteins of *Shigella* species has previously been shown to affect the modal chain length of the lipopolysaccharide O antigen (11).

The *cpsE* gene was also present in all *S. pneumoniae* serotypes tested which contain glucose in their CPS, except type 3, which has a different mode of CPS biosynthesis (1, 5). This gene is also separated into either class I or class II, along with the preceding *cpsC* and *cpsD* genes. However, the two different classes do not appear to affect the function of CpsE, which is a glucose-1-phosphate transferase. Kolkman et al. (14) demonstrated glucose-1-phosphate transferase activity in several pneumococcal serotypes now known to contain either class I or class II *cpsE* genes. In all *S. pneumoniae cps* loci sequenced to date, the gene which follows *cpsE* is serotype or serogroup specific (21).

All *S. pneumoniae cps* loci examined contain highly conserved *cpsA* and *cpsB* genes, indicating that they probably evolved from a common ancestor. However, their *cpsC*, *cpsD*, and *cpsE* genes can be separated into either class I or class II sequences, suggesting that recombination between the original *S. pneumoniae* ancestor (either class I or class II) and exogenous DNA resulted in the formation of two distinct clonal *S. pneumoniae* strains from which all subsequent serotypes have evolved. The presence of DNA homologous to *cps19fA* to *-D*, even though these genes are not functional in type 3 pneumococci (7), probably reflects the common origin between type 3 and other class I pneumococci.

The type 4 and 19B *cpsC* sequences both show evidence of recombination within the *cps* loci. Two recent studies have demonstrated that natural recombination events involving exchange of entire *cps* loci (or major portions thereof) have resulted in switching of capsule type (e.g., from 23F to 19F) by multiply drug-resistant pneumococcal clones on numerous occasions (4, 19). The current study indicates that recombination events involving small fragments within pneumococcal *cps* loci may also be common in nature and may represent a mechanism whereby additional serotype diversity is generated.

Nucleotide sequence accession numbers. The *cps19a* sequence has been deposited with GenBank under accession no. AF094575. The sequences for the 5′ region of *cpsC* from serotypes 2, 6A, 6B, 8, 9N, 19B, and 20 are available under GenBank accession no. AF106132, AF106133, AF106134, AF106135, AF106136, AF106137, and AF106138, respectively.

This work was supported by a grant from the National Health and Medical Research Council of Australia.

REFERENCES

1. **Arrecubieta, C., E. García, and R. López.** 1995. Sequence and transcriptional analysis of a DNA region involved in the production of capsular polysaccharide in *Streptococcus pneumoniae* type 3. Gene **167:**1–7.
2. **Austrian, R.** 1981. Some observations on the pneumococcus and on the current status of pneumococcal disease and its prevention. Rev. Infect. Dis. **3**(Suppl.)**:**S1–S17.
3. **Austrian, R., H. P. Bernheimer, E. E. B. Smith, and G. T. Mills.** 1959. Simultaneous production of two capsular polysaccharides by pneumococcus. II. The genetic and biochemical bases of binary capsulation. J. Exp. Med. **110:**585–602.
4. **Coffey, T. J., M. C. Enright, M. Daniels, J. K. Morona, R. Morona, V. Hryniewicz, J. C. Paton, and B. G. Spratt.** 1998. Recombinational exchanges at the capsular polysaccharide biosynthesis locus lead to frequent serotype changes among natural isolates of *Streptococcus pneumoniae*. Mol. Microbiol. **27:**73–83.
5. **Dillard, J. P., M. W. Vandersea, and J. Yother.** 1995. Characterization of the cassette containing genes for type 3 capsular polysaccharide biosynthesis in *Streptococcus pneumoniae*. J. Exp. Med. **181:**973–983.
6. **Douglas, R. M., J. C. Paton, S. J. Duncan, and D. Hansman.** 1983. Antibody response to pneumococcal vaccination in children younger than five years of age. J. Infect. Dis. **148:**131–137.
7. **García, E., and R. López.** 1997. Molecular biology of the capsular genes of *Streptococcus pneumoniae*. FEMS Microbiol. Lett. **149:**1–10.
8. **Guidolin, A., J. K. Morona, R. Morona, D. Hansman, and J. C. Paton.** 1994. Nucleotide sequence of an operon essential for capsular polysaccharide biosynthesis in *Streptococcus pneumoniae* type 19F. Infect. Immun. **62:**5384–5396.
9. **Henrichsen, J.** 1995. Six newly recognized types of *Streptococcus pneumoniae*. J. Clin. Microbiol. **33:**2759–2762.
10. **Katzenellenbogen, E., and H. J. Jennings.** 1983. Structural determination of the capsular polysaccharide of *Streptococcus pneumoniae* type 19A (57). Carbohydr. Res. **124:**235–245.
11. **Klee, S. R., B. D. Tzschaschel, K. N. Timmis, and C. A. Guzman.** 1997. Influence of different *rol* gene products on the chain length of *Shigella dysenteriae* type 1 lipopolysaccharide O antigen expressed by *Shigella flexneri* carrier strains. J. Bacteriol. **179:**2421–2425.
12. **Kolkman, M. A. B., D. A. Morrison, B. A. M. van der Zeijst, and P. J. M. Nuijten.** 1996. The capsule polysaccharide synthesis locus of *Streptococcus pneumoniae* serotype 14: identification of the glycosyl transferase gene *cps14E*. J. Bacteriol. **178:**3736–3741.
13. **Kolkman, M. A. B., B. A. M. van der Zeijst, and P. J. M. Nuijten.** 1997. Functional analysis of glycosyltransferases encoded by the capsular polysaccharide biosynthesis locus of *Streptococcus pneumoniae* serotype 14. J. Biol. Chem. **272:**19502–19508.
14. **Kolkman, M. A. B., B. A. M. van der Zeijst, and P. J. M. Nuijten.** 1998. Diversity of capsular polysaccharide synthesis gene clusters in *Streptococcus pneumoniae*. J. Biochem. (Tokyo) **123:**937–945.
15. **Kumar, S., K. Tamura, and M. Nei.** 1994. MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. Comput. Appl. Biosci. **10:** 189–191.
16. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

16a. **Morona, J. K.** Unpublished observations.

17. **Morona, J. K., R. Morona, and J. C. Paton.** 1997. Characterization of the locus encoding the *Streptococcus pneumoniae* type 19F capsular polysaccharide biosynthetic pathway. Mol. Microbiol. **23:**751–763.
18. **Muñoz, R., M. Mollerach, R. López, and E. García.** 1997. Molecular organization of the genes required for the synthesis of type 1 capsular polysaccharide of *Streptococcus pneumoniae*: formation of binary encapsulated pneumococci and identification of cryptic dTDP-rhamnose biosynthesis genes. Mol. Microbiol. **25:**79–92.
19. **Nesin, M., M. Ramirez, and A. Tomasz.** 1998. Capsular transformation of a multidrug-resistant *Streptococcus pneumoniae* in vivo. J. Infect. Dis. **177:**707–713.
20. **Ohno, N., T. Yadomae, and T. Miyazaki.** 1980. The structure of the type specific polysaccharide of pneumococcus type XIX. Carbohydr. Res. **80:**297–304.
21. **Paton, J. C., and J. K. Morona.** 1999. *Streptococcus pneumoniae* capsular polysaccharide. *In* V. Fischetti, R. Novick, J. Ferretti, D. Portnoy, and J. Rood (ed.), Gram-positive pathogens, in press. ASM Press, Washington D.C.
22. **Ramirez, M., and A. Tomasz.** 1998. Molecular characterization of the complete 23F capsular polysaccharide locus of *Streptococcus pneumoniae*. J. Bacteriol. **180:**5273–5278.
23. **Tamara, K., and M. Nei.** 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10:**512–526.
24. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
25. **van Dam, J. E. G., A. Fleer, and H. Snippe.** 1990. Immunogenicity and immunochemistry of *Streptococcus pneumoniae* capsular polysaccharides. Antonie Leeuwenhoek **58:**1–47.