CHARACTERISATION OF THE CAPSULAR POLYSACCHARIDE BIOSYNTHESIS LOCI OF *STREPTOCOCCUS PNEUMONIAE* SEROGROUP 19

by

Judy Kay Morona B. Sc. (Adelaide)

A thesis submitted for the degree of Doctor of Philosophy

Department of Microbiology and Immunology
University of Adelaide

November, 1998

ABSTRACT

In this thesis, the genetic loci encoding capsular polysaccharide synthesis (*cps*) have been characterised for all members of *Streptococcus pneumoniae* serogroup (19F, 19A, 19B and 19C). In each serotype, the *cps* locus is located in the *S. pneumoniae* chromosome between *dexB* and *aliA* and appears to be arranged as a single transcriptional unit. The arrangement of the genes within the *cps19* loci is highly conserved with 13 genes (*cps19A-H*, and *K-O*) common to all four serogroup 19 members. These genes encode functions required for the synthesis of the shared trisaccharide component of the group 19 capsular polysaccharide (CPS) repeat unit structures. Furthermore, the genetic differences between the group 19 *cps* loci identified are consistent with the differences in the CPS structures of individual serotypes. Functions have been assigned to nearly all the *cps19* gene products, based on either gene complementation or similarity to other proteins with known functions. This has enabled biosynthetic pathways for production of all four group 19 CPSs to be proposed.

Nearly all of the common genes from types 19F, 19B and 19C are >95% identical to each other. However, closely related homologues of *cps19fI* and *J*, which encode the type 19F polysaccharide polymerase and repeat unit transporter, respectively, are not found in the type 19B and 19C *cps* loci. In type 19B and 19C this region of the *cps* locus (between *cps19bH* and *cps19bK*) contains five genes which encode a unique polysaccharide polymerase and repeat unit transporter, as well as two additional putative glycosyl transferases and a protein which may be involved in synthesis of an activated ribose precursor. Transformation studies indicated that these five genes encode all of the functions required to convert a type 19F pneumococcus to type 19B. The type 19C *cps*

locus differs from the 19B *cps* locus only in the insertion of a glucosyl transferase gene (*cps19cS*) between *cps19cK* and *cps19cL*. Transformation studies have shown that the presence of this gene accounts for the additional glucose side chain in the otherwise identical repeat unit structures. The type 19C *cps* locus contains 19 genes, and at 21 kb it is the largest pneumococcal capsule gene cluster characterised to date.

Although the *cps19a* and *cps19f* loci are identical in the number and arrangement of the genes present, the similarity between individual genes varies from 70% to 99% identity (for both the nucleotide and the deduced amino acid sequences). This sequence divergence is surprising given that the only difference between their CPS repeat units is the glycosidic linkage which joins the repeat units together ($\alpha(1\rightarrow2)$ for 19F and $\alpha(1\rightarrow3)$ for 19A). Theoretically, only a difference in the *cps19aI* gene, which presumably encodes the polysaccharide polymerase responsible for this linkage, is required to change a type 19F pneumococcus into type 19A. Indeed, this was demonstrated by a transformation event in which the region of the *cps19a* locus encoding Cps19aH and Cps19aI replaced the homologous portion of the *cps19f* locus was sufficient to convert CPS type from 19F to 19A. Given that Cps19fH and Cps19aH are >95% identical, it seems probable that Cps19aI (79% identity) is solely responsible for the observed alteration in CPS type.

The serotype specificity of the *cps19f* genes was investigated by Southern hybridisation analysis of chromosomal DNA from other *S. pneumoniae* serotypes. Large variations in the hybridisation patterns were obtained with the different gene-specific probes. Probes specific for sequences flanking *cps19f* hybridised with all the serotypes tested. However, within the *cps* loci, only *cps19fA* and *cps19fB* were common to all serotypes. Based on the Southern hybridisation analysis a protocol for PCR amplification of *cps* loci was developed and used to amplify the *cps* regions from a variety of pneumococcal serotypes. Direct sequencing of the 5' end of the PCR products was undertaken and identified two classes of *cpsC* gene. Southern hybridisation studies with

cps19aC- and *D*-specific gene probes demonstrated that homologues of the first four genes in the *cps* locus, *cpsA-D,* are present in all serotypes and that all the *cps* loci tested evolved from one of two clonal origins which contained either class I or class II *cpsC* and *D* genes. The *cpsE* gene, which encodes a glucose-1-phosphate transferase, is also conserved (in the two distinct classes) in the *cps* loci of all serotypes tested which contain glucose in their CPS, except type 3.

The sequence analysis of the various *cps* loci presented in this thesis provides further evidence that in nature frequent recombination occurs between different *cps* loci resulting in either complete exchange of the *cps* locus or exchange of only part of the *cps* locus, and could potentially result in the expression of new capsular serotypes.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any University or other tertiary institution and to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference is made in the text.

I give consent to this copy of my thesis, when deposited in the University library being available for loan and photocopying.

Judy Kay Morona

ACKNOWLEDGMENTS

I would like to thank my supervisors Assoc. Prof. James Paton and Prof. Paul Manning for the opportunity to undertake my postgraduate studies. In particular, thank you James for your encouragement and your confidence in me, and for your help with this thesis. This project has been both challenging and rewarding.

I would also like to thank Gianni Pozzi, Francesco Iannelli, Rubens López, Ernesto García and Alexander Tomasz for kindly providing copies of their manuscripts prior to publication. This has enabled a much more complete comparison of the various pneumococcal *cps* loci. Also thank you to Mike Gratten, Denise Murphy and Andrew Lawrence for serotyping various pneumococcal isolates and transformants.

To my colleagues, past and present, especially Anne Berry, Rob Lock, David Ogunniyi, David Miller, Ursula Talbot, Angelo Guidolin, Adrienne Paton, Rebekah Miller, Chris Vindurampulle, Matthew Woodrow and Stuart McKessar, thank you for the camaraderie. Those elusive results always seemed more attainable after a celebratory glass of champagne. To David M., thankyou for your help and enthusiasm over the last few months. To Anne, I especially thank you for all of your help and friendship over the years.

To Renato, my husband, a very special thank you, all of your enthusiastic discussions have been invaluable. Thank you for your support and your help with the housework. I couldn't have done it without you.

To my daughter Stephanie, thank you for your patience and love, I know I have locked myself into the study far too often recently.

ABBREVIATIONS

Abbreviations which are acceptable to the Journal of Bacteriology are used in this thesis without definition in the text. Additional abbreviations are defined when first used in the text and are listed below.

Amp	ampicillin
AP	alkaline phosphatase
BA	blood agar
Cml	chloramphenicol
CPS	capsular polysaccharide
cps	capsule biosynthesis genes
C-terminus	carboxy terminus
DIG	digoxigenin
DOC	deoxycholate
ECA	enterobacterial common antigen
EPS	exopolysaccharide
Ery	erythromycin
ExoIII	exonuclease III
Gal	galactose
Glc	glucose
GlcA	glucuronic acid
Glc*N*Ac	*N*-acetyl glucosamine
Hib vaccine	*Haemophilus influenzae* type b vaccine
HPLC	high-performance liquid chromatography
InPCR	inverse PCR
Kan	kanamycin
LB	Luria-Bertani broth
LB+KA	Luria-Bertani broth containing ampicillin and kanamycin
LOS	lipooligosaccharide
LPS	lipopolysaccharide
LR-PCR	long-range PCR
Man*N*Ac	*N*-acetyl mannosamine
N-terminus	amino terminus
ORF	open reading frame
PBP	penicillin binding proteins
Rha	rhamnose
Rib	ribose
Rif	rifampicin
SC	free secretory component
Str	streptomycin
TEMED	N,N,N',N'-tetramethyl-ethylene-diamine
Tet	tetracycline
THY	Todd-Hewitt broth supplemented with 0.5% (wt/v) yeast extract
und-P	undecaprenol-phosphate

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1.1 *Streptococcus pneumoniae*, a Distinguished History

Streptococcus pneumoniae, or the pneumococcus, was first isolated in 1880 by Sternberg in the United States and Pasteur in France. Both independently isolated "roughly lancet-shaped pairs of coccoid bacteria" from the blood of rabbits injected with human saliva. Early intensive scientific investigations of the pneumococcus yielded many important discoveries concerning both cellular and molecular biology (Austrian, 1981a; Watson *et al.,* 1993).

Some of these fundamental achievements, which can be attributed to studies on the pneumococcus, are summarised below.

1.1.1 Development of the Gram stain

Gram (1884) examined sections of lung tissue from people who had died of pneumonia and discovered that these sections contained many coccoid bacteria that retained an aniline-gentian violet stain, "the cocci of croupous pneumonia". Although the significance was not realised at the time, he also noted the presence of an encapsulated coccoid bacterium that did not retain the stain (later identified as *Klebsiella pneumoniae*).

Today, most clinically important bacteria are recognised as either Gram-negative or Gram-positive.

1.1.2 Humoral immunity and phagocytosis

Klemperer and Klemperer (1891) showed that rabbits injected with heat-killed pneumococci were immune to reinfection with the same strain but not to reinfection with different clinical isolates. They also demonstrated that rabbits were protected against primary infection by infusion of serum from an immunised rabbit. Issaeff (1893) demonstrated that the protective serum was not bactericidal, rather it promoted the ingestion of pneumococci by the phagocytic cells of the immune system. Neufeld and Rimpau (1904) demonstrated that this phenomenon results from the exposure of the bacteria, rather than the phagocytes, to the serum. These observations described opsonisation, the interaction of antibodies and complement with the bacterial surface.

1.1.3 The concept of serotyping

Neufeld (1902) demonstrated both macroscopic agglutination and microscopic swelling of the pneumococcal capsule after addition of specific antiserum to a cell suspension. Although the capsule swelling or "quellung" reaction became the accepted technique for routinely serotyping pneumococcal isolates, it was not used for this purpose until much later (Armstrong, 1931).

1.1.4 The observation of antibiotic resistance in bacteria

The pneumococcus was one of the first bacteria to be recognised as being able to acquire resistance to antibiotics. Morgenroth and Kaufman (1912) isolated pneumococci resistant to optichin, a derivative of quinine, from mice which had been treated with this

antibiotic. Moore and Chesney (1917) also described optichin resistant pneumococci isolated from humans. However, these findings pre-dated the development and introduction of safe, effective antibiotic treatments and were largely forgotten. There were reports of antibiotic resistant pneumococcal isolates in the 1940s (Frisch *et al.*, 1943; Eriksen, 1945), however, the significance of clinical resistance to antibiotics was not recognised until twenty years later when a penicillin resistant pneumococcus was isolated in Australia from the sputum of a patient (Hansman and Bullen, 1967). Antibiotic resistant pneumococci are becoming more prevalent and the significance of this will be discussed in section 1.6.

1.1.5 Recognition of the first non-protein antigen

The discovery of a "soluble substance of the Pneumococcus" in the blood and urine from pneumonia patients (Dochez and Avery, 1917) lead to the eventual recognition of the first non-protein antigen, the pneumococcal capsular polysaccharide (CPS). Heidelberger and Avery (1923) identified this soluble substance as a complex carbohydrate or polysaccharide which was responsible for serological reactivity. However, it was widely believed that only proteins could be immunogenic leading to the assumption that the immunogenicity was due to a contaminating protein, even though it was resistant to the effects of boiling or trypsin digestion. Heidelberger (1927) later concluded that this highly reactive substance was distributed over the surface of the bacteria and was responsible for type specificity. Heidelberger and colleagues then showed that this capsule was immunogenic by immunising mice with the capsular material to protect them from subsequent pneumococcal challenge.

1.1.6 Bacterial transformation

The phenomenon of transformation of pneumococci from one serotype to another

was first described by Griffith (1928). His studies were designed to determine the requirement for reversion of unencapsulated or rough pneumococcal variants to the encapsulated or smooth form. When mice were injected with live rough pneumococci derived from one serotype and heat-killed cells from another, a proportion of the mice succumbed to an infection caused by pneumococci of the same serotype as that of the heat-killed inoculate. It wasn't long before capsular transformation was demonstrated *in vitro* by Dawson and Sia (1931), and repeated by Alloway (1932; 1933) using pneumococcal cell extracts.

1.1.7 Discovery of DNA as the genetic material

The hunt for this "transforming principle" led to one of the most significant discoveries in biological science this century. In 1944, Avery, McLeod and McCarty published a paper which showed conclusively that DNA was the determinant of capsular transformation, and therefore, was the carrier of genetic information (Avery *et al.*, 1944). A paper published two years later (McCarty and Avery, 1946) silenced the sceptics by showing that deoxyribonuclease destroyed the biological activity of the "transforming principle". In this paper, McCarty wrote with great insight: "It remains one of the challenging problems for future research to determine what sort of configurational or structural differences can be demonstrated between desoxyribonucleates of separate specificities." McCarty suggested that "the nucleic acid of the pneumococcus is concerned with innumerable other functions of the bacterial cell" thus "discovering the chemical basis of biological specificity of desoxyribonucleic acids becomes extremely complex, since a given preparation will represent a mixture of a large number of entities of diverse specificity." Hotchkiss (1951) demonstrated this "diverse specificity" of DNA by introducing a trait other than capsule type into a pneumococcus. He induced penicillin

resistance in a penicillin-sensitive pneumococcus by transfer of DNA from a penicillin-resistant strain.

1.2 Pneumococcal Disease

More than a century after its initial isolation *S. pneumoniae* is still an important cause of life-threatening, invasive diseases such as pneumonia, bacteraemia and meningitis, with high morbidity and mortality throughout the world. *S. pneumoniae* is also a leading cause of less serious but highly prevalent infections such as otitis media and sinusitis.

An overview of the mechanisms involved in pneumococcal disease and the pathogenesis of *S. pneumoniae* is presented below. Several comprehensive reviews on this topic have been published recently (Musher, 1992; Paton *et al.,* 1993; Tuomanen *et al.,* 1995; AlonsoDeVelasco *et al.,* 1995; Watson and Musher, 1996).

1.2.1 Colonisation of the nasopharynx

S. pneumoniae is carried, asymptomatically, in the upper respiratory tract by many healthy individuals. The transmission of pneumococci from a carrier to another person is dependent on the frequency and intimacy of their contact (Riley and Douglas, 1981). Thus among adults, carrier rates are highest for those living in crowded conditions such as in barracks and dormitories. This person to person spread often occurs concurrently with viral infection of the upper respiratory tract. Carriage rates are highest among preschool children and tend to decrease with increasing age (AlonsoDeVelasco *et al.,* 1995). Day-care centre attendance is associated with increased carriage rates of pneumococci and thus the risk of invasive pneumococcal disease for such children is increased (Takala *et al.,*

1995; Cherian *et al.*, 1994). Virtually everyone is colonised by pneumococci at some stage. The average duration of carriage is about six weeks in an adult, but can be greater than one year in some people (Musher, 1992). There are two consequences of carriage, either seroconversion and subsequent elimination of the pneumococcus, or invasion of the organism leading to pneumococcal disease. Most infections do not result from prolonged carriage, but probably occur within the first week of colonisation (Musher, 1992).

1.2.2 Risk factors for pneumococcal disease

Pneumococcal disease occurs most frequently in the extreme ages of life. High rates of invasive pneumococcal disease occur in children under two. The rate then decreases with increasing age with teenage children and young adults having the lowest rates of disease. The rate then increases with middle age, again reaching a high level in the elderly (Musher, 1992).

People with functional or anatomic asplenia are highly susceptible to pneumococcal bacteraemia because of reduced capacity to clear encapsulated bacteria from the blood. Increased risk of pneumococcal disease is also associated with certain medical conditions that result in either impaired pulmonary clearance mechanisms (such as cigarette smoking, emphysema, chronic bronchitis, chronic pulmonary disease and viral respiratory infections such as influenza), or reduced immune responses (such as leukaemia, multiple myeloma, lymphoma, HIV infection, Hodgkin's disease, organ or bone marrow transplantation, prolonged use of systemic corticosteroids and chronic renal failure). Other medical conditions that increase susceptibility to pneumococcal disease include chronic cardiovascular diseases, chronic liver diseases, diabetes mellitus (which is often associated with either cardiovascular or renal dysfunction), malnutrition, and alcoholism. People with chronic cerebrospinal fluid leakage resulting from congenital lesions, skull fractures or

neurosurgical procedures are also at risk of recurrent pneumococcal meningitis (Centers for Disease Control and Prevention, 1997).

1.2.3 Incidence of pneumococcal disease

S. pneumoniae is the most common cause of acute otitis media in children aged less than five years in the United States (Bluestone *et al.,* 1992). In the United States more than half of all children suffer acute otitis media during their first year of life, and nearly half have had at least three attacks before their third birthday (Teele *et al.,* 1989). Although acute otitis media does not usually progress to invasive disease, it has a considerable impact upon health care costs.

S. pneumoniae is the commonest cause of community-acquired bacterial pneumonia. It is thought that at least 500,000 cases of pneumococcal pneumonia occur annually in the United States. The precise figure is difficult to ascertain because in many cases of pneumonia, the aetiological agent remains unidentified. In the United States, *S. pneumoniae* has been estimated to be responsible for 25-35% of all cases of bacterial pneumonia requiring hospitalisation (Fang, 1990), and concurrent bacteraemia occurs in 10-25% of adult patients. *S. pneumoniae* is also a leading cause of bacterial meningitis in the United States (Wenger *et al.,* 1990).

Invasive pneumococcal disease causes 40,000 deaths a year in the United States alone (Centers for Disease Control and Prevention, 1997). Despite appropriate antimicrobial therapy and intensive medical care, the overall case fatality rates are about 15-20% for bacteraemia and meningitis. This rate increases to 30-40% among elderly patients (Hook *et al.,* 1983; Wenger *et al.,* 1990). As the prevalence of multiply antibiotic resistant pneumococci increases, pneumococcal disease will become more difficult to manage, potentially increasing the mortality rate further.

1.2.4 Pathogenesis of pneumococcal infections

The first stage of pneumococcal infection is nasopharyngeal colonisation. The virulence of the colonising strain and the immune status of the host are important determining factors in the onset of pneumococcal disease. However, the mechanisms which enable the pneumococcus to infect the lung or to migrate directly into the bloodstream are poorly understood (AlonsoDeVelasco *et al.,* 1995). Certain risk factors, such as the presence of a respiratory viral infection, greatly enhance the spread of pneumococci to the lungs. Epithelial damage caused by the virus also aids the spread of the pneumococcus into the bloodstream. From the blood, the pneumococcus may spread to the meninges. The pneumococcus may also enter the meninges directly from the nasopharynx either if the dura mater has been compromised or as a complication of sinusitis (Musher, 1992).

1.2.5. Pneumococcal virulence factors

Many pneumococcal components are known to contribute to the pathogenicity of this organism. Various virulence factors and their role in pneumococcal disease are described below.

1.2.5.1 Capsule

The polysaccharide capsule protects the pneumococcus from phagocytosis and is recognised as the major virulence factor of *S. pneumoniae.* All clinical isolates of *S. pneumoniae* are smooth. Although rough strains can be maintained *in vitro*, such strains are almost completely avirulent. Avery and Dubos (1931) demonstrated that enzymatic depolymerisation of the capsule increased the LD_{50} of a type 3 pneumococcus by 10^6 fold. More recently a similar effect on the virulence of a type 3 pneumococcus was achieved by

transposon mutagenesis of a gene essential for capsule production (Watson and Musher, 1990).

1.2.5.2 Cell Wall

Both cell wall peptidoglycan and teichoic acid induce inflammatory responses in the host. Teichoic acid activates the alternate pathway of complement which enhances vascular permeability, and recruits and activates leucocytes (AlonsoDeVelasco *et al.*, 1995). The cell wall components also induce production of cytokines such as interleukin-1, and stimulate production of platelet-activating factor (Tuomanen *et al.*, 1995). The acute inflammatory response generated is thought to cause the tissue damage responsible for the high morbidity and mortality associated with pneumococcal disease (Musher, 1992). In fact, the tissue damage caused by pneumococcal disease can be mimicked by injecting purified pneumococcal cell wall components into animals (Tuomanen *et al.*, 1985).

1.2.5.3 Autolysin

Autolysin, a N-acetylmuramic acid-L-alanine amidase, cleaves the covalent bond between the glycan chains and the peptide side chains in the cell wall. Autolysin mutants tend to grow in short chains rather than discreet diplococci suggesting that autolysin may be involved in cell separation (Paton *et al.*, 1993). The lytic tendency of pneumococci in stationary phase culture, or in the presence of deoxycholate (DOC), is caused by the activation of autolysin. Autolysin is thought to contribute to pneumococcal pathogenesis because of its ability to lyse the cell. Unrestrained growth of the pneumococcus at the site of infection leads to autolysis and the release of cell wall fragments and virulence proteins such as pneumolysin resulting in an inflammatory response and progression of pneumococcal disease (AlonsoDeVelasco *et al.*, 1995).

1.2.5.4 Pneumolysin

Pneumolysin is a thiol-activated cytolysin which is produced by nearly all clinical

S. pneumoniae isolates. Pneumolysin is a member of the thiol-activated cytolysin family of toxins produced by several Gram-positive genera. However, pneumolysin differs from the other members of this family in that it is not secreted, but remains cytoplasmic until it is released by autolysis. Pneumolysin is a bifunctional toxin, and in addition to its cytotoxic properties it is capable of directly activating the classical complement pathway by binding to the Fc region of human IgG (Paton *et al.,* 1984; Mitchell *et al.*, 1991). It has been shown to have toxic effects on many cell types, which undoubtedly contributes to the pathogenesis of pneumococcal disease (for reviews see Paton *et al.*, 1993; Paton, 1996). These properties include inhibition of the bactericidal activity of leucocytes (Paton and Ferrante, 1983), blockade of proliferative responses and Ig production by lymphocytes (Ferrante *et al.,* 1984), reduction of ciliary beating of human respiratory epithelium (Steinfort *et al.,* 1989), and direct cytotoxicity for respiratory endothelial and epithelial cells (Rubins *et al.,* 1992; 1993). Inactivation of the pneumolysin gene in a *S. pneumoniae* type 2 strain increased the LD_{50} (in mice) approximately 100-fold (Berry *et al.*, 1989) and immunisation with purified pneumolysin increases the survival time of mice challenged intranasally with virulent pneumococci (Paton *et al.,* 1983; 1993).

1.2.5.5 Neuraminidase

All clinical *S. pneumoniae* isolates produce one or more neuraminidases. To date, two different enzymes have been identified in the pneumococcus (Lock *et al.,* 1988a; Camara *et al.,* 1991; Berry *et al.,* 1996). These enzymes cleave the terminal sialic acid residues from glycolipids, glycoproteins, and oligosaccharides on cell surfaces. This has the potential to cause great damage to the host and, indeed, the purified protein has been shown to be toxic for mice (Lock *et al.,* 1988b). Another possible role in pathogenesis for neuraminidase could be the unmasking of cell-surface receptors for pneumococcal adhesins (Paton *et al.,* 1993).

1.2.5.6 Hyaluronidase

Most clinical isolates produce this enzyme which is associated with the cell surface. It degrades hyaluronic acid, which is found associated with mammalian connective tissue. Thus, hyaluronidase may contribute to pathogenesis by degradation of the connective tissue, allowing greater access to host tissues, thereby facilitating invasion (Paton *et al.*, 1993; Berry *et al.*, 1994).

1.2.5.7 Pneumococcal surface protein A (PspA)

PspA and autolysin, are members of a family of pneumococcal choline binding proteins which contain C-terminal repeated choline-binding domains (García *et al.*, 1986; Yother and Briles, 1992). Although its precise function is unknown, PspA is a highly variable protective antigen, both in its molecular weight and its antigenicity (Crain *et al.*, 1990; Waltman *et al.*, 1990). PspA-negative mutant pneumococci are more readily cleared from the blood of mice than wild type strains (McDaniel *et al.*, 1987). The relative importance of PspA to pathogenicity of pneumococci appears to vary from strain to strain. More dramatic reductions in virulence were observed when PspA genes in strains of serotypes 3 and 5 were inactivated compared to a type 2 strain (Briles *et al.*, 1988).

1.2.5.8 PsaA

PsaA is a lipoprotein which was initially thought to be an adhesin based on sequence homology with putative lipoprotein adhesins of *S. sanguis* and *S. parasanguis* (Sampson *et al.*, 1994). Pneumococcal *psaA*-negative mutants, unable to express PsaA, are virtually avirulent for mice (Berry and Paton, 1996), and immunisation with the purified protein protects mice from challenge with virulent pneumococci (Talkington *et al.*, 1996). The *psaA* gene is part of an operon which was recently demonstrated to encode a manganese transporter (Dintilhac *et al.*, 1997). Thus, its role in pathogenesis could be explained either by a requirement for manganese for regulation of expression of other

virulence factors, or by growth retardation due to an inability to scavenge this metal *in vivo* (Paton, 1998).

1.2.5.9 Choline binding protein A (CbpA)

CbpA is also a pneumococcal choline binding protein (Rosenow *et al.,* 1997). *S. pneumoniae cbpA*-negative mutants exhibited reduced adherence to cytokine-activated cells and endothelial cells *in vitro*, and failed to bind to immobilised sialic acid or lacto-N-neotetraose which are known pneumococcal ligands on these cells, suggesting that CbpA may be an adhesin (Rosenow *et al.,* 1997). Hammerschmidt *et al* (1997) independently described a protein, designated SpsA, which binds specifically to human secretory IgA and the free secretory component (SC). Interaction between the pneumococcal cell surface and the SC may interfere with the activity of secretory IgA, and may also directly facilitate adherence, thereby promoting colonisation of the nasopharyngeal mucosa. The demonstrated degree of heterogeneity, in the amino (N-) terminal region, of SpsA among different pneumococcal strains (Hammerschmidt *et al.,* 1997) and the high degree of sequence similarity between CbpA and SpsA indicate that they are the same protein (Paton, 1998). The adhesive properties of CbpA are consistent with the observation that carriage rates of pneumococcal CbpA mutants were reduced 100-fold in an animal model, whereas, there was no reduction of virulence in an intraperitoneal model of sepsis (Rosenow *et al.,* 1997).

1.3 The Cell Surface of the Pneumococcus

A schematic representation of the structure of the pneumococcal cell surface is shown in **Fig. 1.1**. It consists of three layers, the innermost layer is the plasma membrane

which is followed by the peptidoglycan cell wall layer. The outermost layer consists of the polysaccharide capsule which is covalently linked to the peptidoglycan in the cell wall. Properties of each component of the cell surface are described below.

1.3.1 The cell wall

The pneumococcal cell wall is typical of a Gram-positive bacterium. It consists of several layers of peptidoglycan, comprising glycan chains of alternating *N*-acetyl glucosamine (Glc*N*Ac) and *N*-acetylmuramic acid cross-linked with species-specific peptide bridges. These stem peptides have alanine as the first residue, linked to *N*-acetylmuramic acid. High-performance liquid chromatography (HPLC) has revealed the complexity of the peptidoglycan supramolecular structure for both Gram-positive and Gram-negative bacteria. HPLC analysis has identified 18 different muropeptides of highly conserved molar ratios in *S. pneumoniae* (Garcia-Bustos *et al.*, 1987). This profile is

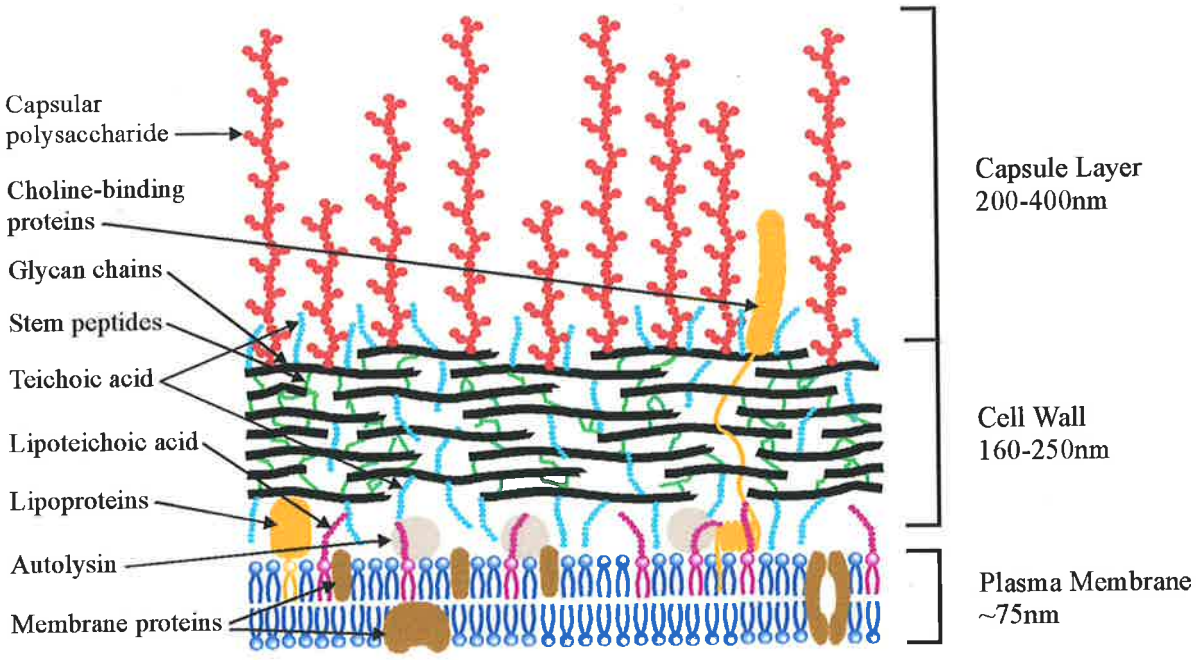


Fig. 1.1. The pneumococcal cell surface. Diagrammatic representation of the pneumococcal cell surface showing possible arrangements of constituents. The thickness of each layer is as described by Tomasz (1981). Representation of cell wall, teichoic acid and capsular polysaccharide adapted from Hammond *et al.* (1984), Fig 2.11.

found in all pneumococci, regardless of serotype, geographic origin or isolation date, with the exception of penicillin- resistant strains which have different muropeptide profiles (Garcia-Bustos and Tomasz, 1990). Resistant pneumococci have alterations in the penicillin binding proteins (PBPs) which result in reduced affinity for the drug. PBPs are enzymes (transpeptidases or carboxypeptidases) that are involved in the assembly of bacterial cell walls. The interaction between PBPs and penicillin is due to the structural similarity between the β–lactam ring of penicillin and the carboxy (C) -terminal D-alanyl-D-alanine residues in the cell wall stem peptides. Thus, it is highly likely the altered PBPs have altered affinities for their natural substrates resulting in an altered peptidoglycan structure in penicillin resistant pneumococci (Severin and Tomasz, 1996).

1.3.2 Teichoic acid

The teichoic acid of the pneumococcus, the C-polysaccharide, was first described by Tillett and Francis (1930), although its chemical structure (**Fig. 1.2**) was not determined until fifty years later (Jennings *et al.*, 1980).

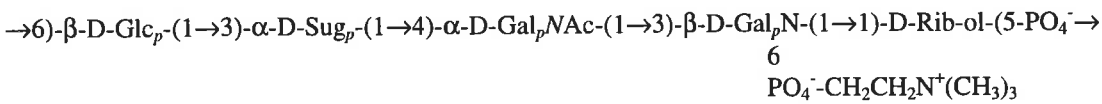


Fig. 1.2. Structure of the C-polysaccharide. Glc_p, glucose; Sug_p, 2-acetamido-4-amino-2,4,6, trideoxy-D-galactose; Gal_pNAc, *N*-acetyl galactosamine; Gal_pN, galactosamine; Rib-ol, ribitol; PO_4^--$CH_2CH_2N^+(CH_3)_3$, phosphorylcholine.

The teichoic acid is covalently linked to the pneumococcal peptidoglycan via *N*-acetylmuramic acid (Tomasz, 1981). The phosphorylcholine in the teichoic acid is a recognition site for autolysin, an *N*-acetylmuramic acid-L-alanine amidase, which cleaves

the bond between alanine and *N*-acetylmuramic acid.

Teichoic acid is uniformly distributed on both the inner and the outer surfaces of the cell wall, and possibly within it as well. The thickness of the teichoic acid layer varies between different pneumococcal strains (Sørensen *et al.*, 1988). The level of teichoic acid substitution depends on the degree of cross-linking within the cell wall peptidoglycan (Fischer and Tomasz, 1985).

Interestingly, the individual components of the teichoic acid can all be found as a constituent in at least one pneumococcal CPS type. Thus, it is a possibility that some of the genes involved in teichoic acid and CPS synthesis may be shared. It is also possible that the capsule (*cps*) genes evolved by duplication of the teichoic acid locus in an ancestral *S. pneumoniae* strain.

1.3.3 The plasma membrane

The plasma membrane consists of a lipid bilayer closely associated with the inner layer of the cell wall. Complex foldings of the plasma membrane, or mesosomal structures, have been observed in the pneumococcus (Tomasz, 1981). However the function of these mesosomal structures is unknown.

The Forssman (F) antigen is a lipid-linked teichoic acid, identical in structure to C-polysaccharide, and is found in the outer leaflet of the plasma membrane. Approximately 20% of the teichoic acid content of the pneumococcus is lipid-linked. The pneumococcal F antigen is a highly specific inhibitor of autolysin, suggesting that it may play an important role in regulating cellular autolysis. When the association between autolysin and the F antigen is disrupted, either by cessation of cell wall biosynthesis or treatment with detergents such as DOC, autolysin becomes free to interact with the choline in teichoic acid. This interaction enables autolysin to cleave the covalent bond between the glycan

chains and the stem peptides, resulting in destruction of the cell wall and cell lysis (Giudicelli and Tomasz, 1984).

1.3.4 Proteins associated with the cell surface

S. pneumoniae is predicted to contain a large number of surface exposed proteins but to date only a small number have been investigated. PspA, PspB and CbpA have been shown to be exposed on the cell surface of pneumococci. PspA and CbpA are choline-binding proteins which are structurally similar, with almost identical C-termini (Yother and Briles, 1992; Rosenow *et al.,* 1997; Hammerschmidt *et al.,* 1997). They are anchored to the cell surface by interaction with the F antigen, which is associated with the plasma membrane, via the repeated choline binding motifs in the C-termini. This was demonstrated when a *S. pneumoniae* mutant lacking phosphocholine in its F antigen no longer retained PspA on the cell surface but released it into the culture medium (Yother *et al.,* 1998). Several other membrane associated proteins have also been identified in the pneumococcus. These include the PBPs which are integral membrane proteins involved with peptidoglycan synthesis, and autolysin, a choline-binding protein, which also interacts with the choline moieties in the F antigen. Lipoproteins, such as PsaA (Sampson, *et al.,* 1994), are also found associated with the pneumococcal plasma membrane.

1.4 The Pneumococcal Capsule

The outermost layer of the pneumococcus consists of a polysaccharide capsule, which is covalently linked to the peptidoglycan in the cell wall (Sørensen *et al.,* 1990).

The nature of this covalent linkage is unknown. The polysaccharide capsule is the thickest layer, completely masking other structures during exponential growth.

1.4.1. Capsule serotyping

Serological analysis of *S. pneumoniae* isolates over the past four decades has resulted in the recognition of 90 structurally distinct capsule types (Henrichsen, 1995). Originally two typing systems were employed; the American system, which simply assigned serotype numbers in the order identified, and the Danish system which groups antigenically similar serotypes. The Danish system is now most widely used and will be used throughout this thesis.

Since the 1930s, serotyping of pneumococcal isolates has been routinely performed using the quellung (or capsular swelling) reaction with type-specific antisera. When the specific antibody binds to the capsule of a pneumococcus, it results in a change in its refractive index; the cells appear swollen and are clearly visible under bright field or phase-contrast microscopy.

A list of type designations and their antigenic formulas are shown in **Table 1.1**. Antigenic formulas are determined by the pattern of antigenic reactivity of factor sera. Factor sera are developed by cross-absorption of type-specific antisera (Henrichsen, 1995). Antigen 'a' is a factor characteristic of the types of a group or an individual type. The letters b, c, d, etc. indicate additional partial antigens which are characteristic of only some types within a group. For example the two factor sera 6b and 6c are required to distinguish types 6A and 6B and are made by reciprocal absorption of type-specific antisera. The factor sera often cross-react with types from a different group, for example, factor serum 7h reacts with serotypes 7B, 7C, 19B, 19C, 24F, 24B and 40. The antigenic cross-reactivity between serotypes is illustrated in **Table 1.1**.

Table 1.1. Antigenic Formulas for the 90 serotypes of pneumococci

Group	Type	Antigenic Formula	Group	Type	Antigenic Formula	Group	Type	Antigenic Formula
1		1a	15	15F	15a, 15b, 15c, 15f	28	28A	28a, 28c, 23d
2		2a		15A	15a, 15c, 15d, 15g	29		29a, 29b, 13b
3		3a		15B	15a, 15b, 15d, 15e, 15h	31		31a, 20b
4		4a		15C	15a, 15d, 15e	32	32F	32a, 27b
5		5a	16	16F	16a, 16b, 11d		32A	32a, 32b, 27b
6	6A	6a, 6b		16A	16a, 16c	33	33F	33a, 33b, 33d
	6B	6a, 6c	17	17F	17a, 17b		33A	33a, 33b, 33d, 20b
7	7F	7a, 7b		17A	17a, 17c		33B	33a, 33c, 33d, 33f
	7A	7a, 7b, 7c	18	18F	18a, 18b, 18c, 18f		33C	33a, 33c, 33e
	7B	7a, 7d, 7e, 7h		18A	18a, 18b, 18d		33D	33a, 33c, 33d, 33f, 6a
	7C	7a, 7d, 7f, 7g, 7h		18B	18a, 18b, 18e, 18g	34		34a, 34b
8		8a		18C	18a, 18b, 18c, 18e	35	35F	35a, 35b, 34b
9	9A	9a, 9c, 9d	19	19F	19a, 19b, 19d		35A	35a, 35c, 20b
	9L	9a, 9b, 9c, 9f		19A	19a, 19c, 19d		35B	35a, 35c, 29b
	9N	9a, 9b, 9e		19B	19a, 19c, 19e, 7h		35C	35a, 35c, 20b, 42a
	9V	9a, 9c, 9d, 9g		19C	19a, 19c, 19f, 7h	36		36a, 9e
10	10F	10a, 10b	20		20a, 20b, 7g	37		37a
	10A	10a, 10c, 10d	21		21a	38		38a, 25b
	10B	10a, 10b, 10c, 10d, 10e	22	22F	22a, 22b	39		39a, 10d
	10C	10a, 10b, 10c, 10f		22A	22a, 22c	40		40a, 7g, 7h
11	11F	11a, 11b, 11e, 11g	23	23F	23a, 23b, 18b	41	41F	41a, 41b
	11A	11a, 11c, 11d, 11e		23A	23a, 23c, 15a		41A	41a
	11B	11a, 11b, 11f, 11g		23B	23a, 23b, 23d	42		42a, 20b, 35c
	11C	11a, 11b, 11c, 11d, 11f	24	24F	24a, 24b, 24d, 7h	43		43a, 43b
	11D	11a, 11b, 11c, 11e		24A	24a, 24c, 24d	44		44a, 44b, 12b, 12d
12	12F	12a, 12b, 12d		24B	24a, 24b, 24e, 7h	45		45a
	12A	12a, 12c, 12d	25	25F	25a, 25b	46		46a, 12c, 44b
	12B	12a, 12b, 12c, 12e		25A	25a, 25c, 38a	47	47F	47a, 35a, 35b
13		13a, 13b	27		27a, 27b		47A	47a, 43b
14		14a	28	28F	28a, 28b, 16b, 23d	48		48a

Adapted from Henrichsen (1995), Table 6.

1.4.2 Chemical structure of the capsular polysaccharide repeat units

The structure of the type 3 repeat unit was determined in 1941 (Reeves and Goebel, 1941), followed by type 8 in 1957 (Jones and Perry, 1957). Since then the repeat unit structures of more than half of the capsule types have been determined, mostly in the 1980s (reviewed by van Dam *et al.*, 1990). The chemical structures of the capsular repeat units vary greatly, ranging from linear polymers of one or two saccharides to complex polysaccharides with additional side chains.

Most pneumococcal capsules are negatively charged and possess acidic components such as: D-glucuronic acid (GlcA) (including types 1, 2, 3, 5, 8, 9A, 9N and 9V), phosphodiester bonds (including types 6A, 6B, 11A, 15F, 19F, 19A, and 23F), and pyruvate (type 4). To illustrate the diversity, the CPS repeat unit structures for some of the clinically most significant serotypes are shown in **Table 1.2** (Lee, 1987).

Table 1.2 Chemical structures of capsular polysaccharide repeat units

Type	Chemical Structure	Reference
1	→3)-α-D-Sug$_p$-(1→4)-α-D-Gal$_p$A-(1→3)-α-D-Gal$_p$A-(1→ (+0.3 OAc)	Lindberg *et al.*, 1980
2	→4)-β-D-Glc$_p$-(1→3)-α-L-Rha$_p$-(1→3)-α-L-Rha$_p$-(1→3)-β-L-Rha$_p$-(1→ 2 ↑ 1 α-D-Glc$_p$A-(1→6)-β-D-Glc$_p$	Jansson *et al.*, 1988
3	→3)-β-D-Glc$_p$A-(1→4)-β-D-Glc$_p$-(1→	Reeves and Goebel, 1941
4	→3)-β-D-Man$_p$*N*Ac-(1→3)-α-L-Fuc$_p$*N*Ac-(1→3)-α-D-Gal$_p$*N*Ac-(1→4)-α-D-Gal$_p$-(1→ 2 3 X H$_3$C COOH	Jones and Currie, 1988
5	→4)-β-D-Glc$_p$-(1→4)-α-L-Fuc$_p$*N*Ac-(1→3)-β-D-Hex$_p$-(1→ 3 ↑ 1 α-L-Pne$_p$*N*Ac-(1→3)-β-D-Glc$_p$A	Jannson *et al.*, 1985
6A	→2)-α-D-Gal$_p$-(1→3)-α-D-Glc$_p$-(1→3)-α-L-Rha$_p$-(1→3)-D-Rib-ol-(5-PO$_4^-$→	Rebers and Heidelberger, 1961
6B	→2)-α-D-Gal$_p$-(1→3)-α-D-Glc$_p$-(1→3)-α-L-Rha$_p$-(1→4)-D-Rib-ol-(5-PO$_4^-$→	Kenne *et al.*, 1979
7F	→6)-α-D-Gal$_p$-(1→3)-β-L-Rha$_p$-(1→4)-β-D-Glc$_p$-(1→3)-β-D-Gal$_p$*N*Ac-(1→ 2 2 4 ↑ OAc ↑ 1 1 β-D-Gal$_p$ α-D-Glc$_p$*N*Ac-(1→2)-α-L-Rha$_p$	Moreau *et al.*, 1988
8	→4)-β-D-Glc$_p$A-(1→4)-β-D-Glc$_p$-(1→4)-α-D-Glc$_p$-(1→4)-α-D-Gal$_p$-(1→	Jones and Perry, 1957
9N	→4)-α-D-Glc$_p$A-(1→3)-α-D-Glc$_p$-(1→3)-β-D-Man$_p$*N*Ac-(1→4)-β-D-Glc$_p$-(1→4)-α-D-Glc$_p$*N*Ac-(1→	Jones *et al.*, 1985
9V	→4)-α-D-Glc$_p$A-(1→3)-α-D-Gal$_p$-(1→3)-β-D-Man$_p$*N*Ac-(1→4)-β-D-Glc$_p$-(1→4)-α-D-Glc$_p$-(1→ +OAc +OAc	Perry *et al.*, 1981
14	→6)-β-D-Glc$_p$*N*Ac-(1→3)-β-D-Gal$_p$-(1→4)-β-D-Glc$_p$-(1→ 4 ↑ 1 β-D-Gal$_p$	Lindberg *et al.*, 1977
18C	→4)-β-D-Glc$_p$-(1→4)-β-D-Gal$_p$-(1→4)-α-D-Glc$_p$-(1→3)-α-L-Rha$_p$-(1→ 2 3 ↑ ↑ 1 PO$_4^-$-1-Glyc-ol AcO3-α-D-Glc$_p$	Lugowski and Jennings, 1984
23F	→4)-β-D-Glc$_p$-(1→4)-β-D-Gal$_p$-(1→4)-β-L-Rha$_p$-(1→ 2 3 ↑ ↑ 1 PO$_4^-$-2-Glyc-ol α-L-Rha$_p$	Richards and Perry, 1988

Abbreviations: Sug$_p$, 2-acetamido-4-amino-2,4,6, trideoxy-D-galactose; Hex$_p$, 2-acetamido-2,6 dideoxy-D-xylo-hexos-4-ulose; Pne$_p$*N*Ac, 2-acetamido-2,6 dideoxy-L-talose; Glc$_p$, glucose; Gal$_p$, galactose; Rha$_p$, rhamnose; Glc$_p$A, glucuronic acid; Gal$_p$A, galacturonic acid; Man$_p$*N*Ac, *N*-acetyl mannosamine; Fuc$_p$*N*Ac, *N*-acetyl fucosamine; Gal$_p$*N*Ac, *N*-acetyl galactosamine; Glc$_p$*N*Ac, *N*-acetyl glucosamine; ribitol; Glyc-ol, glycerol; OAc, O-acetyl; PO$_4^-$, phosphate.

1.4.3 The chemical structure of group 19 CPS

Of particular relevance to this thesis is the chemical structure of the CPS from group 19 pneumococci. Group 19 consists of four immunologically cross-reactive *S. pneumoniae* types (19F, 19A, 19B and 19C). The CPS repeat unit structures for the members of group 19 are shown in **Table 1.3**. Two different structures have been proposed for type 19A. The expression of these two different repeat units has been reported to be dependent on culture conditions (Lee *et al.*, 1987).

Group 19 pneumococci are a clinically significant cause of pneumococcal disease, although prevalence of the component serotypes varies. In one study (Robbins *et al.*, 1983), group 19 pneumococci accounted for 7% of all isolates from cases of invasive disease. Of these, 65% were caused by 19F, 34% by 19A, 1% by 19B; 19C was a very rare cause of disease in this study.

Table 1.3 Chemical structures of the capsular polysaccharide repeat units for group 19

Type	Chemical Structure	Reference
19F	→4)-β-D-Man$_p$*N*Ac-(1→4)-α-D-Glc$_p$-(1→2)-α-L-Rha$_p$-(1-PO_4^-→	Ohno *et al.*, 1980
19A	→4)-β-D-Man$_p$*N*Ac-(1→4)-α-D-Glc$_p$-(1→3)-α-L-Rha$_p$-(1-PO_4^-→	Katzenellenbogen and Jennings, 1983
	→4)-β-D-Man$_p$*N*Ac-(1→4)-α-D-Glc$_p$-(1→2)-α-L-Rha$_p$-(1-PO_4^-→ 2 3 β-D-Glc$_p$*N*Ac-(1→3)-β-D-Gal$_p$-(1- PO_4^- PO_4^-→1)-α-L-Fuc$_p$	Lee and Fraser, 1980
19B	→4)-β-D-Man$_p$*N*Ac-(1→4)-β-D-Glc$_p$-(1→4)-β-D-Man$_p$*N*Ac-(1→4)-α-L-Rha$_p$-(1-PO_4^-→ 3 ↑ 1 β-D-Rib$_p$-(1→4)-α-L-Rha$_p$	Beynon *et al.*, 1991
19C	β-D-Glc$_p$ 1 ↓ 6 →4)-β-D-Man$_p$*N*Ac-(1→4)-β-D-Glc$_p$-(1→4)-β-D-Man$_p$*N*Ac-(1→4)-α-L-Rha$_p$-(1-PO_4^-→ 3 ↑ 1 β-D-Rib$_p$-(1→4)-α-L-Rha$_p$	Beynon *et al.*, 1991

Abbreviations: Glc$_p$, glucose; Rha$_p$, rhamnose; Man$_p$*N*Ac, N-acetyl mannosamine; Gal$_p$, galactose; Fuc$_p$*N*Ac, *N*-acetyl fucosamine; Glc$_p$*N*Ac, *N*-acetyl glucosamine; Rib$_p$, ribose; PO_4^-, phosphate.

1.4.4 The relationship between serotype and virulence

The importance of the capsule in pneumococcal infection is due to its antiphagocytic properties (van Dam *et al.*, 1990). Immunity to pneumococcal infection is serotype dependent and the presence of specific antibody to the capsule results in opsonisation and rapid clearance of the invading pneumococci.

Not all pneumococcal serotypes are equally invasive, and both the capsule type and the amount of CPS expressed affect the relative virulence of pneumococci (Austrian, 1981a). MacLeod and Krauss (1950) showed that the quantity of CPS produced by each of several mutants influenced their virulence in mice. Recently, the CPS serotype expressed by otherwise isogenic pneumococcal strains, generated by *in vitro* or *in vivo* transformation, was shown to directly affect their virulence for mice (Kelly *et al.*, 1994; Nesin *et al.*, 1998). However, the genetic background of these strains was also important, and virulence appeared to be dependent on a combination of CPS and other genetic factors (Kelly *et al.*, 1994). The difference in virulence between serotypes is probably due in part to differences in the ability of the CPS to prevent activation of the alternative complement pathway, to resist phagocytosis and to induce an antibody response (AlonsoDeVelasco *et al.*, 1995).

1.5 Distribution of Capsule Types

In 1979 a worldwide pneumococcal typing surveillance study was initiated by the World Health Organisation. Since then, more than 25,000 strains have been typed in Denmark alone. Even though a large amount of information about serotype distribution is available, analysis of this data is complicated by the fact that serotype prevalence varies

with site of infection, age, time and geographical location (Nielsen and Henrichsen, 1992).

1.5.1 Capsule types associated with paediatric pneumococcal disease

Many epidemiological studies investigating the distribution of pneumococcal serotypes causing paediatric invasive disease have been reported in the literature. The results from many of these studies are summarised in **Table 1.4**. Serotypes/groups 6, 14,

Table 1.4 Incidence of invasive disease in children

Study		No. of	Most common serotypes/groups (% of total)													
Country, Year	Age	Isolates	1	2	3	4	5	6	7	9	11	12	14	18	19	23
Europe[a], 1982-1987	0-14	2,390	7	0.8	1.9	3.4	3.2	15.6	6.9	4.1	0.8	1.7	13.3	8.2	12	6.3
Spain[b], 1979-1993	0-5	167	3	0.6		2.4	7.2	17.4	3	4.8		0	2.6	4,8	17.4	15
Belgium[b], 1991	0-5	77	5.2	0		2.6	1.3	7.8	6.5	9		0	28.6	9.1	15.6	5.2
Denmark[b], 1991-1992	0-14	164	6.7	0		1.8	1.8	20.1	6.7	6.7		0.6	13.4	16.5	10.4	2.4
Finland[b], 1985-1989	0-2	235	0	0		6	0	18.4	8.5	6		0.3	21.5	7.6	17.4	7.3
Boston, Mass.[c], 1957-1969	0-15	71	<2	0	0	8.5	2.8	11.3	<2	2.8	0	2.8	29.6	12.7	12.7	5.6
Chicago, Illinois[d], 1967-1976	0-15	293	4.4	0	3.8	5.8	<2	16	4.1	7.9	0	<2	17.1	8.9	10.6	9.6
Birmingham, Alabama[e], 1975-1978	0-15	114	3.5	0	0	7.1	0.8	20.2	2.6	4.4	0	0	15.8	13.1	9.6	12.3
USA[f], 1978-1994	0-6	3570	1.1	0	0.8	6.7	0.2	17.2	1.6	7.2		0.6	28	8.2	14	6.9
Australia[g], 1970-1979	0-18	219	4.6	0		9.1	0.5	10.5	2.3	6.4		1.4	22.4	11.9	10	7.8
Australian Aborigines[h], 1989-1994	0-14	114	5.3	0	0	6.1	0	13.1	4.4	11.4	0	8.8	9.6	7	5.3	8.8
Israel[i], 1988-1990	0-12	213	13.4	0		2.7	16.5	10.7	5.4	10.7		3.1	14.7	6.7	9.4	3.6
South Africa[b], 1987-1991	0-5	1138	10.5	1.1		3.5	1.7	28	2.8	1.8		1.1	19.2	3.9	13	3.2
Mexico[b], 1992-1993	0-5	120	0.8	2.5		0.8	0.8	17.5	0.8	6.7		0.8	9.2	2.5	14.2	20.8
Brazil[j], 1977-1988	0-2	308	6.4	1.9		2.6	10.3	18.1	1.9	5.1		1.6	10.7	10.5	8	5.4
Uruguay[k], 1987-1989	0-5	48	6.3	0		2.1	14.6	4.2	4.2	10.5		0	39.6	0	0	2.1
Gambia[b], 1991	0-5	59	10.2	0		0	8.5	15.3	0	5.1		8.5	32.2	1.7	5.1	0
Egypt[l,b], 1977-1978, 1992	0-6	86	31.4	2.3		2.3	4.7	8.1	3.5	9.3		5.8	7	3.5	2.3	1.2
Pakistan[m], 1986-1990	0-5	168	0.6	0		0	4.2	10.7	0.6	6.5		0	1.2	2.4	38.7	4.2
Papua New Guinea[n,o], 1980-1987	0-5	151	2.6	7.3		0.7	10.6	6.6	13.2	2		2.6	6.6	0.7	7.3	7.3
Rwanda[p], 1984-1990	0-15	130	22.3	0.7		0	14.6	10	0.8	0.8		0.8	14.6	3.1	3.8	0.8

a. Nielsen and Henrichsen (1992); b. Snaidack *et al.* (1995); c. Burke *et al.* (1971); d. Jacobs *et al.* (1979); e. Gray *et al.* (1979); f. Butler *et al.* (1995); g. Hansman (1983); h. Gratten *et al.* (1996); i. Dagan *et al.* (1992); j. Tauney *et al.* (1990); k. Mogdasy *et al.* (1992); l. Guirguis *et al.* (1990); m. Mastro *et al.* (1991); n. Gratten and Montgomery (1991); o. Gratten *et al.* (1985); p. Bogaerts *et al.* (1993). The six most common serotypes in each study are shaded.

18, 19 and 23 are consistently among the most common disease causing isolates in children, although their rank order differs significantly. Within serogroups 6, 18, 19 and 23, serotypes 6B, 18C, 19F and 23F are the most common isolates (Nielsen and Henrichsen, 1992; Butler *et al.*, 1995). Interestingly, these serotypes are also the most poorly immunogenic in children (Douglas *et al.*, 1983). Other serotypes/groups, such as type 1, 4, 5, 7F, 9 and 12F, are a considerable cause of disease in only some of the studies. For example, type 1 is a common cause of disease in many developing countries and in parts of Europe but not in the USA, whereas type 4 is most commonly isolated in the USA. In the USA, serotypes 6B, 14, 18C, 19F and 23F along with types 3 and 19A are the most common pneumococcal isolates causing otitis media as shown in **Table 1.5** (Butler *et al.*, 1995). Generally the most common disease causing serotypes are also those most frequently carried in the nasopharynx (**Table 1.6**). However, there are some exceptions, 11A, for example, is a rare cause of pneumococcal disease in children but it is frequently carried. In Papua New Guinea, groups 15 and 33 (not shown in **Table 1.6**) are also

Table 1.5 Incidence of otitis media in children

Study		No. of	Most common serotypes/groups (% of total)													
Country, Year	Age	Isolates	1	2	3	4	5	6	7	9	11	12	14	18	19	23
USA[a], 1978-1994	0-6	249			9	4		14		3			15	2	22	13
Birmingham, Alabama[b] 1975-1978	0-15	396	1.8	0.3	10.6	3.5	0.3	9.6	2.5	4.3	1.8	0	9.6	4.5	28.2	11.9

a. Butler *et al.* (1995); b. Gray *et al.* (1979). The six most common serotypes in each study are shaded.

Table 1.6 Carriage of pneumococcal strains

Study		No. of	Most common serotypes/groups (% of total)													
Country, Year	Age	Isolates	1	2	3	4	5	6	7	9	11	12	14	18	19	23
Birmingham, Alabama[a] 1975-1978	0-15	245	2	0	14.3	5.7	0	22.9	1.2	3.7	8.2	0.4	8.2	10.2	20.4	16.3
Birmingham, Alabama[b] 1974-1975	0-2	573	0.2	0	2.8	3.3	0	27	1.2	4.5	4.2	0	7	2.6	17.4	15.4
Papua New Guinea[c] 1985-1987	0-7	1449	0.1	0.1	1.6	1.9	0.3	26.6	1.7	3.2	1	0.8	8.1	1.5	23.3	14.1

a. Gray *et al.* (1979); b. Gray, *et al.* (1980); c. Montgomery *et al.* (1990). The six most common serotypes in each study are shaded.

commonly isolated from the nasopharynx with carriage rates of 5.6% and 4.8% respectively (Montgomery *et al.*, 1990), but these serotypes rarely cause disease.

1.5.2 Capsule types associated with invasive pneumococcal disease in adults

Although there have been many epidemiological studies investigating the distribution of pneumococcal serotypes causing disease, few report serotype prevalence only in adults. The serotypes responsible for most adult pneumococcal disease are quite different to those that cause paediatric disease. Serotypes 1, 3, 4, 7F, 8 and 14 were associated with most disease in adults in the studies summarised in **Table 1.7**.

Table 1.7 Incidence of invasive disease in adults

Study		No. of	Most common serotypes (% of total)														
Country, Year	Age	Isolates	1	2	3	4	5	6	7	8	9	11	12	14	18	19	23
Europe and Israel[a], 1982-1987	>14	6,376	8.9	0.4	9.1	6.6	1.5	6.3	7.8	5.1	6.8	2.2	3.2	7.9	2.2	6.6	4.4
Australian Aborigines[b], 1989-1994	>14	107	15.9	0	12.1	7.5	0	0.9	13.1	6.5	3.8	0.9	9.3	1.9	0.9	0.9	0.9

a. Nielsen and Henrichsen (1992); b. Gratten *et al.* (1996). Shaded boxes represent the six most common serotypes in each study.

1.5.3 Geographic distribution of invasive pneumococci

The serotypes that most commonly cause invasive pneumococcal disease in different countries are summarised in **Table 1.8**. Serotypes/groups 6, 14 and 19, are consistently important in all geographic areas. Others occur more frequently in some countries. Types 1 and 5 are a major cause of disease in developing countries, such as Africa, China and Israel, whereas group 18 is only important in developed countries (Snaideck *et al.*, 1995). The serotype distribution among indigenous and disadvantaged minorities in developed countries reflects that of the country to which they belong.

However, they often live in poor conditions and their disease rates are far greater than the rest of the population (McIntyre, 1997).

Table 1.8 Incidence of invasive disease (all ages)

Study	No. of	Most common serotypes (% of total)														
Country, Year	Isolates	1	2	3	4	5	6	7	8	9	11	12	14	18	19	23
Japan[a], 1984	430	1.2	0	12.7	2.4	1.2	11.8	1.4	0.3	5.6	4.1	1.4	4.9	2.5	13	6.8
Taiwan[a], 1984	170	2.9	0	5.6	3.5	0.6	9.5	1.8	1.8	4.1	2.9	0	10	1.2	10.5	7.1
China[a], 1984	448	9.8	7.8	5.6	1.1	14.1	11.6	2.5	2	1.6	0.9	2.7	7.6	1.8	9.2	7.4
Europe and Israel[b], 1982-1987	10,298	8	0.6	6.7	5.3	2.3	9.1	7.6	3.7	5.9	1.8	2.7	8.7	3.8	8.2	5.2
Australian Aborigines[c], 1989-1994	221	11.3	0	6.4	7.4	0	9.4	5.9	4.4	8.4	0.5	8.9	6.4	4.4	3.4	5.4
South Africa[d], 1979-1986	4766	24.2	4	5.9	4.4	3.7	18.5	5	6.3	2.7	0.7	2.2	9.3	4.3	10.2	1.6
USA[e], 1970-1983	4676	5	0.2	7.1	10.6	1.6	6.4	5.9	9.8	5.4	1	4	9.2	4.4	5.5	4.7
CDC[e], 1978-1983	1900	3.2	0	4.5	9.4	0.8	11.7	3.5	3.5	7.8	0.7	2.2	14.7	7.5	11	7.1

a. Lee (1987); b. Nielsen and Henrichsen (1992); c. Gratten *et al.* (1996); d. Klugman and Koornhof (1988); e. Robbins *et al.* (1983). Shaded boxes represent the six most common serotypes in each study.

1.5.4 Changes in serotype distribution over time

Surveys of the serotype distribution in the 1930s showed that most of the pneumococcal disease recorded was caused by a small number of types, principally 1, 2, 3, 5, 7 and 8 (Finland and Barnes, 1977). The apparent small number of serotypes causing disease may have been partly due to the limitations in the number of serotypes that could be positively identified at the time. Types 1, 2 and 3 were the most commonly isolated types in Boston City Hospital before 1950 (Finland and Barnes, 1977), but collectively accounted for less than 5% of all isolates in 1979-1982 (Barry *et al.*, 1984). The occurrence of type 5 was also declining. Recent investigations in the developing countries have shown a type distribution similar to that in the industrialised countries before 1950 (Klugman and Koornhof, 1988; Snaideck *et al.*, 1995; Scott *et al.*, 1996; Sankilampi, 1997).

1.6 Antimicrobial Resistance in the Pneumococcus

The emergence of clinically significant resistance to antibiotics was first noted in 1967 when a penicillin resistant pneumococcus was isolated, in Australia, from the sputum of a patient (Hansman and Bullen, 1967). Subsequently, penicillin resistant pneumococci were isolated in New Guinea, and the prevalence of penicillin resistant isolates from this region increased dramatically from 12% in 1970 (Hansman *et al.*, 1974) to 33% in 1980 (Gratten *et al.*, 1980). Multiply resistant pneumococci (resistant to three or more classes of antibiotics) were first isolated in South Africa (Jacobs *et al.*, 1978); these strains showed a range of resistance patterns to several antibiotics including penicillin, tetracycline and chloramphenicol. Today, antibiotic resistant pneumococci are common throughout the world and are rapidly increasing in prevalence (Appelbaum, 1992). This is becoming a major clinical problem as management of multiply resistant pneumococcal infections become more complicated and requires the use of more expensive alternative antimicrobial agents. The impact of increased antimicrobial resistance on the mortality rate has not yet been clearly defined (Centers for Disease Control and Prevention, 1997).

1.6.1 Geographical distribution of antimicrobial resistance

Antibiotic resistance in pneumococci is increasing all over the world. The pattern of antibiotic resistance is uneven both between countries and within countries. South Africa, Spain and Eastern Europe report rates of penicillin resistance of up to 50% of all isolates, compared with less than 10% in most of Western Europe and USA (Schreiber and Jacobs, 1995). In Australia, resistance to antibiotics is also increasing. In 1989 less than 2% of all isolates were penicillin resistant, but by 1994 that had increased to 5.8% (Collignon and Bell, 1996). Interestingly, 8% of all isolates were found to be resistant to at

least three antibiotics, including penicillin, erythromycin, tetracycline, and/or chloramphenicol. In some Aboriginal communities, levels of antibiotic resistant pneumococci are much higher. One study, in Darwin, showed penicillin resistance to be as high as 30% (Skull *et al.*, 1996). Recent data from the Women's and Children's Hospital (WCH) in Adelaide shows that 40% of *S. pneumoniae* isolates from WCH patients are resistant to penicillin, compared with only 15% two years ago (J. Bell, WCH, personal communication).

1.6.2 The spread of antibiotic resistance

There is a link between antibiotic usage and the increasing prevalence of antibiotic resistance in pneumococci. One study from Iceland (Aarson *et al.*, 1996) confirms this linkage among children attending day-care. In this study, the overall pneumococcal carriage rate was 52.7%, with 9.7% of isolates being penicillin resistant. Only 1.8% of children who had no courses of antibiotics in the previous year carried penicillin resistant strains compared with 14% who had antibiotics, and 61% of those who had received antibiotics in the previous two to seven weeks.

1.6.3 Antibiotic resistance among different pneumococcal serotypes

Antibiotic resistance is most prevalent in *S. pneumoniae* serotypes/groups 6, 9, 14, 19 and 23, which are commonly associated with carriage in children (**Table 1.4**, Schreiber and Jacobs, 1995). These serotypes accounted for more than 80% of the penicillin resistant infections among Aboriginal children in Darwin (Skull *et al.*, 1996).

1.7 The Pneumococcal Polysaccharide Vaccine

The first commercially available pneumococcal vaccine was licensed in the USA in 1977. This vaccine consisted of 50 μg of purified capsular polysaccharide of 14 of the most common disease related serotypes (Hilleman *et al.,* 1981). These 14 serotypes (1, 2, 3, 4, 6A, 7F, 8, 9N, 12F, 14, 18C, 19F, 23F and 25F) accounted for 68% of all cases of pneumococcal disease in the USA. In 1983, the 14-valent vaccine was replaced by a 23-valent formulation containing 25 μg of the purified capsular polysaccharide from serotypes 1, 2, 3, 4, 5, 6B, 7F, 8, 9N, 9V, 10A, 11A, 12F, 14, 15B, 17F, 18C, 19F, 19A, 20, 22F, 23F and 33F, which are responsible for more than 85% of all cases of pneumococcal disease in North America and Europe (Robbins *et al.,* 1983). The halving of the amount of each capsular polysaccharide did not affect the antibody response to the vaccine.

1.7.1 Antibody response

Pneumococcal CPS induces type-specific antibodies that enhance opsonisation, phagocytosis and killing of pneumococci. After vaccination, a type-specific antibody response develops in >80% of healthy young adults (Musher *et al.,* 1990) and this may persist for up to 10 years (Mufson *et al.,* 1987). However, immune responses are usually not consistent among all 23 serotypes in the vaccine and the level of antibody response required for protection against pneumococcal disease has not been clearly defined.

The immune response to the pneumococcal vaccine is dependent on the age, health and immunocompetency of the individual. The antibody response to the vaccine is generally good in healthy young adults, but it is often diminished or absent in young children aged <2 years and immunocompromised individuals (Centers for Disease Control

and Prevention, 1997). Bacterial polysaccharides induce antibodies primarily by T-cell independent mechanisms and are therefore poorly immunogenic in children aged <2 years whose immune systems are immature. The immune response to some common paediatric pneumococcal serotypes is also suboptimal in children aged 2-5 years (Douglas *et al.*, 1983) and does not reach "adult levels" until age 8 to 10 years (Paton *et al.*, 1986).

1.7.2 Conjugate polysaccharide-protein vaccines

Conjugation to a protein carrier dramatically increases the immunogenicity of polysaccharide antigens in young children, converting the polysaccharide from a T-cell independent to a T-cell dependent antigen, capable of eliciting immunological memory. The success of this approach has been demonstrated with *Haemophilus influenzae* type b (Hib). Since the licensing of the first conjugate Hib vaccine in 1987, for use in children at least 18 months of age, the incidence of Hib meningitis has significantly decreased. An American study (Adams *et al.*, 1993) shows a sharp decrease of 56% in the incidence of Hib meningitis between 1989 and 1991. Interestingly, there was also a substantial decrease in the incidence of Hib meningitis in infants aged less than one year during this same period even though the vaccine was not licensed for use in this age group until 1990. Hib conjugate vaccination has been shown to lower the rate of nasopharyngeal carriage (Takala *et al.*, 1991), thus, vaccination of older children can reduce the probability of spread of Hib to younger unvaccinated children.

Development of a pneumococcal conjugate vaccine is more complicated than for Hib because of the numerous pneumococcal serotypes that cause disease. The number of pneumococcal serotypes that could be included in a conjugate vaccine formulation is probably limited. The pneumococcal serotypes that cause most paediatric pneumococcal disease are also those that are commonly antibiotic resistant, thus initial vaccination against

these serotypes would provide the greatest benefit to the community. Several pneumococcal conjugate formulations are currently being evaluated.

An Israeli study, involving vaccination of children 12 to 18 months with a conjugate vaccine consisting of polysaccharides of types 4, 6B, 9V, 14, 18C, 19F and 23F conjugated to the outer membrane complex of *Neisseria meningitidis* serogroup B, showed a significant reduction in the nasopharyngeal carriage rates one year after vaccination (Dagan *et al.*, 1996). The carriage rates of pneumococcal types included in the vaccine and antibiotic resistant pneumococci were significantly lower than in the control group which had received the current 23-valent vaccine. However, as expected, the carriage rate of pneumococcal serotypes not included in the conjugate vaccine was unaffected.

Another study, investigating the immunogenicity in Gambian infants of a pentavalent vaccine consisting of types 6B, 14, 18C, 19F and 23F conjugated to the diphtheria toxin mutant protein CRM_{197}, also showed a significant reduction in the nasopharyngeal carriage of pneumococci belonging to the vaccine serotypes two years after vaccination (Leach *et al.*, 1996; Obaro *et al.*, 1996). However, in this study, the carriage rate of non-vaccine serotypes was significantly higher than in the unvaccinated control group. Thus, widespread use of conjugate pneumococcal vaccines may alter the pneumococcal serotype distribution associated with colonisation, and potentially also the types associated with invasive disease.

1.8 Polysaccharide Biosynthesis

Bacteria produce a large variety of polysaccharides associated with the outer surface of the cell. These include capsular polysaccharide (CPS), lipopolysaccharide O-antigen

(LPS), lipooligosaccharide (LOS), exopolysaccharide (EPS), enterobacterial common antigen (ECA), and cell wall polysaccharides such as peptidoglycan and teichoic acid. These polysaccharides are diverse in their structure and are often species and type-specific antigens.

1.8.1 Some common features in polysaccharide biosynthesis

Most polysaccharides are synthesised from activated sugar intermediates. Monosaccharides can be activated by the addition of nucleoside diphosphates by nucleotidyltransferases. Specific activated nucleotide sugars can also be synthesised by enzymatic modification of the monosaccharide part of a nucleotide sugar to produce derivatives such as deoxy sugars, aminodeoxy sugars and glycuronic acids (Shibaev, 1986). However, the biosynthesis of some polysaccharides does not involve nucleotide sugars, such as glucan biosynthesis in *Streptococcus mutans* (as discussed in section 1.8.5).

Biosynthesis of many polysaccharides involves a reversible first step, that is transfer of sugar 1-phosphate to the lipid carrier undecaprenol phosphate (und-P). Only a limited number of sugars are known to be transferred to und-P. RfbP from *Salmonella enterica* serovar typhimurium (Jiang *et al.*, 1991) has been shown to transfer galactose (Gal)-1-phosphate to und-P, and several functional homologues have been identified in other polysaccharide systems. Some of the identified homologues have been shown to transfer a different sugar, glucose (Glc)-1-phosphate, to a lipid carrier, possibly und-P (Kolkman *et al.*, 1996). Another enzyme known to initiate polysaccharide synthesis in Gram-negative bacteria is Rfe which transfers Glc*N*Ac-1-phosphate to und-P (Meier-Dieter *et al.*, 1990).

Polymerisation of the polysaccharide usually occurs via one of two different mechanisms. These two main mechanisms differ in the direction of chain elongation,

which occurs at either the reducing or the non-reducing end of the polysaccharide chain (Whitfield, 1995). They are described in greater detail below. However, at this stage, the precise mechanisms involved in subsequent transport across the outer membrane (in Gram-negative bacteria) and covalent attachment (if required) of the polysaccharide to the outer surface are poorly understood.

1.8.2 Polysaccharide biosynthesis via lipid-linked repeat unit intermediates

Biosynthesis of many polysaccharides, including: LPS O-antigen in *S. enterica*, *Escherichia coli* and *Shigella flexneri* (Whitfield, 1995), peptidoglycan (Fuchs-Cleveland and Gilvarg, 1976), the EPS xanthan gum in *Xanthomonas campestris* (Ielpi *et al.*, 1993), EPS from *Lactococcus lactis* (van Kranenburg *et al.*, 1997), succinoglycan from *Rhizobium meliloti* (Reuber and Walker, 1993) and CPS from *Aerobacter aerogenes* (Troy *et al.*, 1971), involves assembly of an oligosaccharide repeat unit intermediate on und-P (**Fig. 1.3A**). Biosynthesis of the repeat units occurs through a series of interdependent and sequential reactions, each mediated by a specific glycosyl transferase.

The initial reaction involves the transfer of sugar 1-phosphate from the nucleotide-diphosphate precursor to und-P (step 1). The subsequent transferase reactions involve the formation of glycosidic linkages to the und-P-P-sugar intermediate by transfer of the sugar moiety only, releasing the nucleotide-phosphate (steps 2 & 3). This method allows the synthesis of complex repeat unit structures with complete fidelity, due to the substrate specificity of each of the glycosyl transferases. The und-P-P-oligosaccharide repeat units are synthesised on the cytoplasmic side of the plasma membrane and are translocated to the outer surface by the repeat unit transporter, an integral membrane protein. The polysaccharide polymerase then links the repeat units to form a polysaccharide chain. The polysaccharide polymerase transfers the growing polysaccharide chain from one und-P

carrier to the non-reducing end of a newly synthesised und-P-P-oligosaccharide repeat unit (steps 4 & 5). Growth of the polymer therefore occurs at the reducing terminus (the end closest to und-P) in a block-wise fashion (Whitfield, 1995).

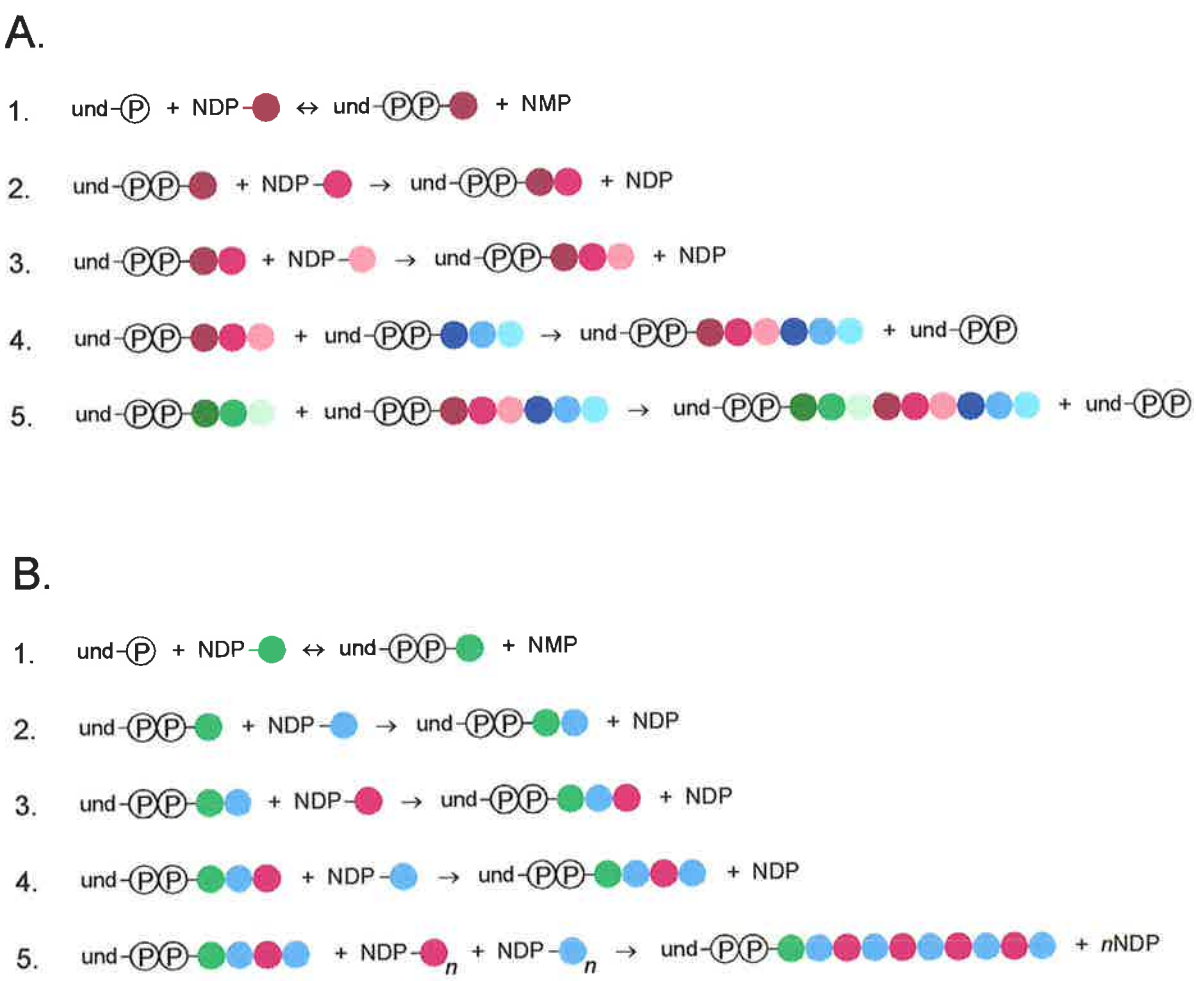


Figure 1.3. Polymerisation of polysaccharides. Two distinct polymerisation models have been identified which can be distinguished by the direction of chain elongation. **A.** Polysaccharide biosynthesis via lipid-linked repeat unit intermediates. Polymerisation of a polysaccharide with a trisaccharide repeat unit in a block-wise fashion where the growth of the polymer occurs at the reducing terminus or the end closest to und-P. Coloured circles represent sugars and groups of three similarly coloured circles represent the trisaccharide repeat units. **B.** Polysaccharide biosynthesis involving ABC transporters. Polymerisation of a polysaccharide with a disaccharide repeat unit is illustrated. Polymerisation is initiated by the addition of a sugar-1-phosphate, that is not necessarily present in the repeat unit structure of the polysaccharide, and continues by sequential addition of sugars at the non-reducing terminus or the end furthest from und-P. Different coloured circles represent different sugars. (Adapted from Whitfield, 1995)

1.8.3 Polysaccharide biosynthesis involving ABC transporters

Polymerisation at the non-reducing end (**Fig. 1.3B**) involves the sequential transfer of sugars to the non-reducing end of the und-P polymer. Polysaccharides synthesised by this method tend to have simple repeat unit structures of one or two sugars. Some polysaccharides synthesised in this way include: CPS from *E. coli* K1, K4, K5, K92 and *Haemophilus influenzae* type b (Roberts, 1996), and *N. meningitidis* group B (Frosch *et al.*, 1989), O-polysaccharide from *K. pneumoniae* serotype O1 (Bronner *et al.*, 1994), teichoic acid from *Bacillus subtilis* (Lazarevic and Karamata, 1995), amylovoran from *Erwinia amylovora* (Bugert and Geider, 1995), and Nod factors from *R. meliloti* (Vázquez *et al.*, 1993). The pathway is initiated by the transfer of a sugar-1-phosphate residue to und-P (step 1). This sugar is not necessarily present in the repeat unit structure of the polysaccharide. The sugars are then transferred directly from NDP-sugar precursors to the growing polysaccharide chain by specific glycosyl transferases (steps 2 to 5). Growth of the polymer occurs at the end furthest from the und-P carrier (Whitfield, 1995).

Transport of the completed polysaccharide chain across the cytoplasmic membrane into the periplasm is facilitated by ABC transporters. ABC transporters are ubiquitous and are involved in a diverse range of import and export systems. Those involved in polysaccharide export form a closely related sub-class within this large group of proteins (Reizer *et al.*, 1992). They are a two component system, comprising an integral membrane protein and a hydrophilic ATP-binding protein. The transport complex consists of two of each component (Reizer *et al.*, 1992).

1.8.4 Polysaccharide synthesis via processive glycosyl transferases

LPS O:54 from *S. enterica* serovar borreze (Keenleyside and Whitfield, 1996), cellulose from *Acetobacter xylinum* and *Agrobacterium tumifaciens* (Saxena *et al.*, 1994),

hyaluronic acid from *Streptococcus pyogenes* (DeAngelis *et al.,* 1993), type 3 CPS from *S. pneumoniae* (Arrecubieta *et al.*, 1995; Dillard *et al.,* 1995), alginate from *Pseudomonas aeruginosa* (Maharaj *et al.,* 1993) and chitin from *Saccharomyces cerevisiae* and *Candida albicans* (Cabib *et al.,* 1983) are all polysaccharides with simple mono- or di-saccharide repeat unit structures. Biosynthesis of these polysaccharides requires a processive glycosyl transferase or synthase. These glycosyl transferases share some common structural features including four transmembrane domains and a large hydrophilic domain with two distinct catalytic sites capable of forming two different glycosidic linkages (Keenleyside and Whitfield, 1996). However, there may be some differences in the mechanism of polysaccharide biosynthesis between different synthases. Biosynthesis of hyaluronic acid in *S. pyogenes* and type 3 CPS from *S. pneumoniae* have been shown to require only the synthase itself for production of CPS (Arrecubieta *et al.,* 1996; DeAngelis and Weigel, 1994). Whereas, studies on the biosynthesis of O:54 LPS from *S. enterica* serovar borreze have shown that the initial steps of polysaccharide synthesis resemble those for ABC-transporter dependent pathways (Keenleyside and Whitfield, 1996). Biosynthesis of O:54 LPS is dependent on the presence of a functional *rfe* gene, thus the initial step in polysaccharide synthesis involves the transfer of sugar 1-phosphate to und-P. The polysaccharide chain is then polymerised by the synthase. No dedicated export systems have been identified for these polysaccharides. These proteins all have four C-terminal transmembrane domains (Cabib *et al.,* 1983) which could conceivably form a pore in the cytoplasmic membrane, enabling extrusion of the polysaccharide chain as it is synthesised.

1.8.5 Polysaccharide synthesis not involving lipid-linked intermediates

Some EPSs, such as glucan (a Glc polymer) and fructan (a fructose polymer) from *S. mutans* and levan (a homopolymer of fructose) from *E. amylovora* are synthesised

extracellularly and do not involve activated sugar precursors or lipid-linked intermediates (Gross *et al.*, 1992; Mukasa, 1986).

In *S. mutans*, glucan is synthesised from sucrose by glucosyl transferases (GTFs) which are exported from the cell via a signal peptide and convert sucrose to glucan and fructose. They have a catalytic site thought to be involved in sucrose hydrolysis (Mooser *et al.*, 1991) and a series of C-terminal direct repeats thought to be involved in glucan binding (Mooser and Wong, 1988). These repeats are similar to those identified in the C-termini of other Gram-positive ligand-binding proteins, such as those found in pneumococcal choline binding proteins (Wren, 1991).

1.8.6 Capsule gene clusters

Capsule gene clusters have been cloned and sequenced from a variety of Gram-negative and Gram-positive bacteria. These include *E. coli* K1 (Silver et al., 1984), K4 (Drake *et al.*, 1990), K5 (Petit *et al.*, 1995) K7, K12, K92 (Roberts *et al.*, 1986), K10, and K54 (Pearce and Roberts, 1995), *Haemophilus influenzae* type b (Kroll *et al.*, 1989), *N. meningitidis* group B (Frosch *et al.*, 1989), *Salmonella typhi* Vi antigen (Hashimoto *et al.*, 1993), *E. amylovora* (Bugert and Geider, 1995), *Erwinia stewartii* (Dolph *et al.*, 1988), *Pseudomonas solanacearum* (Huang and Schell, 1995), *K. pneumoniae* (Arakawa *et al.*, 1995), and *Staphylococcus aureus* types 1 (Lin *et al.*, 1994), 5 (Sau *et al.*, 1997), and 8 (Sau and Lee, 1996). The genes involved in capsule biosynthesis for all of these organisms are clustered together on the chromosome. The organisation of these gene clusters share some similarities. Generally, the genes which are common to all serotypes are located near the 5' and 3' ends of the locus, with the serotype specific genes in the centre. This arrangement may have contributed to the high degree of capsule diversity, with recombination events between the conserved regions resulting in acquisition

of new genes and the expression of a new capsule serotype.

1.9 The Capsule Locus of *S. pneumoniae*

As outlined above, capsule production requires a complex pathway including transport into the cell and/or synthesis of the component monosaccharides, activation of each to a nucleotide precursor, co-ordinated transfer of each sugar, in sequence, to the repeating oligosaccharide and subsequent polymerisation, export and attachment to the cell surface. Thus, the loci encoding capsule production would be expected to consist of a large number of genes. However, when the work for this thesis was initiated, very little was known about these loci. Only a single gene from the type 3 capsule locus (Arrecubieta *et al.,* 1994) and the first seven genes of the type 19F locus (Guidolin *et al.,* 1994) had been cloned and sequenced. Since then the entire sequences for the capsule loci of several different pneumococci have been completed. These will be discussed in detail in chapter 8.

1.9.1 Pneumococcal capsular transformation

Classical genetic studies carried out by Austrian *et al.,* (1959) demonstrated that the *S. pneumoniae* genes required for biosynthesis and expression of capsular polysaccharide are closely linked on the pneumococcal chromosome and are transferred as a cassette during transformation. In the vast majority of cases, incorporation of the donor *cps* locus into the recipient cell chromosome was accompanied by loss of the original *cps* locus, consistent with genetic recombination between homologous sequences flanking the serotype-specific *cps* genes.

Transformation of a non-encapsulated derivative of a type 3 strain with DNA from

a type 1 strain occasionally resulted in encapsulated transformants expressing both type 1 and type 3 CPS. Such binary encapsulated strains contained the defective type 3 *cps* locus as well as the functional type 1 locus (Austrian *et al.*, 1959). Binary encapsulation could also be achieved by transformation of the type 3 mutant with DNA from other serotypes whose CPS contained GlcA, a component of type 3 CPS. The recipient strains used in these studies had mutations in the gene which encoded the enzyme required for synthesis of UDP-GlcA. Thus, in the binary transformants, the UDP-GlcA required for synthesis of type 3 CPS could be supplied by the enzyme encoded by the second *cps* locus (Bernheimer *et al.*, 1967).

Binary transformants always expressed abundant amounts of the type 3 CPS but only small amounts of the second CPS and represented atypical recombination of the second *cps* locus into the type 3 chromosome (Bernheimer *et al.*, 1967). However, unrelated pneumococci could only be transformed to the binary phenotype by using DNA from binary strains, never by using a mixture of the DNA from the two component serotypes. This suggests a degree of linkage between the two *cps* loci in the binary strains, and possibly a tandem arrangement in some cases (Bernheimer *et al.*, 1967). Binary transformants were only ever achieved using type 3 pneumococci, thus the type 3 *cps* locus and/or its flanking sequences may differ in some way to that of other pneumococci.

1.9.2 Localisation of the type 19F capsule locus

The fortuitous discovery of the insertion sequence IS*1202* closely linked to the type 19F capsule locus (Morona *et al.*, 1994a) enabled the isolation and characterisation of part of this locus. As shown in **Fig. 1.4**, the capsule locus is located downstream from *dexB*, a gene which is part of the multiple sugar metabolism locus and encodes a glucan-1,6-α-glucosidase. IS*1202* is located in the 1 kb intergenic region and is only 87 bp from the start

of the first gene in the capsule locus.

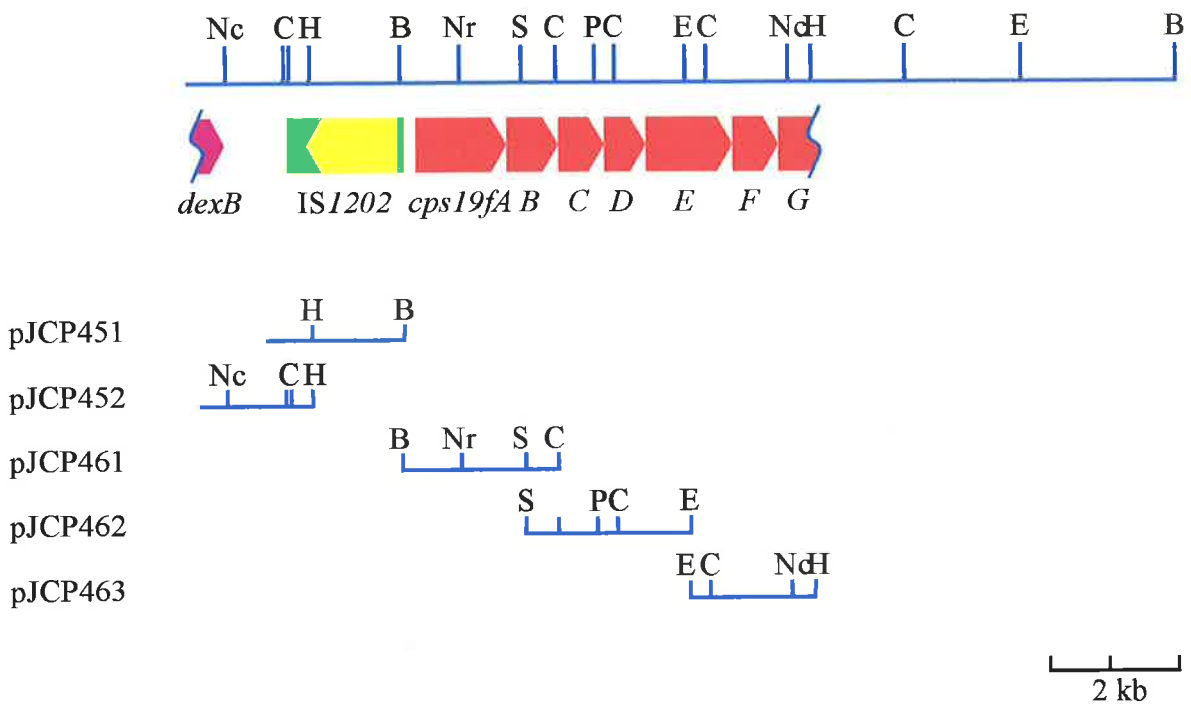


Fig. 1.4 Physical map of the *S. pneumoniae* chromosome in the vicinity of the 19F *cps* locus (Guidolin *et al.*, 1994). Arrows represent potential ORFs and the box represents IS*1202*. Gene designations are indicated below the map; genes *cps19fB-G* are abbreviated to *B-G*, respectively. Restriction sites are as follows: B, *Bam*HI; C, *Cla*I; E, *Eco*RI; H, *Hin*dIII; Nc, *Nco*I; Nr, *Nru*I; P, *Pst*I; S, *Sph*I. Various subclones of this region, designated pJCP451-2, pJCP461-3, are as shown.

1.9.3 Partial cloning and sequencing of the 19F capsule locus

The first part of the type 19F capsule locus was cloned and sequenced (Guidolin *et al.*, 1994). Examination of the sequence data revealed the presence of six complete open reading frames (ORFs) and one incomplete ORF, as shown in **Fig. 1.4**. An almost perfect consensus promoter sequence was detected 30 bp upstream of the first ORF in this operon. Each of the ORFs is preceded by a ribosome binding site and they are all closely coupled, being separated by 1-15 nucleotides.

1.9.4 Characterisation of *cps19fA-G*

The genes *cps19fA-G* are part of an operon essential for type 19F capsule biosynthesis as disruption of these ORFs resulted in the loss of type 19F capsule production (Guidolin *et al.*, 1994). The predicted functions of the seven *cps19f* ORFs, based on database homology searches (summarised in **Table 1.9** and discussed in detail below), are also consistent with the involvement of these proteins in capsule biosynthesis. Thus the genes *cps19fA-G* form the first part of the type 19F capsule locus.

Table 1.9 Predicted functions of *S. pneumoniae cps19f* ORFs

ORF	Predicted function	Similar proteins (% identity)
Cps19fA	regulation?	*B. subtilis* LytR[a] (27.6%)
Cps19fB	unknown	*S. agalactiae* CpsA[b] (63.8%)
Cps19fC	chain length regulation /export	*S. agalactiae* CpsB[b] (46%) *R. meliloti* ExoP[c] (22.5% to N-terminus)
Cps19fD	chain length regulation /export	*S. agalactiae* CpsC[b] (59.2%) *R. meliloti* ExoP[c] (29.9% to C-terminus)
Cps19fE	glucose-1-phosphate transferase	*S. agalactiae* CpsD[b] (45.6%) *S. enterica* serovar typhimurium RfbP[d] (31.4%) *X. campestris* Xps2a[e] (24.1%)
Cps19fF	*N*-acetyl mannosamine transferase	*B. subtilis* TagA[f] (31.2%) *E. coli* K12 RffM[g] (33%) *S. enterica* serovar typhimurium RffM[h] (32.5%)
Cps19fG[j]	unknown	*Haemophilus influenzae* LicD[i] (38%)

a. Lazarevic *et al.* (1992); b. Rubens *et al.* (1993); c. Glucksmann *et al.* (1993); d. Jiang *et al.* (1991); e. Ielpi *et al.* (1993); f. Mauel *et al.* (1991); g. Daniels *et al.* (1992); h. Lu and Abdelal (1993); i. Weiser *et al.* (1989); j. Cps19fG ORF is truncated.

1.9.4.1 Cps19fA

Cps19fA exhibits significant similarity (27.6% identity) to the entire *B. subtilis lytR* gene product (Lazarevic *et al.*, 1992). LytR is a basic protein that is thought to act as a transcriptional attenuator of *lytABC* (autolysin) expression and is thought to be membrane bound via a N-terminal anchoring domain. Cps19fA is also basic and contains three

hydrophobic segments near its N-terminus which may be membrane anchoring domains. Interestingly, the predicted LytR protein lacks this hydrophobic region of the sequence, but has a single hydrophobic region near its N-terminus. Thus Cps19fA may play a role as a regulator of capsule gene expression although there is no direct evidence for this.

1.9.4.2 Cps19fB

The *cps19fB* gene product has 63.8% identity with the product of the *Streptococcus agalactiae* (group B streptococcus) *cpsA* gene. The *S. agalactiae cpsA* gene has been shown to be involved in the production of type III capsular polysaccharide in group B streptococci (Rubens *et al.*, 1993). No other significant sequence similarities were detected, and the functions of either of these gene products is not known.

1.9.4.3 Cps19fC

The *cps19fC* gene product has 46% identity with the *S. agalactiae* CpsB protein (Rubens *et al.*, 1993), and 22.5% identity with the N-terminal portion of the *R. meliloti* ExoP protein (Glucksmann *et al.*, 1993). Cps19fC also has some similarity to proteins involved in chain length regulation of polysaccharides, including a conserved motif (amino acids 168-187 in Cps19fC) found near the C-terminus in this type of protein (Kolkman *et al.*, 1997b; Becker *et al.*, 1995). Cps19fC probably regulates the chain length of type 19F polysaccharide.

1.9.4.4 Cps19fD

The *cps19fD* gene product has significant similarity (59.2% identity) to the *S. agalactiae cpsC* gene product (Rubens *et al.*, 1993). In addition, the Cps19fD protein exhibits similarity (29.9% identity) to the C-terminal portion of the *R. meliloti* ExoP protein (Glucksmann *et al.*, 1993). Cps19fC and Cps19fD are homologous to first and last thirds of ExoP, respectively. This may indicate that ExoP has multiple functional domains, which are encoded by separate genes in streptococci. Cps19fD and the C-terminal domain

of ExoP contain the ATP-binding motifs GXGKTT and YIIVD (amino acids 46-51 and 148-152 in Cps19fD) which have been described previously (Guidolin *et al.,* 1994; Kolkman *et al.,* 1997b). *R. meliloti* ExoP-truncated mutants (missing the C-terminal domain) produce less succinoglycan with a reduced molecular weight (Becker *et al.,* 1995). This suggests that Cps19fD may function, in conjunction with Cps19fC, to regulate the export and chain length of type 19F CPS.

1.9.4.5 Cps19fE

The *cps19fE* gene product is relatively hydrophobic and exhibits significant similarity to known glycosyl transferases. The galactosyl transferase encoded by the *S. agalactiae cpsD* gene (Rubens *et al.,* 1993) shows 45.6% identity to the *cps19fE* gene product. Cps19fE also exhibits similarity to RfbP (31.4%) from *S. enterica* serovar typhimurium strain LT2 (Jiang *et al.,* 1991), an und-P Gal-1-phosphate transferase which catalyses the initial step in O-antigen biosynthesis, and the *X. campestris xps2a* (*gumD*) gene product (24.1%), a Glc-1-phosphate transferase which catalyses the first step in xanthan gum synthesis (Ielpi *et al.,* 1993). Examination of the hydrophobicity plot of Cps19fE revealed that the N-terminal portion of the protein has 3 hydrophobic segments which may be potential membrane spanning domains; this suggests that Cps19fE is anchored to the bacterial membrane. Interestingly, although the N-terminal portions of Cps19fE, RfbP, and Xps2a have little amino acid sequence similarity, the hydropathy plots for these regions are very similar. Studies with RfbP (Wang *et al.,* 1996) have demonstrated that the hydrophilic C-terminal region is required for the transferase activity of these proteins. Kolkman *et al* (1996) demonstrated that Cps14E, which is almost identical (95.8%) to Cps19fE, is a glucosyl transferase. Thus Cps19fE is a glucosyl-transferase which adds Glc, the first sugar in the polysaccharide repeat unit, to the lipid carrier.

1.9.4.6 Cps19fF

The *cps19fF* gene product has similarity (31-33% identity) to three other proteins, namely the *B. subtilis tagA* gene product, an *N*-acetyl mannosamine (Man*N*Ac) transferase which is needed for cell wall teichoic acid biosynthesis (Mauel *et al.,* 1991), and the *E. coli* K12 and *S. enterica* serovar typhimurium strain LT2 *rffM* gene product, which is a putative *N*-acetyl-D-mannosaminuronic acid transferase, involved in the synthesis of ECA (Daniels *et al.,* 1992; Lu and Abdelal, 1993). Since the type 19F CPS contains Man*N*Ac, Cps19fF is probably the transferase which catalyses the addition of this sugar in the synthesis of the type 19F polysaccharide.

1.9.4.7 Cps19fG

The *cps19fG* gene (incomplete) encodes a truncated protein which exhibits 38% identity with the N-terminal end of the LicD protein of *Haemophilus influenzae* (Weiser *et al.,* 1989). However, the precise function of LicD is unknown.

1.10 Aims of this Thesis

As mentioned previously (section 1.9), when this study commenced, the only capsule biosynthesis genes from *S. pneumoniae* that had been analysed were a single gene from the type 3 locus (Arrecubieta *et al.,* 1994) and what appeared to be the first six genes of the type 19F locus (Guidolin *et al.,* 1994). Knowledge of the factors determining or regulating capsular polysaccharide biosynthesis is likely to provide new insights into the mechanisms of pathogenesis of pneumococcal disease. Moreover, comparison of the genetic structure of the loci encoding synthesis of immunologically-related capsular polysaccharides may provide information on the mechanism whereby *S. pneumoniae*

acquired the capacity to synthesise 90 distinct capsular serotypes. Accordingly, the work described in this thesis was aimed at complete characterisation of the *cps* loci for all four members of *S. pneumoniae* serogroup 19 (serotypes 19F, 19A, 19B and 19C).

Chapter 2

MATERIALS AND METHODS

2.1 Bacterial strains and cloning vectors

2.1.1 Bacterial strains

The *E. coli* strains used in this study are described in **Table 2.1**. Bacteriophage N4 was obtained from Prof. L.B. Rothman-Denes and Dr. D.R. Kiino (Kiino *et al.*, 1993).

Table 2.1 ***E. coli*** **K12 strains**

Strain	Relevant genotype	Source / Reference
DH1	F⁻, *gyrA96, recA1, relA2, endA1, thi1, hsdR1, supE44,* λ⁻	Hanahan (1983)
DH5	F⁻, *deoR, recA1, endA1, hsdR17*(r_k^-,m_k^-), *supE44, thi1, gyrA96, relA1,* λ⁻	Bethesda Research Laboratories, Gaithersburg, MD
DH5α	F⁻, ϕ*80dlacZΔM15,* Δ(*lacZYA-argF*), *U169, deoR, recA1, endA1, hsdR17*(r_k^-,m_k^-), *supE44, thi1, gyrA96, relA1,* λ⁻	Bethesda Research Laboratories, Gaithersburg, MD
MC4100	F⁻, *araD139,* Δ(*argF-lac*)*U169, rpsL150, rclA1, flbB5301, deoC1, ptsF25, rbsR, thiA1*	*E. coli* Genetic Stock Center, Yale University, CT (Silhavy *et al.*, 1984)
KI8828	MC4100 *nfrC2*	Prof. L.B. Rothman-Denes and Dr. D.R. Kiino (Kiino *et al.*, 1993)
SØ874	F⁻, *lacZ2286, trp-49,* Δ(*sbcB-rfb*)86, *upp-12, relA1, rspL150,* λ⁻	Neuhard and Thomassen (1976)

S. pneumoniae strains and clinical isolates used in this study are described in **Table 2.2**, all other clinical isolates were from the Department of Microbiology, Women's and Children's Hospital, Adelaide, South Australia.

Table 2.2 *S. pneumoniae* strains

Strain	Serotype	Source / Reference
Rx1	rough	Shoemaker and Guild (1974)
Rx1-19F	19F	Morona *et al.* (1994a)
SSZ-19F	19F	Dr. Chi-Jen Lee, Center for Biologics, FDA, Bethesda, Md., USA
1777/39 (19A1)	19A	Dr. Jorgen Henrichsen, Statens Seruminstitüt, Copenhagen, Denmark
19A2	19A	Dr. Chi-Jen Lee, Center for Biologics, FDA, Bethesda, Md., USA
clinical isolates	19A	Mike Gratten, ARI Research and Reference Unit, Centre for Public Health Sciences, Queensland Health, Brisbane, Australia
19B	19B	Dr. Chi-Jen Lee, Center for Biologics, FDA, Bethesda, Md., USA
19C	19C	Dr. Chi-Jen Lee, Center for Biologics, FDA, Bethesda, Md., USA

FDA, Food and Drug Administration; ARI, Acute Respiratory Infections.

2.1.2 Cloning vectors

The bacterial cloning vectors used in this study are listed in **Table 2.3**.

Table 2.3 Bacterial plasmids and cloning vectors

Plasmid/Vector	Antibiotic Resistance	Reference
pBluescript SK+	Amp	Stratagene, La Jolla, CA.
pBluescript KS+	Amp	Stratagene, La Jolla, CA.
pGEM-7Zf(+)	Amp	Promega Corporation, Madison, WI.
pK184	Kan	Jobling and Holmes (1990)
pK194	Kan	Jobling and Holmes (1990)
pVA891	Ery, Cml	Macrina *et al.* (1983)
pGP1-2	Kan	Tabor and Richardson (1985)

Amp, ampicillin; Kan, kanamycin; Ery, erythromycin; Cml, chloramphenicol.

2.1.3 Growth media

E. coli strains were grown in Luria-Bertani (LB) broth with or without 1.5% (w/v) Bacto-agar (Difco Laboratories, Detroit, MI, USA). LB broth was prepared by dissolving 10 g tryptone (Difco), 5 g yeast extract (Difco), and 5 g NaCl in deionised water to a final volume of 1 litre, adjusted to pH 7.5 with NaOH, and sterilised by autoclaving at 121°C for 15 min. LB agar was prepared similarly, but with the addition of 15 g agar before autoclaving. Where appropriate, chloramphenicol (Cml), ampicillin (Amp), kanamycin (Kan), or erythromycin (Ery) was added to the growth medium at concentrations of 25, 50, 25 and 10 µg/ml, respectively.

E. coli strains were preserved as follows: a 1.5 ml aliquot from a fresh overnight culture was mixed with 375 µl 80% (v/v) glycerol and stored at -80°C.

Pneumococci were routinely grown in either Todd-Hewitt broth supplemented with 0.5% (w/v) yeast extract (THY), or on blood agar (BA). THY was prepared by dissolving 36.4 g of Todd-Hewitt broth (Oxoid Limited, Basingstoke, England) and 5 g yeast extract (Difco) in deionised water to a final volume of 1 litre, and sterilised by autoclaving. BA was prepared by dissolving 40 g blood agar base no. 2 (Oxoid) in deionised water to a final volume of 1 litre, sterilised by autoclaving and cooled to 55°C before the addition of 25 ml of defibrinated horse blood (Amadeus International, Adelaide, Australia). Where appropriate, Ery was added to media at a concentration of 0.2 µg/ml. When pneumococci were grown to stationary phase in THY, 1% (w/v) choline (Aldrich Chemical Company, NSW, Australia) was added to the medium.

Pneumococcal strains were grown in serum broth for 4 h at 37°C and then stored at -80°C. Serum broth was prepared by dissolving 10 g Bacto peptone (Difco), 5 g NaCl and 8 g Lab Lemco powder (Oxoid) into 1 litre of deionised water, and the pH was adjusted to 7.4. After sterilisation by autoclaving, 10% (v/v) donor horse serum (CSL, Victoria,

Australia) was added, and the broth was filter sterilised.

2.2 Chemicals and reagents

2.2.1 General chemicals

Most chemicals used were AnalaR grade and were purchased from Ajax Chemicals, (NSW, Australia). Tris was purchased from Progen Industries (Queensland, Australia). Maleic acid, SDS, DTT and N,N,N',N'-tetramethyl-ethylene-diamine (TEMED) were purchased from Sigma Chemical Company (St. Louis, MD., USA). Acrylamide and ammonium persulphate were purchased from Bio-Rad Laboratories (Hercules, CA, USA). The deoxyribonucleoside triphosphates (dATP, dCTP, dGTP and dTTP), X-gal, IPTG, and herring sperm DNA were purchased from Boehringer Mannheim (Mannheim, Germany). DOC was purchased from BDH Biochemicals (Poole, England).

2.2.2 Antibiotics

Amp was purchased from CSL (Victoria, Australia); Kan sulphate was purchased from Progen Industries; Cml and Ery were purchased from Boehringer Mannheim; rifampicin (Rif), tetracycline (Tet) and streptomycin sulphate (Str) were purchased from Sigma.

2.2.3 Enzymes

RNase A, lysozyme, pronase and proteinase K were purchased from Boehringer-Mannheim. All restriction endonucleases were purchased from either Boehringer

Mannheim, or Progen and used according to the manufacturers' recommendations. Other DNA modifying enzymes, T4 DNA ligase, Klenow fragment of DNA polymerase I and Taq polymerase were purchased from Boehringer Mannheim. Sequencing kits using either dye-labelled primer or dye-labelled terminators were purchased from Applied Biosystems.

2.2.4 Oligodeoxynucleotides

The oligodeoxynucleotides used are listed in **Table 2.4** and were purchased from Life Technologies (Gaithersburg, MD, USA).

Table 2.4 Oligodeoxynucleotides

Name	Sequence (5'-3')	Homologous to:
DEXB	ataaagcTTCCATGGGATGCTTTCTGT	*dexB* (35-16 nt from 3' end)
CPSA2	ataggatCCCTAGCAAGGCAACTAGTA	<*cps19fA* (nt 335-354)
CPS5'	TGATGTTCAAGGTATAGGTGTTAATCA	before *cps19fA* (nt 146-169)
CPSX1	atactcgAGATGGAGCACCAAATCATTC	*cps19fG* (nt 6,245-6,265)
J5	atgggatCCACTACACTTGATGGGGGT	*cps19fF* (nt 5,224-5,243)
J7	TTATTATATTTTgcATgCCTTTAATGGC	*cps19fI* (nt 7,772-7,799)
J8	TAAAACTAGACaAGCTtTAGCAAAA	<*cps19fE* (nt 4,584-4,608)
J9	TTCCTTCTaGATTTGTAAAGATATT	*cps19fJ* (nt 9,502-9,526)
J10	CTAAGAACAAGcTTGTATATTTCCT	<*cps19fI* (nt 7,919-7,943)
J11	GATggATCCCATCAATCCAACCCAAGT	<*cps19fJ* (nt 9,108-9,129)
J12	CTCGCGCTGCAgCAAAACAAC	*cps19fL* (nt 11,597-11,614)
J14	CAACTTTgAAtTCACCAGTTATTTTTTG	*cps19fH* (nt 7,515-7,542)
J15	GACTCAgGTAcCAAATTAAAAGAAGAA	*cps19fI* (nt 8,881-8,907)
J16	AAATGGATccAAAGGTAGAAATGTAATC	<*cps19fJ-K* (nt 10,379-10,406)
J17	CACGGgGATcCTACTACGACATAT	*cps19fK* (nt 10,697-10,720)
J18	AATGATCTGcATgCGTTCTGTATCTC	<*cps19fK* (nt 11,166-11,191)
J19	AATAGTCAGTCgaCCTACTATTGGAC	*cps19fO* (nt 15,340-15,364)
J20	AAGTAgGAtCCTGATATTTCCCAGC	<*cps19fG* (nt 6,096-6,120)

(continued overleaf)

Table 2.4 Oligodeoxynucleotides (continued)

Name	Sequence (5'-3')	Homologous to:
J21	TGCTGGAAgcTTGAAATTGGTTGGG	*cps19fK* (nt 11,341-11,365)
J22	AATTGaaTTCTTTTATAGATTTAACACAAG	<*cps19fH* (nt 6,743-6,772)
J24	TTGAGAAAGcTTTTGAAGAGGTGAAA	*cps19fO* (nt 14,260-14,285)
J25	GTACGgaTCCAAGTCGGACGACC	<*cps19fO* (nt 14,682-14,704)
J26	TAGTGAGAAtTcTCTATCCTATCTTCTA	*aliA* (nt 16,127-16,154)
J27	GGCTgGatCCAAACCATCTGGACTT	<*cps19fL* (nt 11,787-11,811)
J28	GAAAAAAaGCTTCCACTTGGATTTCC	*cps19fM* (nt 12,541-12,566)
J29	GCTTGGATCcGCGTAGTTCACAAA	<*cps19fM* (nt 12,880-12,903)
J30	TGACTTAaGCTtGAAACCGCGCGA	*cps19fN* (nt 13,136-13,159)
J31	GCCTgGAtCCCAGCTTCAAAGTTG	<*cps19fN* (nt 13,946-13,946)
J32	CGTCTTTTacTAGTTGTTGGATATCAAT	<*aliA* (nt 17,853-17,880)
J33	ACAAGAaTCgATTACGTATTTAGTTGGT	*aliA* (nt 17,339-17,366)
J34	CTATTTGGATCCGTAGCTTTGGCA	<*aliA* (nt 17,057-17,080)
J36	CAATAATGTCACGCCCGCAAGGGCAAGT	<*aliA* (nt 16,463-16,490)
J37	TGGTGGAAAGCttAGAAAGAAGCTG	*cps19fN* (nt 14,002-14,026)
J38	TCTAGCGGaTCcCAGTTTACCAAGC	<IG3' region (nt 15,568-15,592)
J39	TAGTTCATGTAGTTGCAAGTGACATGCACAA	*cps19fB* (nt 2,190-2,220)
J42	AAGTCaTcGATGAAACTATTTCTTGTG	*cps19bH* (nt 1,499-1,525)
J43	TATATGATTGTATcGATTATCATGTGGC	*cps19bP* (nt 2,693-2,720)
J44	TAATCAGCTGATcgATCAGCTCCGCTC	<*cps19bP* (nt 3,395-3,421)
J45	AATGGAtcCATTTTTATCGCTTCTGGAC	<*cps19bK* (nt 9,273-9,302)
J47	CCTATACCATTTGaATTCAATATATTCTAG	<*cps19bJ* (nt 7,830-7,857)
J49	AATGTGgAtCCTACATATGTATTAACAG	*cps19bR* (nt 6,886-6,913)
J70	GATGGTggaTCCTGTTTTAGATTTATTTGG	*cps19aK* (nt 12,114-12,143)
J72	TATCGTCAtCGATAAATTCTTCACCG	< *cps19aL* (nt 13,515-13,540)
J87	CTCATACTaGTCCACTGTTGGTG	<*cps19aE* (nt 6,458-6,480)
J88	CTCAGGatCCCGCCTGTACCC	*cps19aL* (nt 13,266-13,286)
J92	TTGGATAAGCTtGAAAAAAATCGGATT	*cps19aB* (nt 3,441-3,467)
J93	TACTGAAaGCTTTGCTAGTTTTTCAC	< *cps19aE* (nt 5,330-5,355)
J94	AACACGATAGAAATCGATGTATTTC	*cps19aC* (nt 3,932-3,956)
J95	TTTTGgAtCCTGCACGCGCAAAAG	< *cps19aD* (nt 4,797-4,820)

< denotes primer is complementary to the sequence of the indicated gene. Lower case letters indicate base substitutions to insert restriction sites. Cleavage sites for restriction endonucleases *Eco*RI (GAATTC), *Bam*HI (GGATCC), *Hin*dIII (AAGCTT), *Cla*I (ATCGAT), *Xba*I (TCTAGA), *Pst*I (CTGCAG), *Sph*I (GCATGC), *Sal*I (GTCGAC), *Kpn*I (GGTACC), *Spe*I (ACTAGT) and *Xho*I (CTCGAG) are underlined. The sequences to which the primers are homologous are available as follows: *dexB-cps19fA* and *cps19fA-G* are available under GenBank accession nos U04047 and U09239, respectively; *cps19fG-O, cps19b* and *cps19a* are as described in Appendices I, II and III, respectively, nt denotes nucleotide position.

2.3 Serotyping of pneumococcal strains

Production of capsule by pneumococci was assessed by quellung reaction, using diagnostic pneumococcal typing sera produced by Statens Seruminstitüt, Copenhagen, Denmark. Serum broth was inoculated with cells from an overnight BA plate and incubated at 37°C for 4 h. Then 10 µl serum broth culture was mixed with 10 µl of the typing sera and microscopic examination for capsular swelling confirmed the production of type-specific capsule. This was performed by staff from the Dept. of Microbiology and Infectious Diseases, Women's and Children's Hospital, Adelaide.

Pneumococci belonging to serogroup 19 were serotyped (19F, 19A, 19B or 19C) by quellung reaction, using factor specific antisera obtained from Statens Seruminstitüt, Copenhagen, Denmark, and was performed by Mr. M. Gratten and Ms. Denise Murphy, Public Health Bacteriology Laboratory, Centre for Public Health Sciences, Queensland Health.

2.4 Bacterial transformation

2.4.1 Preparation of competent *E. coli* strains

2.4.1.1 $RbCl_2$ method

E. coli K12 DH5α was made competent as follows. A 100 ml LB broth was inoculated with 1 ml of an overnight culture of DH5α and grown to very early log phase (A_{600} 0.04). The cells were then centrifuged in a Hettich Universal/K2S centrifuge at 4,000 x g for 15 min at 4°C. The pelleted cells were resuspended in 40 ml of TfbI (30 mM

KOAc, 100 mM $RbCl_2$, 10 mM $CaCl_2$, 50 mM $MnCl_2$, adjusted to pH 5.8 with 0.2 M acetic acid and filter sterilised) and placed on ice for 5 min. The cells were again centrifuged at 4,000 × g for 15 min at 4°C. The cells were then resuspended in 2 ml of TfbII (10 mM MOPS, 75 mM $CaCl_2$, 10 mM $RbCl_2$, 15% (v/v) glycerol, adjusted to pH 6.5 with KOH and filter sterilised) and placed on ice for 15 min. The cells were dispensed into sterile 1.5 ml microcentrifuge tubes in 50 µl volumes, quickly frozen on ethanol/dry ice for 5 min, and transferred to -70°C. The bacterial cells remained viable with no apparent loss of competence for at least 6 months. The cells were thawed at room temperature before being placed on ice and transformed as described in section 2.4.2.

2.4.1.2 Electroporation method

DH5(pGP1.2) was made competent for transformation by electroporation, as follows. Briefly, 100 ml fresh LB broth was inoculated with 10 ml of an overnight broth culture. The culture was grown at 30°C with vigorous shaking until a density of A_{600} = 0.5 to 0.8 was achieved. Before harvesting, the bacterial cells were chilled on ice for 15 min and then centrifuged at 4,000 × g for 15 min at 4°C. The supernatant was drained away as much as possible and the pellet was resuspended in 100 ml ice-cold water and centrifuged as before. The last step was repeated, and the cells were resuspended in 50 ml ice cold water. The resultant pellet was then resuspended in 20 ml of ice-cold 10% (v/v) glycerol, centrifuged as before and resuspended in a final volume of 2 ml of ice-cold 10% (v/v) glycerol. Cells were quickly frozen in ethanol/dry ice in 100 µl aliquots in 1.5 ml microfuge tubes, and stored at -70°C. The cells were stable for about six months under these conditions.

2.4.1.3 $CaCl_2$ method

All other *E. coli* strains were made competent for transformation with plasmid DNA according to the method described by Brown *et al.* (1979). An overnight culture (in

LB broth) was diluted 1:20 into fresh LB broth and incubated with aeration for 2 h until the culture reached an A_{600} of 0.04 (very early log phase). The bacterial cells were pelleted by centrifugation at 4,000 × g for 15 min at 4°C and resuspended in 0.5 volumes of cold 100 mM $MgCl_2$, centrifuged again and resuspended in a 0.1 volumes of cold 100 mM $CaCl_2$. The bacterial cells were allowed to stand for 1 to 2 h on ice before use.

2.4.2 Transformation of competent *E. coli* strains

Competent cells (50 µl) were mixed with 15 µl DNA and kept on ice for 30 min. The cell/DNA mixture was then heat-shocked at 42°C for 3 min, and then again placed on ice for a further 10 min. The bacterial cells were incubated at 37°C for 1 h after the addition of 0.5 ml LB. The bacterial cell suspension was then plated onto selection plates.

2.4.3 Electroporation of *E. coli* strains

Transformation of electroporation competent cells was performed as described in the Electrocell Manipulator® 600 (ECM® 600) Operation Manual (BTX, San Diego, CA., USA). Frozen bacteria were gently thawed at room temperature and placed on ice before the addition of 1 to 2 µg of DNA (dissolved in Hi-pure water), the cells were mixed thoroughly and allowed to stand for 1 min. The cell suspension was transferred to a chilled sterile BTX disposable electroporation cuvette (P/N 620, 2 mm gap) and this was then placed in the safety chamber of the ECM® 600 which was set at 2.50 kV/resistance high voltage and 129 Ω resistance. The cell suspension was then pulsed for 5-6 msec at a charging voltage of 2.45 kV. After electroporation the cell suspension was mixed with 960 µl LB, incubated at 30°C for 1 h without shaking, and then plated onto selective medium.

2.4.4 Transformation of pneumococcal strains

S. pneumoniae strains were grown at 37°C, in 10 ml THY broth, to a density of $3x10^8$ cells/ml (A_{600} of 0.27). The culture was diluted 100-fold into competence medium containing 10% (v/v) glycerol and stored at -70°C in 500 µl aliquots. After thawing a 500 µl aliquot of cells, 500 µl of competence factor was added and incubated at 37°C for 20 min. The cells were then mixed with 100 µl of DNA prepared as described in section 2.5.2 or 2.5.3 and incubated at 37°C for 2 h. Transformants were selected by plating the cells on blood agar plates containing 0.2 µg/ml erythromycin and incubating at 37°C for up to 48 h.

2.4.5 Preparation of competence factor for pneumococcal transformation

Pneumococcal strains produce one or more competence-inducing pheromones which activate competence and the uptake of exogenous DNA (Pozzi *et al.*, 1996). Production of these pheromones occurs spontaneously at a particular point in the growth cycle and when harvested can be used to induce competence in other pneumococcal strains. The highly transformable *S. pneumoniae* strain Rx1 was used to prepare competence factor, as described previously by Yother *et al.* (1986).

A 10 ml culture of Rx1 in THY was grown, at 37°C, to a density of $3x10^8$ cells (A_{600} of 0.27). A 5 ml aliquot, with 10% (v/v) glycerol, was stored at -70°C, for 16 h. The 5 ml aliquot was thawed at 37°C and diluted into 500 ml competence medium (THY; 0.2% (w/v) bovine serum albumin; 0.01% $CaCl_2$; adjust to pH 7.8). The 500 ml culture was incubated at 37°C for 30 min and then 18 ml samples were taken at 10 min intervals. Each 18 ml sample was mixed with 2.6 ml 80% (v/v) glycerol and stored at -70°C immediately; a 100 µl aliquot of each time point was stored separately in a 1.5 ml microfuge tube for further analysis.

To assess competence, the 100 µl samples were incubated at 37°C for 10 min, then 10 µl chromosomal DNA (from an Ery resistant pneumococcus) was added and the samples were incubated for a further 30 min at 37°C. After incubation, serial 10-fold dilution to 10^{-6} were carried out in a microtitre tray and plated onto BA plates containing 0.2 µg/ml Ery, and incubated at 37°C for 16 h. The time points with the highest number of Ery-resistant transformants and thus the highest transformation frequencies were prepared for use as competence factor. The frozen aliquots were thawed and the bacteria pelleted by centrifugation at 4,000 × g for 15 min at 4°C. The supernatant (or competence factor) was carefully removed and filter sterilised through a 0.45 µm filter, dispensed into 500 µl aliquots and stored at -70°C.

2.5 DNA extraction procedures

2.5.1 Plasmid isolation

Plasmid isolation procedures were based on the alkaline lysis method (Morelle, 1989). A 1.5 ml aliquot of an *E. coli* overnight culture was transferred to a microcentrifuge tube, pelleted by centrifugation at 13,000 rpm for 2 min in a microfuge (Biofuge 13, Heraeus Instruments, Germany), and resuspended in 100 µl of a solution of 50 mM Glc, 25 mM Tris-HCl (pH 8.0), 10 mM EDTA, and lysozyme at a final concentration of 4 µg per ml. After 5 min incubation at room temperature, 200 µl of a solution of 1% (w/v) SDS and 200 mM NaOH was added. The tubes were then mixed by inverting several times and incubated on ice for 5 min. After the addition of 400 µl 7.5 M ammonium acetate, pH 4.8, the tubes were again mixed by inversion, incubated for 10 min on ice, and cellular debris

was then removed by centrifugation at 13,000 rpm for 10 min. Supernatants were then transferred to a clean microfuge tube and the plasmid DNA precipitated by the addition of 500 µl of ice-cold isopropanol. The tubes were then mixed by inverting several times and left standing at room temperature for 15 min. The plasmid DNA was pelleted by centrifugation at 13,000 rpm for 15 min, the pellet was washed with 70% (v/v) ethanol, and dried under vacuum. The pellet was resuspended in 20-50 µl of sterile water (Baxter, NSW, Australia) and stored at 4°C.

Plasmid DNA to be used in sequencing reactions (Applied Biosystems) was further purified as follows. The plasmid DNA prepared as above was mixed with 1 µl RNase A (10 mg/ml in H_2O) and incubated at 37°C for 15-30 min. The DNA was then mixed with an equal volume of 7.5 M ammonium acetate and 2 volumes of isopropanol. The sample was mixed by vortex, left standing at room temperature for 15 min, and then centrifuged for 15 min at 13,000 rpm. The supernatant was discarded, the DNA pellet was washed with 75% ethanol, 25% 50 mM Na acetate and microfuged for 10 min. The pellet was then washed in 100% ethanol and microfuged for 5 min. The pellet was dried under vacuum and resuspended in the original volume of TE (10 mM Tris, 1 mM EDTA, pH 8.0). Alternatively, they were prepared using the plasmid mini-kit (Qiagen), following the procedure outlined in the Qiagen plasmid handbook.

2.5.2 Plasmid isolation for pneumococcal transformation

A 100 ml overnight culture was centrifuged at 4,000 × g for 15 min at 4°C and the pellet was resuspended in 4 ml 50 mM Glc, 25 mM Tris (pH 8.0), 10 mM EDTA (pH 8.0), 2 mg/ml lysozyme and incubated on ice for 10 min, then 8 ml 0.2 M NaOH, 1% (w/v) SDS was added, mixed by inversion and placed on ice for 10 min. Next, 6 ml 1.5 M Na acetate (pH 4.8) was added, mixed by inversion and placed on ice for a further 10 min. The debris

was pelleted by centrifugation at 4,000 × g for 15 min at 4°C and the supernatant was carefully decanted into a clean tube. The supernatant was mixed with 6 ml of ice-cold isopropanol and placed on ice for 15 min. The DNA was pelleted by at 4,000 × g for 20 min at 4°C, resuspended in 2 ml of 0.1 M Na acetate, 0.05 M Tris (pH 8.0) and transferred to microfuge tubes (500 µl per tube) and re-precipitated by the addition of 1 ml 100% alcohol. The tubes were placed on ice for 15 min and the DNA was pelleted by centrifugation at 4,000 × g for 15 min at 4°C. The DNA was resuspended in a total volume of 400 µl TE and used to transform *S. pneumoniae*.

2.5.3 Preparation of pneumococcal chromosomal DNA

S. pneumoniae chromosomal DNA was prepared using the Wizard® Genomic DNA Purification Kit (Promega Corporation). The protocol was modified to optimise recovery of chromosomal DNA from pneumococci as follows.

The growth from a BA plate incubated at 37°C overnight was resuspended in 10 ml THY containing 1% (w/v) choline and grown at 37°C for 3-4 h. The culture was pelleted by centrifugation at 4,000 × g for 20 min at 4°C, resuspended in 200 µl 50 mM EDTA (pH 8.0) plus 0.1% (w/v) DOC and the suspension was divided between two microfuge tubes before incubation at 37°C for 10 min. After 10 min, the suspension was translucent, indicating that most of the bacteria had lysed, then 200 µl 50 mM EDTA (pH 8.0) and 300 µl Nuclei Lysis Solution (from the kit) was added and gently mixed by pipette. The lysates were then incubated at 80°C for 10 min to ensure all the bacteria had lysed. The lysates were cooled to room temperature, 1.5 µl of a 10 mg/ml RNase A solution was added, mixed by inversion and incubated at 37°C for 15-60 min. The lysates were then cooled to room temperature and 100 µl Protein Precipitation Solution (from the kit) was added and

mixed by inversion. The lysates were placed on ice for 5 min and then centrifuged at 13,000 rpm for 10 min. The supernatant was transferred to a clean microfuge tube and 600 µl isopropanol was added and mixed by inversion. The precipitated chromosomal DNA appeared as white fluffy strands and was pelleted by centrifugation at 13,000 rpm for 10 min. The supernatant was carefully removed and the pellet was washed with 600 µl 70% (v/v) ethanol. The DNA was again pelleted by centrifugation at 13,000 rpm for 5 min and the supernatant was drained from the tube, removing as much as possible. The pellet was resuspended in 50-100 µl TE and allowed to resuspend at 4°C overnight.

2.6 Analysis and manipulation of DNA

2.6.1 Restriction endonuclease digestion of DNA

DNA samples were digested using the restriction enzyme buffers recommended by the manufacturers. About 100 to 500 ng of plasmid DNA, or purified restriction fragments, were incubated with 2 units of each restriction enzyme in a final volume of 10 µl at 37°C for 2 h, with the exception of *Sma*I digests, which were carried out at 30°C. The reactions were terminated by heating at 65°C for 10 min. Prior to electrophoresis on an agarose gel, a one-tenth volume of tracking dye consisting of 25% (w/v) Ficoll (type 400), 0.25% (w/v), bromophenol blue, 0.25% (w/v) xylene cyanol, was added. For the digestion of chromosomal DNA, 4 µg of DNA was incubated with 4 units of each enzyme in a final volume of 20 µl at 37°C for 16 h.

2.6.2 Calculation of sizes of restriction fragments

The sizes of restriction enzyme fragments were calculated by comparing their

relative mobility with that of *Eco*RI-digested *B. subtilis* bacteriophage SPP1 DNA, and/or with that of *Hin*dIII-digested λ phage DNA obtained from Bresatec (Adelaide, South Australia). The calculated molecular sizes of the *Eco*RI-digested SPP1 DNA fragments were: 8.51 kb, 7.35 kb, 6.11 kb, 4.84 kb, 3.59 kb, 2.81 kb, 1.95 kb, 1.86 kb, 1.51 kb, 1.39 kb, 1.16 kb, 0.98 kb, 0.72 kb, 0.48 kb, and 0.36 kb (Ratcliffe et al., 1979). The calculated molecular sizes of the *Hin*dIII-digested phage λ DNA fragments were: 23.13 kb, 9.42 kb, 6.56 kb, 4.36 kb, 2.32 kb, 2.03 kb, 0.56 kb and 0.13 kb (Sanger *et al.*, 1983).

2.6.3 Analytical and preparative separation of restriction fragments

Electrophoresis of digested DNA was carried out at room temperature on horizontal, 0.8% -1.5% (w/v) agarose (DNA grade agarose, Progen Industries) gels with a Tris-borate-EDTA (TBE; 67 mM Tris, 22 mM boric acid and 2 mM EDTA at pH 8.0) buffer system, containing 0.5 μg/ml ethidium bromide, as described by Maniatis *et al.* (1982). DNA bands were visualised by trans-illumination with UV light and photographed using either a Polaroid MP4 camera with type 667 positive film or Mitsubishi thermal paper (K65HM) on a Mitsubishi video copy processor fitted to a Tracktel GDS-2 camera.

2.6.4 Isolation of restriction fragments from agarose gels

DNA restriction fragments were excised from agarose gels and purified as detailed in the Qiaex handbook using the Qiaex DNA Gel Extraction Kit (Qiagen).

2.6.5 *In vitro* cloning

The DNA fragment to be subcloned (about 200 ng) was cleaved in either single or double restriction enzyme digests. This was combined with 20 ng of similarly cleaved

vector DNA, then ligated with 1 unit of T4 DNA ligase in a volume of 15 µl in a buffer containing a final concentration of 20 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 10 mM DTT, and 600 nM ATP at 4°C for 16 h. The ligated DNA was then used directly for transformation of *E. coli* strains. Wherever possible, transformants were screened for insertional inactivation of any interrupted genetic marker prior to plasmid DNA isolation. If the interrupted marker conferred antibiotic resistance, the transformants were screened for sensitivity to the antibiotic. Where the DNA insert was cloned into a multicloning site near the start of the β-galactosidase gene, transformants were selected using indicator plates containing IPTG and X-Gal. The transformants containing only vector DNA formed blue colonies whereas those with additional DNA inserted in the vector formed white colonies in these plates.

2.7 Southern hybridisation

2.7.1 Labelling of DNA probes

DNA probes were labelled with digoxigenin (DIG) using the DNA Labelling Kit purchased from Boehringer Mannheim. The DNA to be labelled was added to a microfuge tube and denatured by heating to 95°C for 10 min and then rapidly cooled on ice for 1 min. Then 2 µl hexanucleotide mixture, 2 µl dNTP mixture (containing DIG-11-dUTP) and 1 µl Klenow enzyme was added to the DNA. Water was added to give a final volume of 20 µl. The mix was incubated at 37°C overnight. The labelled DNA mixture was heated at 95°C for 10 min and cooled rapidly on ice before being added to the hybridisation mix.

2.7.2 DNA hybridisation by Southern blotting

2.7.2.1 Transfer of DNA to nylon membrane

Restriction fragments from single and/or double digests of DNA were separated electrophoretically on a 1% (w/v) agarose gel. DIG labelled lambda DNA, restricted with *Hin*dIII, was used as a DNA molecular size marker. Unidirectional transfer of DNA from agarose gels to Hybond-N^+ nylon membrane (Amersham, England) was performed as described by Southern (1975) and modified by Maniatis *et al.* (1982). The DNA was fixed to the nylon membrane by placing the membrane on a piece of Whatman filter paper soaked in 0.4 M NaOH for 5-30 min. The membrane was then washed in 5x SSC (150 mM NaCl, 15 mM Na_3 citrate, pH 7.0 for 1x SSC), and then in deionised water before prehybridisation.

2.7.2.2 High stringency hybridisation

The membrane was prehybridised for 1-6 h at 42°C in a prehybridisation solution consisting of 50% (v/v) formamide, 5x SSC (pH 7.0), 2% (w/v) blocking reagent (Boehringer Mannheim) and 0.01% (w/v) N-lauroylsarcosine. The membrane was then hybridised overnight at 42°C in hybridisation solution consisting of 50% (v/v) formamide, 5x SSC (pH 7.0), 5x Denhardt's reagent, 1% (w/v) SDS, 100 µg/ml heparin and 100 µg/ml of single stranded herring sperm DNA (Sigma) (Maniatis *et al.*, 1982) to which the denatured DIG labelled DNA probe had been added. After hybridisation, high stringency washing of the membrane was carried out as follows. The membrane was washed twice for 5 min with 2x SSC at 65°C, and then twice for 5 min with a solution of 0.2x SSC and 0.1% (w/v) SDS at 65°C. These conditions were able to detect DNA sequences with ≥90% homology to the DIG labelled DNA probe, based on the actual DNA sequences detected.

2.7.2.3 Low stringency hybridisation

Low stringency hybridisation was essentially the same as high stringency hybridisation, except it was performed at room temperature. After hybridisation, the membrane was washed twice for 15 min with 2× SSC at 50°C. These conditions were able to detect DNA sequences with ≥70% homology to the DIG labelled DNA probe, based on the actual DNA sequences detected.

2.7.2.4 Immunological detection

DIG labelled probe which annealed to the DNA on the membrane was detected as follows. After stringency washing of the membranes, they were washed briefly in DIG Buffer 1 (100 mM maleic acid, 150 mM NaCl, pH 7.5) and then incubated for 30 min at room temperature in 100 ml of DIG Buffer 2 (1% blocking reagent in DIG Buffer 1). The membrane was then incubated with 6 µl anti-DIG Fab fragment, alkaline phosphatase (AP) conjugate in 30 ml DIG Buffer 2 for 1 h. The membrane was then washed twice in 100 ml DIG Buffer 1 and then briefly in DIG Buffer 3 (100 mM Tris, 100 mM NaCl, 50 mM $MgCl_2$, pH 9.5). The DIG label was detected by incubating the membrane in the dark with 10 ml DIG Buffer 3 containing 45 µl 4-nitro blue tetrazolium (NBT; 75 mg/ml in 70% (v/v) dimethylformamide) and 35 µl 5-bromo-4-chloro-3-indolylphosphate, toluidinium salt (X-phosphate; 50 mg/ml in 100% dimethylformamide).

2.8 DNA sequencing and analysis

2.8.1 Nested deletions

In order to obtain sequence of both strands of various clones containing

pneumococcal DNA, a series of nested deletion derivatives were constructed by the method of Henikoff (1984) using an `Erase-a-base' kit (Promega, Madison, WI) according to the manufacturer's instructions.

Approximately 5 µg of DNA was digested, in a total volume of 20 µl, with two restriction enzymes. One which generated exonuclease I resistant 3' ends (such as *Bst*XI, *Kpn*I and *Sph*I) and one which generated exonuclease I sensitive ends. This ensured that exonuclease digestion would proceed in only one direction. An aliquot of 1 µl was analysed by agarose gel electrophoresis to ensure that digestion was complete and that the plasmid had been linearised.

Prior to digestion with exonuclease III (Exo III), the restricted DNA was diluted to 60 µl in Exo III 1x buffer (66 mM Tris-HCl (pH 8.0), 0.66 mM $MgCl_2$) and incubated at 37°C. After the addition of 400 units of Exo III, 2.5 µl aliquots were removed at 30 s intervals and added to 7.5 µl of S1 nuclease mix (40.5 mM potassium acetate, pH 4.6, 338 mM NaCl, 1.35 mM $ZnSO_4$, 6.75% (v/v) glycerol containing 2.25 units of S1 nuclease) and placed on ice. After all the samples had been taken, the tubes were incubated at room temperature for 30 min. After the addition of 1 µl of S1 stop buffer (300 mM Tris, 50 mM EDTA) the samples were incubated at 65°C for 10 min to inactivate the S1 nuclease. A 2 µl aliquot from each of the time points was analysed by agarose gel electrophoresis to determine the extent of Exo III digestion. The samples which had been appropriately digested were processed further. The samples were incubated at 37°C for 1 min prior to the addition of 1 µl of Klenow mix (30 µl of 20 mM Tris (pH 8.0), 100 mM $MgCl_2$ with 5 units of Klenow DNA polymerase). The samples were incubated at 37°C for 3 min before the addition of 1 µl of dNTP mix (0.125 mM of each of dATP, dCTP, dGTP and dTTP) followed by a further 5 min incubation at 37°C. The samples were then transferred to room

temperature and mixed with 40 µl of ligase mix (50 mM Tris, pH 7.6, 10 mM $MgCl_2$, 1 mM ATP, 5% (w/v) PEG 8000, 1 mM DTT, with 5 units/ml T4 DNA ligase). After 1 h the samples were placed at 4°C overnight.

The samples were then transformed (as described in section 2.4.2) into competent *E. coli* DH5α, prepared using the $RbCl_2$ method (section 2.4.1.1), and plated onto selection plates. Plasmid DNA from the transformants was characterised by restriction enzyme analysis. Plasmids which contained appropriate deletions were subjected to DNA sequencing.

2.8.2 DNA Sequencing using dye-labelled primers

Sequencing reactions were carried out using 1 µg of double stranded plasmid DNA with the Applied Biosystems Prism dye-primer cycle sequencing ready reaction kit as detailed below.

Reagent	A	C	G	T
Ready Reaction premix	4	4	8	8
DNA template (µl)	1	1	2	2
Total volume (µl)	5	5	10	10

Samples were placed in a Hybaid® Touchdown® Thermal Cycler. The program used consisted of: rapid thermal ramp to 95°C for 30 sec, rapid thermal ramp to 55°C for 30 sec, rapid thermal ramp to 70°C for 60 sec and was repeated for 15 cycles.

After the cycles were completed, the four separate reactions for each DNA sample were mixed together with 80 µl of 95% (v/v) ethanol, mixed thoroughly, and kept on ice for 15 min to precipitate the DNA. The DNA was pelleted at 15,000 rpm for 30 min in a

microcentrifuge, washed in 250 µl of 70% (v/v) ethanol, centrifuged for another 10 min, dried *in vacuo* for 3 min and stored at -20°C.

2.8.3 Sequencing with dye-labelled terminators

Double-stranded DNA template (1 µg in a volume of up to 10 µl), 8 µl of Applied Biosystems Prism dye-terminator premix and 3.2 pM of primer were added and the volume adjusted to 20 µl with deionised H_2O. Sequencing reactions were carried out using 1 µg of double-stranded plasmid DNA in a Hybaid® Touchdown® Thermal Cycler as follows; 96°C for 30 sec, 50°C for 30 sec, 60°C for 4 min repeated for 25 cycles.

After the cycles were completed, each reaction mixture was precipitated with 50 µl of 95% (v/v) ethanol and 2 µl of 3 M sodium acetate (pH 4.8) and placed on ice for 10 min. The DNA was pelleted at 15,000 rpm for 30 min in a microfuge at 4°C, and the pellet was washed in 250 µl of ice-cold 70% (v/v) ethanol. Samples were dried *in vacuo* and stored at -20°C.

2.8.4 DNA sequencing

The sequencing reactions were then sent to the DNA sequencing service provided by the Molecular Pathology Unit, IMVS, Adelaide, Australia. DNA sequence data were obtained using an Applied Biosystems model 373A automated DNA sequencer.

2.8.5 DNA sequence analysis

The sequence was analysed using DNASIS and PROSIS Version 7.0 software (Hitachi Software Engineering, San Bruno, CA.). The programs BLASTX (Altschul *et al.*, 1990) and gapped BLAST (Altschul *et al.*, 1997) was used to translate DNA sequences and

conduct homology searches of the protein databases available at the National Center for Biotechnology Information, Bethesda, Md. Amino acid sequence alignments were performed with the program CLUSTAL (Higgins and Sharp, 1988) and CLUSTAL W (Thompson *et al.*, 1994). The program PROFILEGRAPH (Hofmann and Stöffel, 1989) was used to align hydropathy plots generated according to the method of Kyte and Doolittle (1982). Phylogenetic analysis was carried out using the program MEGA (Kumar *et al.*, 1994).

2.9 PCR Amplification

2.9.1 PCR amplification using Taq polymerase

PCR amplification was performed in a Hybaid® Touchdown® Thermal Cycler and the 50 µl reaction volume contained PCR buffer (1.5 mM $MgCl_2$, 10 mM Tris (pH 8.4), 50 mM KCl, and 100 µg per ml of gelatine), 1.5 U of Taq polymerase, 1 µM of each primer, 100 ng of DNA template, and 200 mM of each of the four dNTPs. The program consisted of 25 cycles, and each cycle consisted of: denaturation at 95°C for 30 sec, annealing at 50°C for 30 sec, and elongation at 72°C for 1-4 min, 1 min for each kb of expected product. A 5 µl sample of the PCR product was analysed by agarose gel electrophoresis.

2.9.2 Long range PCR

The Expand™ Long Template PCR System (Boehringer Mannheim, Germany) was used for long range PCR (LR-PCR), according to the manufacturer's instructions, in a Hybaid® Touchdown® Thermal Cycler. The LR-PCR program used consisted of 25 cycles,

after an initial incubation at 92°C for 2 min to allow complete denaturation of the DNA template. The first 10 cycles consisted of denaturation at 92°C for 30 sec, annealing initially at 55°C for 1 min, decrementing by 1°C per cycle to 45°C at cycle 10, and elongation at 68°C for 5-20 min. Cycles 11-25 consisted of denaturation at 92°C for 30 sec, annealing at 55°C for 1 min at cycle 11, decrementing to 45°C at cycle 25, and elongation at 68°C for 5-20 min at cycle 11, increasing by 1 min per cycle to cycle 25. The length of the extension time varied with the expected size of the PCR product, approximately 1 min for every kb to be amplified, up to a maximum of 20 min.

2.9.3 Inverse PCR

The method used for inverse PCR (InPCR) was that described by Ochman *et al.* (1988). Briefly, chromosomal DNA (5 µg) was digested to completion with an appropriate restriction enzyme. The cleaved DNA was then self-religated at 16°C for 16 h at a concentration of 10 µg/ml. Amplification was performed using 1 µg of ligated DNA and 50 pM of each appropriate primer using either Taq polymerase or LR-PCR as described above.

2.9.4 Purification of PCR products

PCR products were purified for further analysis using a PCR Clean Up Kit (QIAGEN), according to the manufacturer's instructions. If the PCR product required concentration, several individual PCR reactions were pooled and purified and eluted into a final volume of 50 µl.

2.10 Insertion-duplication mutagenesis

This method was used to mutagenise pneumococci by insertion of pVA891 (**Table 2.3**) into pneumococcal ORFs. This required the cloning an appropriate internal fragment of the pneumococcal ORF into pVA891, which encodes Cml and Ery resistance, but cannot replicate in *S. pneumoniae* (Macrina *et al.*, 1983), and transforming this plasmid construct into Rx1-19F as described in section 2.4.4. As the pVA891 replicon cannot function in pneumococci, Ery-resistant transformants are the result of a homologous recombination event directed by the cloned fragment of pneumococcal DNA that leads to the integration of the pVA891 plasmid into the host chromosome, and consequent disruption of the gene of interest. Furthermore, the cloned segment of pneumococcal DNA is duplicated and flanks the integrated copy of pVA891 as illustrated in **Fig. 2.1**.

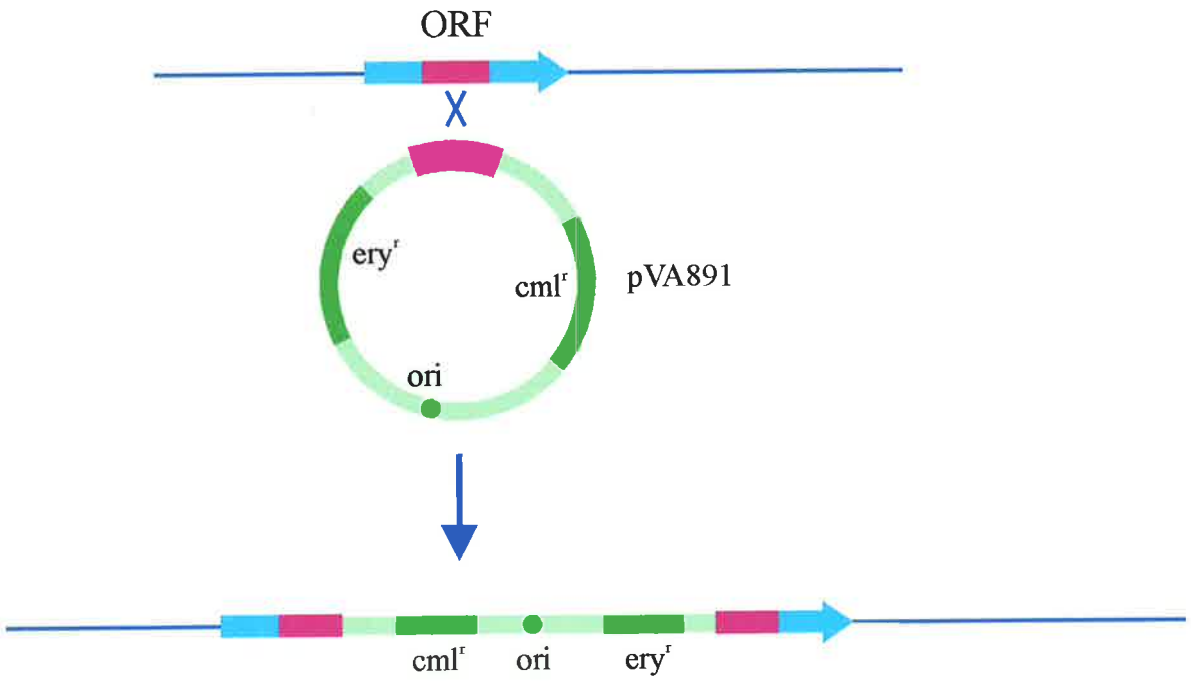


Fig. 2.1. Insertion-duplication mutagenesis. Recombination between the pneumococcal chromosome and the homologous DNA fragment (shown in pink) inserted into pVA891 (shown in green) results in the insertion of pVA891 into the chromosome and interruption of the ORF. The relative positions of the origin of replication and the antibiotic resistance markers for Cml and Ery are shown in dark green.

2.11 Plasmid insertion/rescue

This procedure was used to isolate pneumococcal DNA which flanks the pVA891 insertion site of an insertion-duplication mutant. Plasmid sequences along with flanking DNA were recovered from the mutants by digestion of chromosomal DNA (5 µg) with an appropriate restriction enzyme, self-religation of the cleaved DNA (at a concentration of 10 µg/ml) as shown in **Fig. 2.2**, followed by transformation into *E. coli* DH5α (section 2.4.2) and selection for either Cml or Ery resistance. The pneumococcal DNA was subsequently subcloned into pGEM-7Zf(+), pBluescript SK+, or pBluescript KS+ for further analysis.

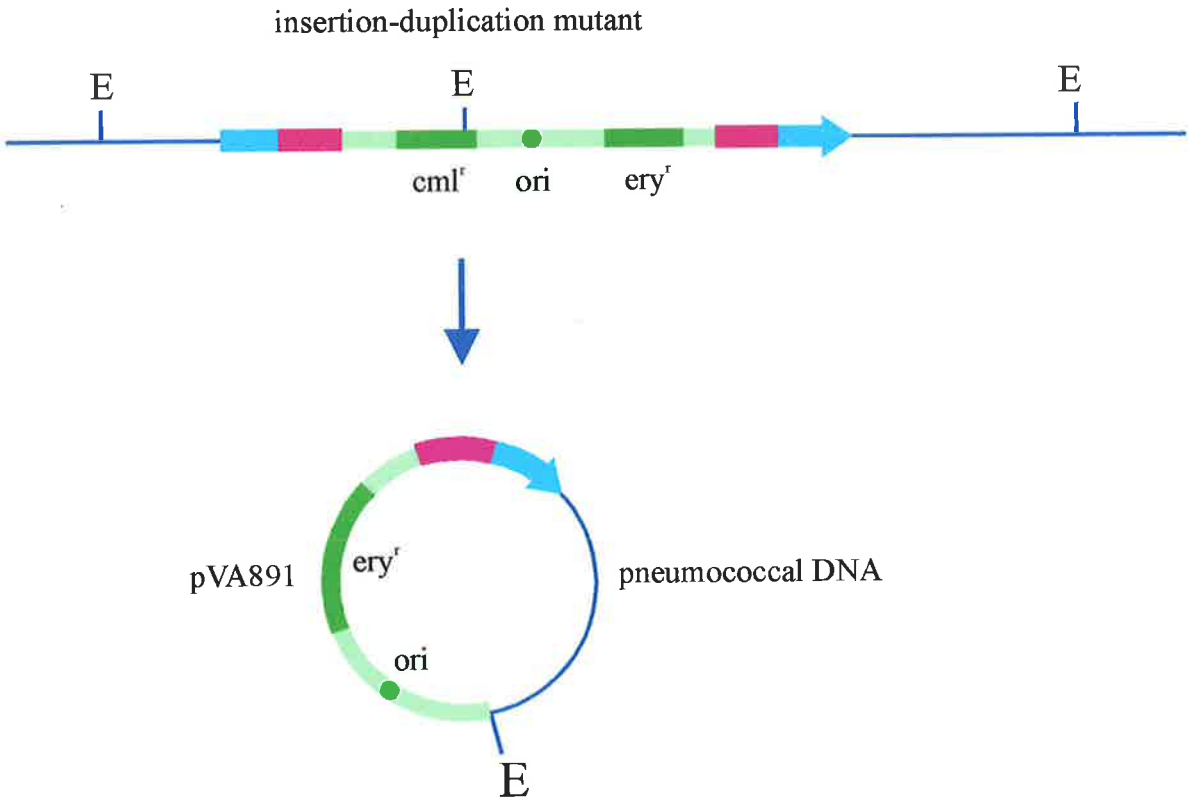


Fig. 2.2. Plasmid insertion/rescue. Chromosomal DNA from a pneumococcal insertion-duplication mutant is digested with an appropriate restriction endonuclease (E) and self-ligated to recircularise the pVA891 replicon (shown in green) containing additional pneumococcal DNA from the adjoining region. The relative positions of the origin of replication and the antibiotic resistance markers for Cml and Ery are shown in dark green.

2.12 Protein and LPS analysis

2.12.1 T7 expression

DH5(pGP1-2) derivatives were grown in 10 ml LB+KA (LB broth containing Amp and Kan) at 30°C for 16 h. The bacteria were then washed in 10 ml LB+KA and resuspended in 10 ml LB+KA. One ml was diluted to 30 ml in LB+KA and grown at 30°C to an OD of 0.3. The culture was then divided in two. Half was heat shocked at 42°C for 30 min to induce T7 expression and then grown at 37°C for 2 h after the addition of 200 μg/ml Rif. The other 15 ml was uninduced and grown at 30°C for the duration of the experiment. The bacteria were pelleted by centrifugation at 4,000 × g for 15 min at 4°C and resuspended in 50 μl water and 350 μl loading buffer (consisting of 62.5 mM Tris-HCl (pH 6.8) 2% (w/v) SDS, 10% (v/v) glycerol, 4% (v/v) β-mercaptoethanol, and 0.1% (w/v) bromophenol blue), heated at 100°C for 5 min, and stored at -20°C before SDS-PAGE analysis.

2.12.2 LPS preparation

One ml of an overnight culture was pelleted at 15,000 rpm for 2 min in a microfuge, the pellet was resuspended in 50 μl loading buffer and heated at 100°C for 5 min. Then 10 μl of proteinase K solution (2.5 mg/ml proteinase K in loading buffer) was added and incubated at 56°C for 4 h. The samples were stored at -20°C before SDS-PAGE analysis.

2.12.3 SDS-PAGE

SDS-PAGE was performed as described by Laemmli (1970) using Protean II gel apparatus (Bio-Rad Laboratories) and a Model 2000/200 power supply (Bio-Rad Laboratories). The 10 or 15% separating polyacrylamide gels consisted of 10 or 15% acrylamide, 142 mM Tris (pH 8.8), 0.05% SDS, 0.04% ammonium persulphate and 0.014% TEMED. The stacking gel consisted of 3% acrylamide 62 mM Tris (pH 6.8), 0.05% SDS, 0.04% ammonium persulphate and 0.011% TEMED. The polyacrylamide gels were 15 cm long, 14 cm wide, and 1.5 mm thick. The prepared samples were loaded using a loading syringe (Hamilton Company, Reno, Nevada, USA) prior to electrophoresis in electrode buffer (25 mM Tris, 25 mM glycine, 0.1% SDS, pH 8.3) at 350 V for 2-3 h. Proteins were stained overnight at room temperature in 0.1% (w/v) Coomassie Brilliant Blue R250, 10% (v/v) acetic acid, 25% (v/v) methanol, with gentle agitation. The gel was destained by gentle agitation in several changes of 10% (v/v) acetic acid, 10% (v/v) isopropanol at 65°C. The molecular weight markers (Bio-Rad) were myosin (200 kDa), β-galactosidase (116.25 kDa), phosphorylase b (97.4 kDa), serum albumin (66.2 kDa), ovalbumin (45 kDa), carbonic anhydrase (31 kDa), trypsin inhibitor (21.5 kDa), lysozyme (14.4 kDa) and aprotinin (6.5 kDa).

2.12.4 Western blotting

Samples were electrophoretically transferred from SDS-PAGE gels onto nitrocellulose filters as described by Towbin *et al.* (1979), in transblot buffer (25 mM Tris-HCl, 200 mM glycine, 5% (v/v) methanol, pH 8.3) at 350 mA for 2 h in a BioRad Transblot apparatus fitted with a water-cooled coil.

The nitrocellulose filter was then agitated for 20 min in 100 ml of TTBS (20 mM Tris-HCl, 150 mM NaCl, .05% (v/v) Tween-20, pH 7.4) containing 5% (w/v) skim milk.

This solution was then discarded and replaced with 50 ml TTBS containing 0.02% (w/v) skim milk and 10 µl of anti-*S. flexneri* type 4 serum (Murex, UK), the filter was then incubated at room temperature, with agitation, overnight. The filter was then washed for 10 min in three changes of 100 ml TTBS. The filter was then incubated at room temperature with agitation for 60 min in 50 ml TTBS containing 25 µl goat anti-rabbit IgG-AP conjugate. The filter was then washed for 5 min in four changes of 100 ml TTBS and once in DIG Buffer 3. The filter was then developed as described in section 2.7.2.4 except that colour development occurred much faster and it was not performed in the dark.

Chapter 3

CHARACTERISATION OF THE COMPLETE TYPE 19F CAPSULE LOCUS

3.1 Introduction

The cloning and sequencing of the first six genes of the type 19F *cps* locus has been reported previously (section 1.9.2; Guidolin *et al.*, 1994). At that time, no complete pneumococcal *cps* locus had been described for any serotype. Accordingly, the work described in this chapter was aimed at isolating, sequencing and where possible, functionally characterising the remainder of the pneumococcal type 19F *cps* locus.

3.2 Results

3.2.1 Isolation, cloning and DNA sequencing of the type 19F *cps* locus

The strategy used to obtain the remainder of the type 19F locus was to sequentially isolate DNA fragments from *S. pneumoniae* Rx1-19F (Morona *et al.*, 1994a) using either plasmid insertion/rescue or InPCR, and then clone them into pBluescript SK+ or pGEM-7Zf(+) for further analysis. The resultant recombinant plasmids pJCP464, pJCP465,

pJCP466, pJCP467, pJCP468 and pJCP469 are shown in **Fig. 3.1** and described in detail below.

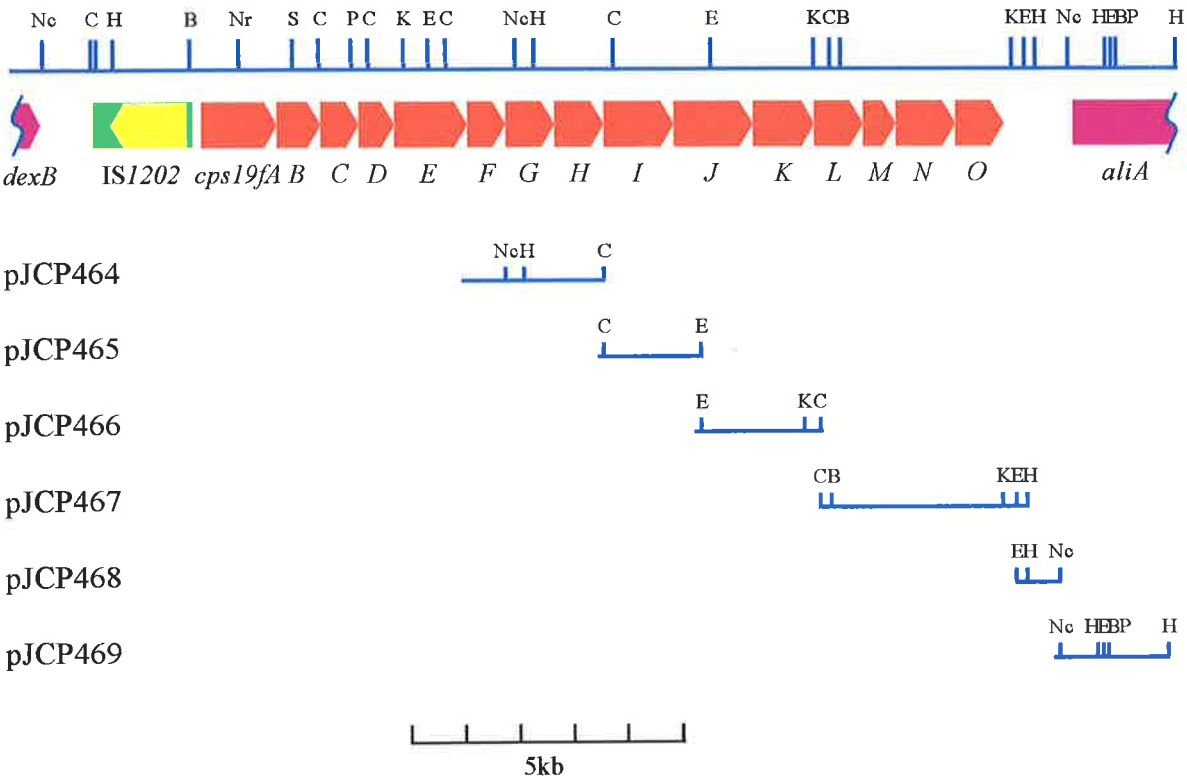


Fig. 3.1. Physical map of the chromosome of *S. pneumoniae* Rx1-19F in the vicinity of the *cps19f* locus. Boxed arrows represent potential ORFs and the closed box represents the insertion element IS*1202*. Gene designations are indicated below the map; *cps19fB-O* are abbreviated to *B-O*, respectively. 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. The regions of DNA subcloned into various recombinant plasmids are shown below the map.

The plasmid pJCP464 (**Fig. 3.1**) was constructed using plasmid insertion/rescue (described in section 2.11). The pVA891 replicon was excised, along with flanking pneumococcal DNA, from a derivative of *S. pneumoniae* Rx1-19F in which *cps19fF* had been interrupted by insertion of pVA891 (Guidolin *et al.*, 1994). The *ClaI* restricted DNA was recircularised and transformed into *E. coli* DH5α. The 2.65 kb pneumococcal DNA insert was excised from the rescued pVA891 derivative and subcloned into pBluescript

SK+. This plasmid was used to generate a series of nested deletion derivatives (section 2.8.1) which were then sequenced (section 2.8.2).

The available sequence data was then used to design primers J7 and J8 (**Table 2.4**) which were used for InPCR amplification (section 2.9.3) of *Eco*RI restricted and circularised chromosomal DNA to obtain a 2.0 kb segment of downstream DNA, overlapping the insert of pJCP464 by 100 bp. This PCR product was cloned into pGEM-7Zf(+) to generate pJCP465. Sequencing of the insert of this plasmid enabled design of primers J9 and J10 (**Table 2.4**) which were then used for InPCR amplification of *Cla*I restricted and circularised chromosomal DNA to obtain a further 2.2 kb of flanking DNA (again overlapping by 100 bp). This PCR product was also cloned into pGEM-7Zf(+) to generate pJCP466 (**Fig. 3.1**) and sequenced.

The next plasmid, pJCP467 (**Fig. 3.1**), was obtained by rescue of the pVA891 replicon plus flanking DNA from a derivative of *S. pneumoniae* Rx1-19F in which *cps19fJ* had been interrupted (described in section 3.2.4). The *Hin*dIII-restricted DNA was recircularised and transformed into *E. coli* DH5α, followed by subcloning of the 3.85-kb *Cla*I-*Hin*dIII pneumococcal DNA insert into pGEM-7Zf(+). Nested deletion derivatives of pJCP467 were sequenced. Sequence data across the *Cla*I site was obtained by dye terminator sequencing (section 2.8.3) of the rescued pVA891 based plasmid with the primer J12 (**Table 2.4**).

Two further rounds of InPCR were then used to construct pJCP468 and pJCP469. The PCR product obtained from *Nco*I restricted and circularised chromosomal DNA with primers J19 and J20 (**Table 2.4**) was cloned in pGEM-7Zf(+) to generate pJCP468 (**Fig. 3.1**) and then sequenced. Primers J26 and J27 (**Table 2.4**) were designed for inverse PCR amplification of *Cla*I restricted and circularised chromosomal DNA. The 4-kb PCR product was cloned into pBluescript SK+ to generate pJCP469 (**Fig. 3.1**). However only

the first 2.1 kb (to the *Hin*dIII site) was sequenced using the M13 forward primer and primers J32, J33 and J34 (**Table 2.4**) as well as sequence of a subclone containing the internal 1.2-kb *Hin*dIII fragment using the M13 forward and reverse primers. Only the part of pJCP469 which was sequenced is shown in **Fig. 3.1.**

3.2.2 Analysis of the *cps19f* DNA sequence

Both strands of the pneumococcal DNA inserts of each of the above plasmids (or nested derivatives thereof) were subjected to sequence analysis in order to compile the complete sequence of the remainder of *cps19f*, as shown in **Appendix I.** Examination of the compiled sequence revealed the presence of a further nine potential open reading frames (ORFs), which have been designated *cps19fG* to *cps19fO*. Each ORF is preceded by a ribosome binding site and the majority are very closely linked. The only potentially significant intergenic gaps occur between *cps19fJ* and *cps19fK* (63 nucleotides) and between *cps19fN* and *cps19fO* (65 nucleotides). However, potential stemmed-loop structures were not found immediately downstream of *cps19fJ* or *cps19fN* and no obvious promoter sequences were seen immediately upstream of *cps19fK* or *cps19fO*. The first six genes of the *cps19f* locus are also closely linked (intergenic distances 1-15 nucleotides) (Guidolin *et al.*, 1994) and so co-transcription of the entire locus remains a possibility.

A large region (1,458 nucleotides) downstream of *cps19fO* did not appear to contain any significant ORFs on either DNA strand, but the region from nucleotides 15,390-15,630 contained numerous stemmed-loop structures reminiscent of transcription terminators. There was, however, an additional ORF commencing at nucleotide 16,446, which was preceded by a ribosome binding site and by -10 and -35 promoter sequences. Comparison of sequence data for this gene with those deposited on GenBank indicated 97.3% DNA identity with that reported for the pneumococcal *aliA* gene, which encodes an

oligopeptide binding protein (Alloing *et al.*, 1994) and so is unrelated to CPS biosynthesis. This information, combined with earlier results (section 1.9) show that in *S. pneumoniae* type 19F, the *cps* locus is located between *dexB* and *aliA* in the chromosome, as shown in **Fig. 3.1**.

3.2.3 Characterisation of *cps19fG-O*

The locations and several properties of each of the ORFs designated *cps19fG-O* are summarised in **Table 3.1**. Significant similarities with other known proteins, revealed by comparison with sequence databases, are described below.

Table 3.1. Summary of ORFs *cps19fG-O*.

ORF	Location in sequence	Predicted MW	No. amino acids	Hydrophobicity index[a]	Predicted pI	%G+C content[b]
cps19fG	5,883-6,693	31,647	269	-0.39	8.43	36.3
cps19fH	6,694-7,572	34,474	292	-0.54	7.80	30.3
cps19fI	7,573-8,910	51,734	445	0.68	9.59	29.7
cps19fJ	8,933-10,354	55,055	473	0.81	9.83	29.7
cps19fK	10,418-11,506	40,950	362	-0.30	5.48	35.2
cps19fL	11,545-12,414	32,215	289	-0.21	4.69	42.3
cps19fM	12,415-13,011	22,379	198	-0.40	5.05	41.5
cps19fN	13,021-14,070	39,053	349	-0.45	5.16	42.1
cps19fO	14,136-14,986	32,330	283	-0.50	4.71	41.5

a. According to Kyte and Doolittle (1982), as implemented in PROSIS.
b. Percent guanine plus cytosine (G+C) of coding region.

3.2.3.1 *cps19fG*

The *cps19fG* gene encodes a putative 31.6 kDa protein which exhibits 25.9% identity with the LicD protein of *H. influenzae*, which is encoded by the *licD* gene of its LPS locus (Weiser *et al.*, 1989). However, the precise function of LicD is unknown.

Cps19fG also has a low level of similarity to the central portion of the Mnn4 gene product which is required for the addition of mannosylphosphate to *N*-linked oligosaccharides in *S. cerevisiae* (Odani *et al.*, 1996). The alignment between these three proteins is shown in **Fig. 3.2**; the similarity is greatest at the N-terminal end of Cps19fG where a conserved motif has been identified. This motif includes aspartate (D) residues, which are predicted to be present in the catalytic sites of glycosyl transferases (Saxena *et al.*, 1995). The structure of the *H. influenzae* LPS is highly variable, but Phillips *et al.* (1993) suggest it may be phosphorylated. The type 19F CPS and the cell wall mannans in *S. cerevisiae* are both phosphorylated. Thus it is possible that Cps19fG, LicD and Mnn4 are all involved in the phosphorylation of polysaccharides.

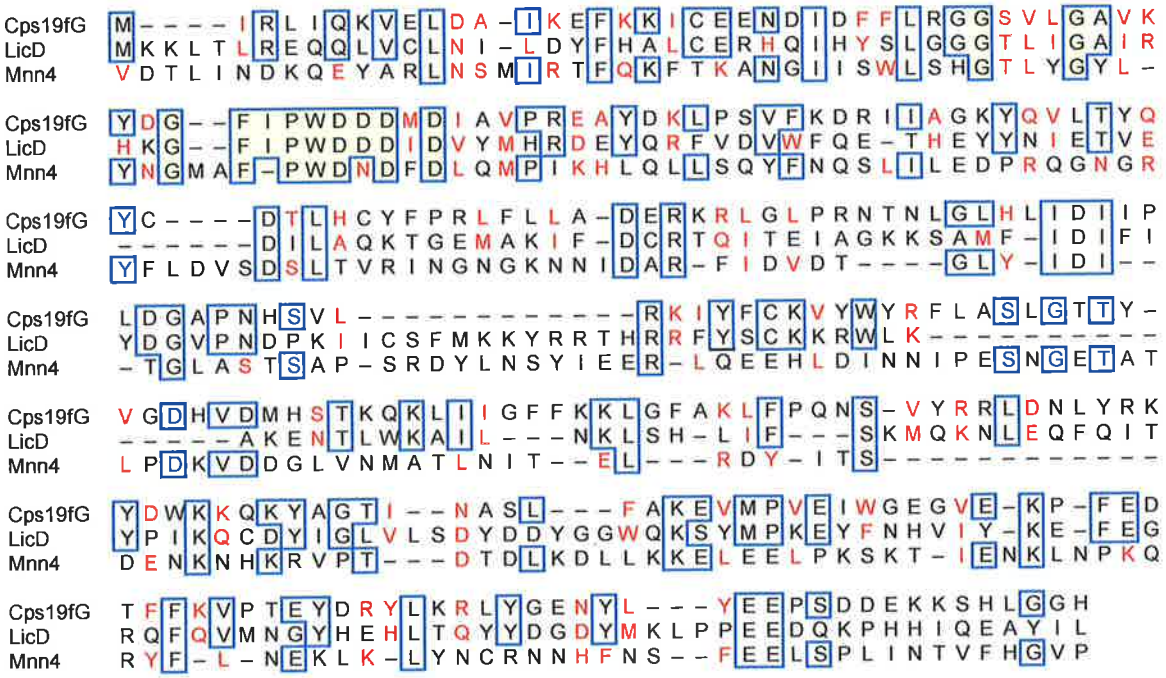


Fig. 3.2. Alignment of Cps19fG. Alignment of Cps19fG with *H. influenzae* LicD (LicD) (Weiser *et al.*, 1989) and amino acids 465-730 of *S. cerevisiae* Mnn4 (Mnn4) (Odani *et al.*, 1996) using the default settings of the program CLUSTAL (Higgins and Sharp, 1988) and enhanced by manual adjustment. . Residues identical to those in Cps19fG are boxed; similar residues are shown in red; - indicates absence of a residue. The yellow shaded regions correspond to an apparently conserved motif in these proteins.

3.2.3.2 *cps19fH*

The *cps19fH* gene encodes a putative 34.5 kDa protein, assuming that translation is initiated at the TTG codon, as shown in **Appendix I**. The nearest in-frame ATG codon is a further 90 codons downstream (the expected translation product would be approximately 20 kDa) and is not preceded by a ribosome binding site. Cps19fH has a limited degree of identity with rhamnosyl transferases from *Yersinia enterocolitica* and *S. flexneri*, as shown in the alignment of these proteins in **Fig. 3.3**. Interestingly, the more conserved regions correspond to a motif previously identified in a number of rhamnosyl and other 6-deoxyhexosyl transferases (Morona *et al.*, 1995). This suggests that Cps19fH is likely to be the rhamnosyl transferase involved in incorporation of rhamnose (Rha) into the type 19F CPS.

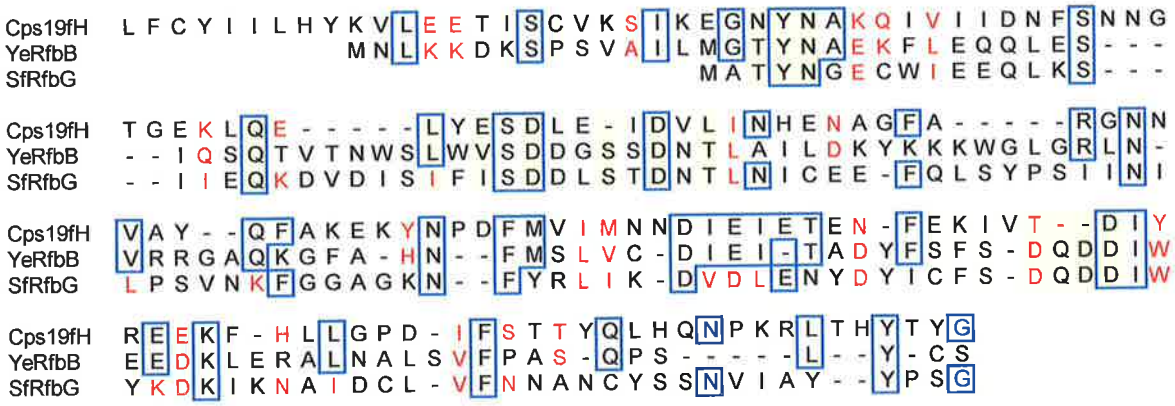


Fig. 3.3. Alignment of Cps19fH. Alignment of Cps19fH, from amino acid position 1 to 143, with *Y. enterocolitica* RfbB (YeRfbB) (Zhang *et al.*, 1993) and *S. flexneri* RfbG (SfRfbG) (Morona *et al.*, 1995) using the default settings of the program CLUSTAL (Higgins and Sharp, 1988) and enhanced by manual adjustment. Residues identical to those in Cps19fH are boxed; similar residues are shown in red; - indicates absence of a residue. The yellow shaded regions correspond to those found to be most conserved amongst a variety of Gram-negative rhamnosyl and 6-deoxy-hexosyl transferases (Morona *et al.*, 1995).

3.2.3.3 *cps19fI*

The *cps19fI* gene encodes a putative 51.7 kDa protein, which has similarity to Rfc proteins (O-antigen polymerases) from a variety of Gram-negative bacteria, as shown in **Table 3.2**. Although the overall similarity between Cps19fI and the various Rfc proteins is low (14.0 - 20.1% identity) it is as strong as the degree of identity within the Gram-negative species (15.5-19.8%). The relationship between Cps19fI and Rfc proteins is even more apparent when similar as well as identical amino acids are considered, as exemplified by the alignment with *E. coli* K12 Rfc (Stevenson *et al.*, 1994) shown in **Fig. 3.4**. The hydropathy plots for the various proteins (**Fig. 3.5**) also illustrate the marked similarity, each having 10-12 hydrophobic, potentially membrane-spanning, domains. It therefore seems probable that Cps19fI is the polysaccharide polymerase.

Table 3.2. Similarity of Cps19fI to other proteins.

	% Identity[a]						
	Cps19fI	K12Rfc	SdRfc	StRfc	SfRfc	M67Rfc	M40Rfc
Cps19fI[b]	100	16.5 [394]	19.6 [382]	15.6 [417]	14 [350]	18.3 [394]	20.1 [318]
K12Rfc[c]		100	17.5 [382]	19.1 [397]	19.4 [309]	16.5 [358]	17 [335]
SdRfc[d]			100	15.5 [381]	16.5 [315]	18.5 [211]	16.5 [363]
StRfc[e]				100	18.8 [308]	19.7 [269]	17.5 [341]
SfRfc[f]					100	17.5 [331]	19.8 [349]
M67Rfc[g]						100	17.5 [360]
M40Rfc[h]							100

a. Percentage of identical amino acids determined with FASTA as implemented in PROSIS. Numbers in parentheses indicate the number of amino acids over which the % identity occurs. b. *S. pneumoniae* Cps19fI. c. *E. coli* K12 Rfc (Stevenson *et al.*, 1994). d. *Shigella dysenteriae* type 1 Rfc (Klena and Schnaitman, 1993). e. *S. enterica* serovar typhimurium Rfc (Collins and Hackett, 1991). f. *S. flexneri* Rfc (Morona *et al.*, 1994b). g. *S. enterica* serovar muenchen strain M67 Rfc (Brown *et al.*, 1992). h. *S. enterica* serovar montevideo strain M40 Rfc (Lee *et al.*, 1992).

Fig. 3.4. Alignment of Cps19fI. Alignment of Cps19fI with *E. coli* K12 Rfc (K12Rfc) (Stevenson *et al.*, 1994), using the default settings of the program CLUSTAL (Higgins and Sharp, 1988). Identical residues are boxed; similar residues are shown in red; - indicates absence of a residue.

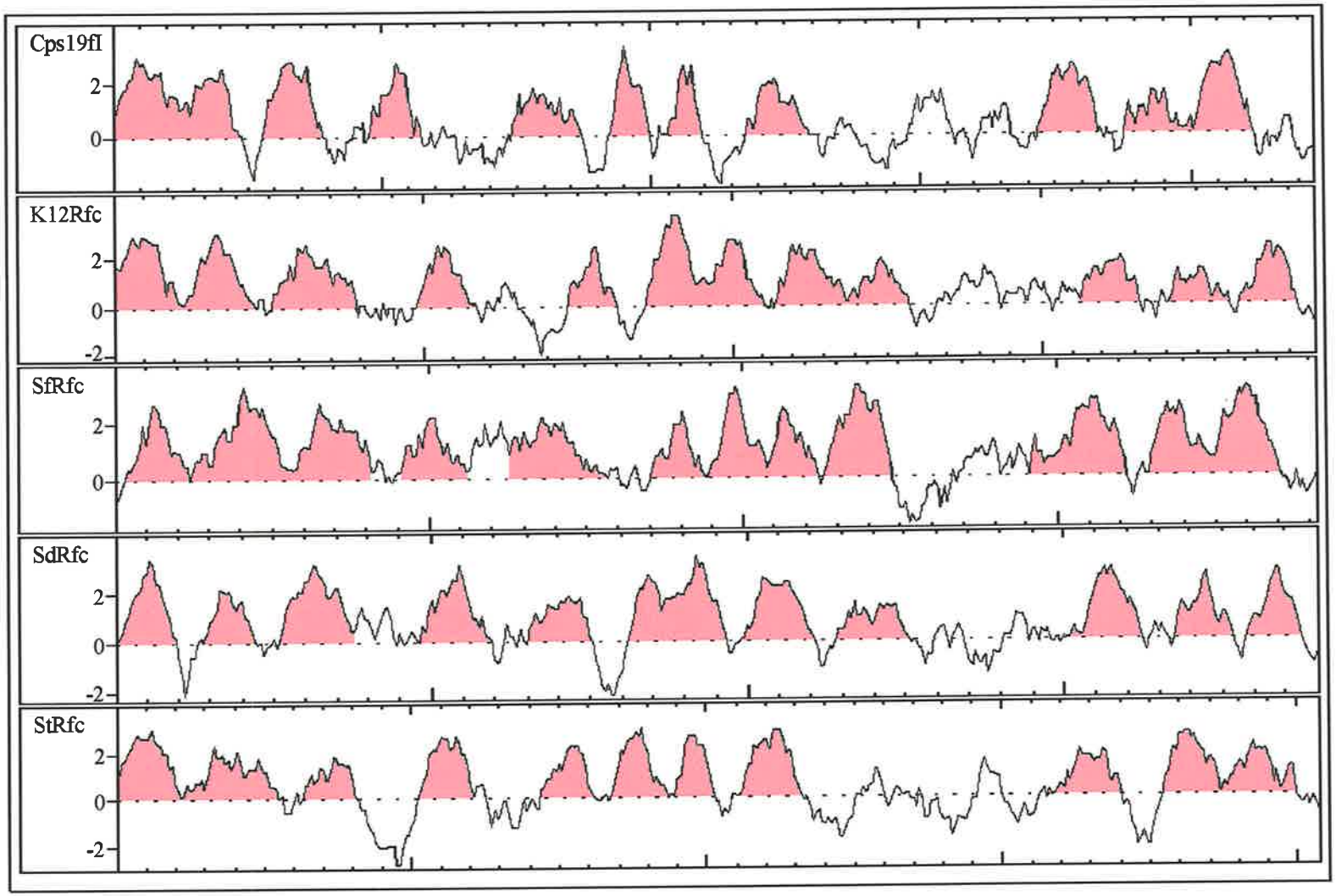


Fig. 3.5. Hydropobicity analysis of Cps19fI. Hydropathy plots of Cps19fI, *E. coli* K12 Rfc (K12Rfc) (Stevenson *et al*, 1994), *S. flexneri* Rfc (SfRfc) (Morona *et al*, 1994b), *S. dysenteriae* Rfc (SdRfc) (Klena and Schnaitman, 1993) and *S. enterica* serovar typhimurium Rfc (StRfc) (Collins and Hackett, 1991) were generated by the method of Kyte and Doolittle (1982) and aligned using PROFILEGRAPH (Hofmann and Stöffel, 1989). Positive numbers on the Y-axis indicate hydrophobic regions and putative membrane-spanning domains are shaded. The position of every 10th amino acid is marked on each X-axis.

3.2.3.4 *cps19fJ*

The *cps19fJ* gene encodes a putative 55.1 kDa protein with similarity to RfbX proteins of *E. coli, Shigella* sp. and *Yersinia* sp., as well as to the CapF protein of *S. aureus*, as shown in **Table 3.3**. The RfbX proteins are known to be involved in transport of O-antigen repeat units across the membrane (Liu *et al.*, 1996; Macpherson *et al.*, 1995). Again, the overall similarity between Cps19fJ and the various Gram-negative proteins is low (15.7-19.4% identity), but the degree of identity within the Gram-negative species is similar (16.4-31.4%). Moreover, the hydropathy plots for the Cps19fJ and the various RfbX-related proteins are very similar (**Fig. 3.6**). These are all integral membrane proteins with 10-12 hydrophobic, potentially membrane-spanning, domains Thus, Cps19fJ is probably the polysaccharide repeat unit transporter.

Table 3.3. Similarity of Cps19fJ to other proteins.

	% Identity[a]						
	Cps19fJ[b]	SfRfbX[c]	K12RfbX[d]	SdRfbX[e]	YeTrgA[f]	YpRfbX[g]	SaCapF[h]
Cps19fJ	100	19.2	18.1	19.4	19.2	15.7	16.4
		[421]	[414]	[402]	[416]	[396]	[397]
SfRfbX		100	19.7	22.5	19.8	19.9	19
			[396]	[409]	[424]	[403]	[410]
K12RfbX			100	31.4	28.3	17.4	21.5
				[401]	[406]	[373]	[395]
SdRfbX				100	28.4	18.6	21.1
					[401]	[269]	[393]
YeTrsA					100	20.8	20.4
						[394]	[401]
YpRfbX						100	16.1
							[367]
SaCapF							100

a. Percentage of identical amino acids determined with FASTA as implemented in PROSIS. Numbers in parentheses indicate the number of amino acids over which the % identity occurs.
b. *S. pneumoniae* Cps19fJ.
c. *S. flexneri* RfbX (Macpherson *et al.*, 1995).
d. *E. coli* K12 RfbX (Stevenson *et al.*, 1994).
e. *S. dysenteriae* RfbX (Klena and Schnaitman, 1993).
f. *Y. enterocolitica* TrsA (Skurnik *et al.*, 1995).
g. *Y. pseudotuberculosis* RfbX (Kessler *et al.*, 1993).
h. *S. aureus* CapF (Lin *et al.*, 1994).

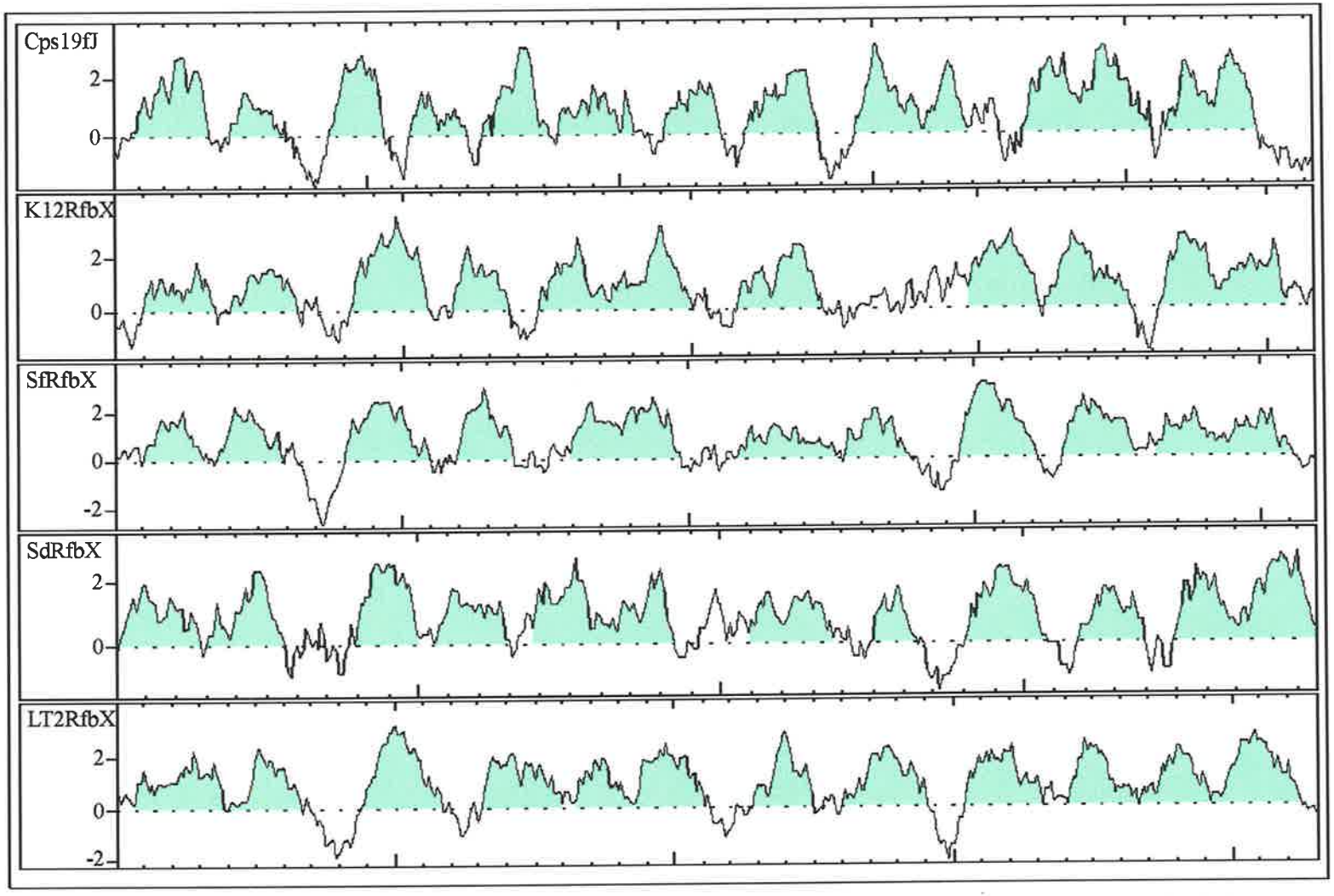


Fig. 3.6. Hydropobicity analysis of Cps19fJ. Hydropathy plots of Cps19fJ, *E. coli* K12 RfbX (K12RfbX) (Stevenson *et al*, 1994), *S. flexneri* RfbX (SfRfbX) (Macpherson *et al*, 1995), *S. dysenteriae* RfbX (SdRfbX) (Klena and Schnaitman, 1993) and *S. enterica* serovar typhimurium LT2 RfbX (LT2RfbX) (Jiang *et al*, 1991) were generated by the method of Kyte and Doolittle (1982) and aligned using PROFILEGRAPH (Hofmann and Stöffel, 1989). Positive numbers on the Y-axis indicate hydrophobic regions and putative membrane-spanning domains are shaded. The position of every 10th amino acid is marked on each X-axis.

3.2.3.5 *cps19fK*

The *cps19fK* gene encodes a putative 40.9 kDa protein. Cps19fK has a high degree of similarity (49% to 63% identity,) with a family of proteins including the *E. coli rffE* gene product (**Table 3.4**). The alignment of Cps19fK with these proteins is shown in **Fig. 3.7**. RffE is a UDP-Glc*N*Ac-2-epimerase, and functions in the synthesis of UDP-Man*N*Ac, a precursor required for the synthesis of UDP-*N*-acetyl mannosaminuronic acid in ECA biosynthesis (Meier-Dieter *et al.*, 1990). Man*N*Ac is a component of the type 19F CPS and thus *cps19fK* is presumed to encode the enzyme needed to synthesise UDP-Man*N*Ac in type 19F pneumococci.

Table 3.4. Similarity of Cps19fK to other proteins.

	%Identity[a]					
	Cps19fK[b]	BsYvyH[c]	K12RffE[d]	PsEpsC[e]	054RfbC[f]	Syn0624[g]
Cps19fK	100	63.2 [359]	50.8 [368]	50.3 [372]	49.7 [374]	49 [357]
BsYvyH		100	54.9 [368]	52 [373]	52.9 [367]	47 [362]
K12RffE			100	65 [369]	54.8 [367]	47.3 [364]
PsEpsC				100	52.6 [367]	45.8 [369]
054RfbC					100	44.4 [365]
Syn0624						100

a. Percentage of identical amino acids determined with FASTA as implemented in PROSIS. Numbers in parentheses indicate the number of amino acids over which the % identity occurs.
b. *S. pneumoniae* Cps19fK
c. *B. subtilis* YvyH (Soldo *et al.*, 1993)
d. *E. coli* K12 RffE (Daniels *et al.*, 1992)
e. *P. solanacearum* EpsC (Huang and Schell, 1995)
f. *S. enterica* O54 RfbC (Keenleyside and Whitfield, 1995)
g. *Synechocystis* sp. ORF 0624 (Kaneko *et al.*, 1995)

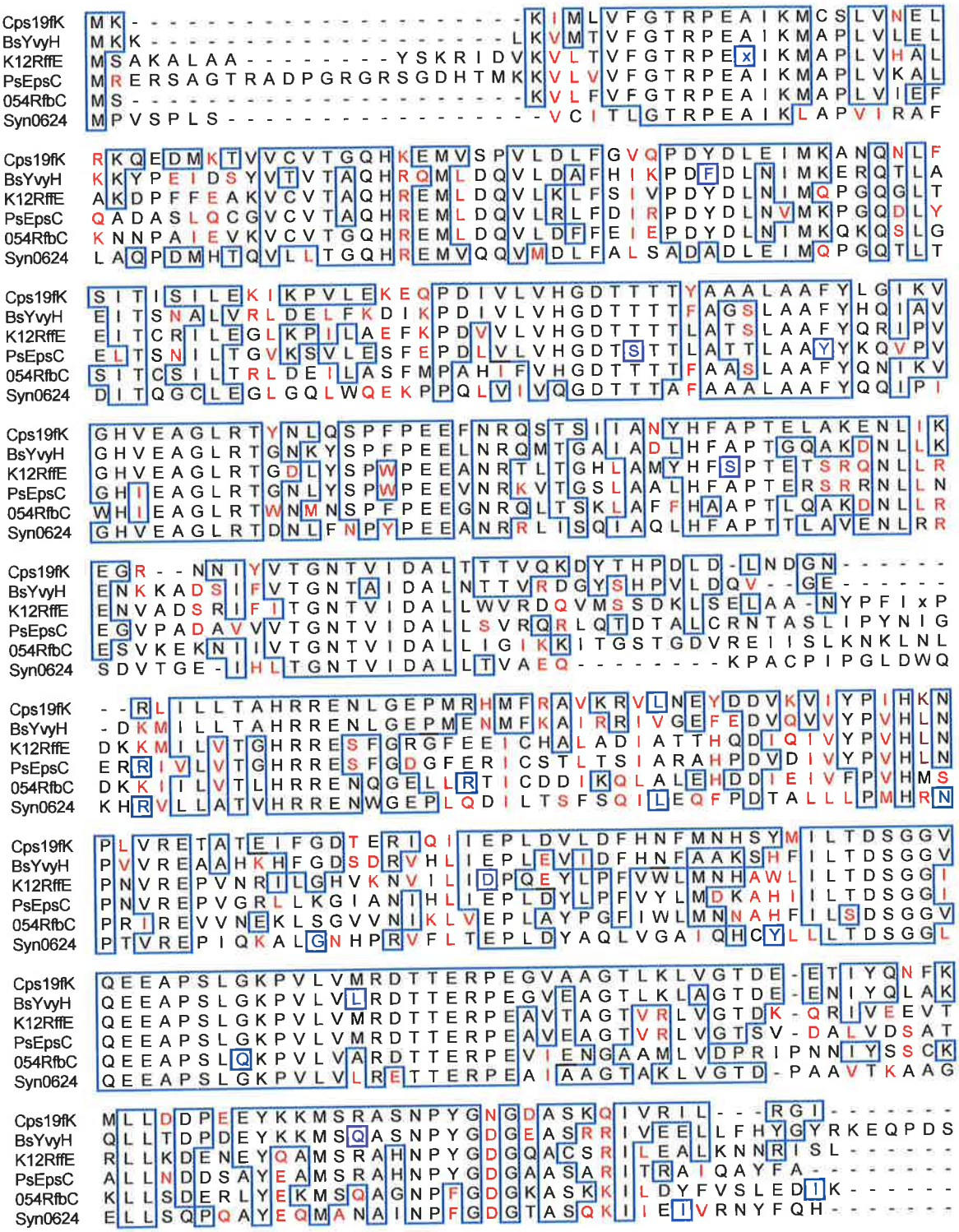


Fig. 3.7. Alignment of Cps19fK. Alignment of Cps19fK with *B. subtilis* YvyH (BsYvyH) (Soldo *et al.*, 1993), *E. coli* K12 RffE (K12RffE) (Daniels *et al.*, 1992), *P. solanacearum* EpsC (PsEpsC) (Huang and Schell, 1995), *S. enterica* O54 RfbC (O54RfbC) (Keenleyside and Whitfield, 1995), and *Synechocystis* sp. ORF 0624 (Syn0624) (Kaneko *et al.*, 1995), as determined using the default settings of the program CLUSTAL (Higgins and Sharp, 1988). Residues identical to those in Cps19fK are boxed; similar residues are shown in red; - indicates absence of a residue.

3.2.3.6 *cps19fLMNO*

The *cps19fL, cps19fM, cps19fN* and *cps19fO* genes encode proteins of 32.2, 22.4, 39.1 and 32.3 kDa, respectively. These four genes have extensive similarity to a portion of the *S. flexneri rfb* gene cluster (*rfbBDAC*) which encodes the enzymes involved with dTDP-Rha biosynthesis (Macpherson *et al.*, 1994).

Cps19fL has a high degree of similarity (69.1-70.6% identity) to Glc-1-phosphate thymidylyl transferases (RfbA) from *E. coli*, *S. flexneri* and *S. enterica* (**Table 3.5**). RfbA catalyses the first step in the dTDP-Rha biosynthesis pathway, suggesting *cps19fL* may also encode this enzyme. The alignment between Cps19fL and the RfbA proteins is shown in **Fig. 3.8**. Interestingly, there is also significant nucleotide sequence similarity (67-70% identity) between the *cps19fL* and *rfbA* genes.

Cps19fM has 24.3% to 34.6% identity to RfbC from various Gram-negative bacteria and StrM from *Streptomyces griseus* (**Table 3.6**, **Fig. 3.9**). Thus, *cps19fM* probably encodes dTDP-4-keto-6-deoxyglucose-3,5-epimerase, the third enzyme in the dTDP-Rha biosynthesis pathway.

Cps19fN has similarity to a large family of dehydratases, in particular to RfbB from various Gram-negative bacteria and StrE from *S. griseus* (**Table 3.7**, **Fig. 3.10**). An important feature of these closely related proteins is the presence of a NAD-binding domain (shaded residues in **Fig. 3.10**; Macpherson *et al.*, 1994). Thus, Cps19fN is probably a dTDP-Glc-4,6-dehydratase, the second enzyme in the dTDP-Rha biosynthesis pathway.

Cps19fO exhibits 27% to 39.7% identity to RfbD from various Gram-negative bacteria and StrL from *S. griseus* (**Table 3.8**, **Fig. 3.11**). These proteins also have a NAD-binding domain (shaded residues in **Fig. 3.11**). Thus Cps19fO is probably dTDP-L-Rha synthase, which catalyses the last step in dTDP-Rha biosynthesis.

Table 3.5. Similarity of Cps19fL to other proteins.

	% Identity[a]															
	Sf RfbA	K12 RfbA	O7 RfbA	M32 RfbA	LT2 RfbA	Nm RfbA	Ng RfbA	Ye RfbA	Xc RfbA	Ec RffM	Bs SpsI	Stf TylA	Sg StrD	Sp DnrL	Sn SnoD	Sv GraD
Cps19fL[b]	68.8 [285]	69.4 [288]	70.6 [286]	70.5 [285]	69.1 [285]	66.9 [285]	67.6 [287]	66.1 [289]	64.1 [284]	64.3 [286]	45.3 [243]	54.4 [285]	35.4 [237]	33.6 [259]	33.3 [249]	36.2 [232]
SfRfbA[c]	100	89.3 [291]	93.2 [293]	91.1 [292]	89.3 [291]	63.2 [285]	63.5 [285]	64 [286]	65.9 [290]	62.4 [287]	43.1 [211]	50.7 [286]	33.9 [242]	37.9 [198]	33.5 [218]	34.9 [215]
K12RfbA[d]		100	94.2 [292]	92.4 [291]	92.1 [291]	65.1 [284]	65.1 [284]	65.6 [288]	68.3 [290]	64 [286]	43 [237]	51.7 [288]	34.6 [243]	31.4 [280]	33.7 [243]	36 [228]
O7RfbA[e]			100	95.6 [293]	93.5 [292]	65 [286]	65 [286]	65.9 [287]	67.7 [291]	63.9 [288]	42.9 [238]	52.4 [288]	35.7 [244]	32.8 [268]	33.6 [244]	35.8 [229]
M32RfbA[f]				100	95.2 [291]	66 [285]	66 [285]	64.7 [286]	67.2 [290]	63.1 [287]	43 [237]	51.6 [287]	35.4 [243]	33 [267]	32 [266]	36 [228]
LT2RfbA[g]					100	64.2 [285]	64.2 [285]	64.3 [286]	67.2 [290]	61.7 [287]	43.5 [237]	50.9 [287]	35 [243]	32.7 [266]	31.5 267	36 [228]
NmRfbA[h]						100	91.3 [288]	69.1 [285]	63.4 [284]	69.9 [286]	44.4 [239]	50.5 [285]	35.9 [237]	35.3 [232]	34.9 [232]	34.1 [232]
NgRfbA[i]							100	69.5 [285]	63.4 [284]	70.6 [286]	42.7 [239]	48.8 [285]	35.9 [237]	36.2 [232]	34.5 [232]	33.6 [232]
YeRfbA[j]								100	61.2 [286]	72.1 [287]	42.6 [244]	50 [286]	35.2 [244]	35.5 [234]	32.6 [233]	33.2 [232]
XcRfbA[k]									100	61.9 [286]	43.7 [245]	50.5 [289]	38.1 [247]	37.9 [232]	35.1 [248]	33.3 [243]
EcRffM[l]										100	42.2 [244]	51.2 [285]	35.4 [243]	36.1 [233]	32.1 [243]	33.3 [228]
BsSpsI[m]											100	39.5 [248]	28.9 [239]	30.8 [247]	32.5 [240]	31.3 [243]
StfTylA[n]												100	37.7 [236]	37.7 [236]	28.7 [268]	34.2 [231]
SgStrD[o]													100	59.2 [355]	54.1 [355]	51 [355]
SpDnrL[p]														100	49 [355]	49.2 [355]
SnSnoD[q]															100	57.2 [355]
SvGraD[r]																100

a. Percentage of identical amino acids determined with FASTA as implemented in PROSIS. Numbers in parentheses indicate the number of amino acids over which the % identity occurs.
b. *S. pneumoniae* Cps19fL
c. *S. flexneri* RfbA (Macpherson *et al.*, 1994)
d. *E. coli* K12 RfbA (Stevenson *et al.*, 1994)
e. *E. coli* O7 RfbA (Marolda and Valvano, 1995)
f. *S. enterica* E1 RfbA (Wang *et al.*, 1992)
g. *S. enterica* serovar typhimurium strain LT2 RfbA (Jiang *et al.*, 1991),
h. *N. meningitidis* RfbA (Hammerschmidt *et al.*, 1994)
i. *Neisseria gonorrhoeae* RfbA (Robertson *et al.*, 1994)
j. *Y. enterocolitica* RfbA (Zhang *et al.*, 1993)
k. *X. campestris* RfbA (Koplin *et al.*, 1993)
l. *E. coli* RffM (Daniels *et al.*, 1992)
m. *B. subtilis* SpsI P39629 (GenBank accession no. Z99123)
n. *Streptomyces fradiae* TylA (Merson-Davies and Cundliffe, 1994)
o. *S. griseus* StrD (Distler *et al.*, 1987)
p. *Streptomyces peucetius* DnrL (Gallo *et al.*, 1996)
q. *Streptomyces nodosus* SnoD (GenBank accession no. A25110)
r. *Streptomyces violaceoruber* GraD (Bechthold *et al.*, 1995)

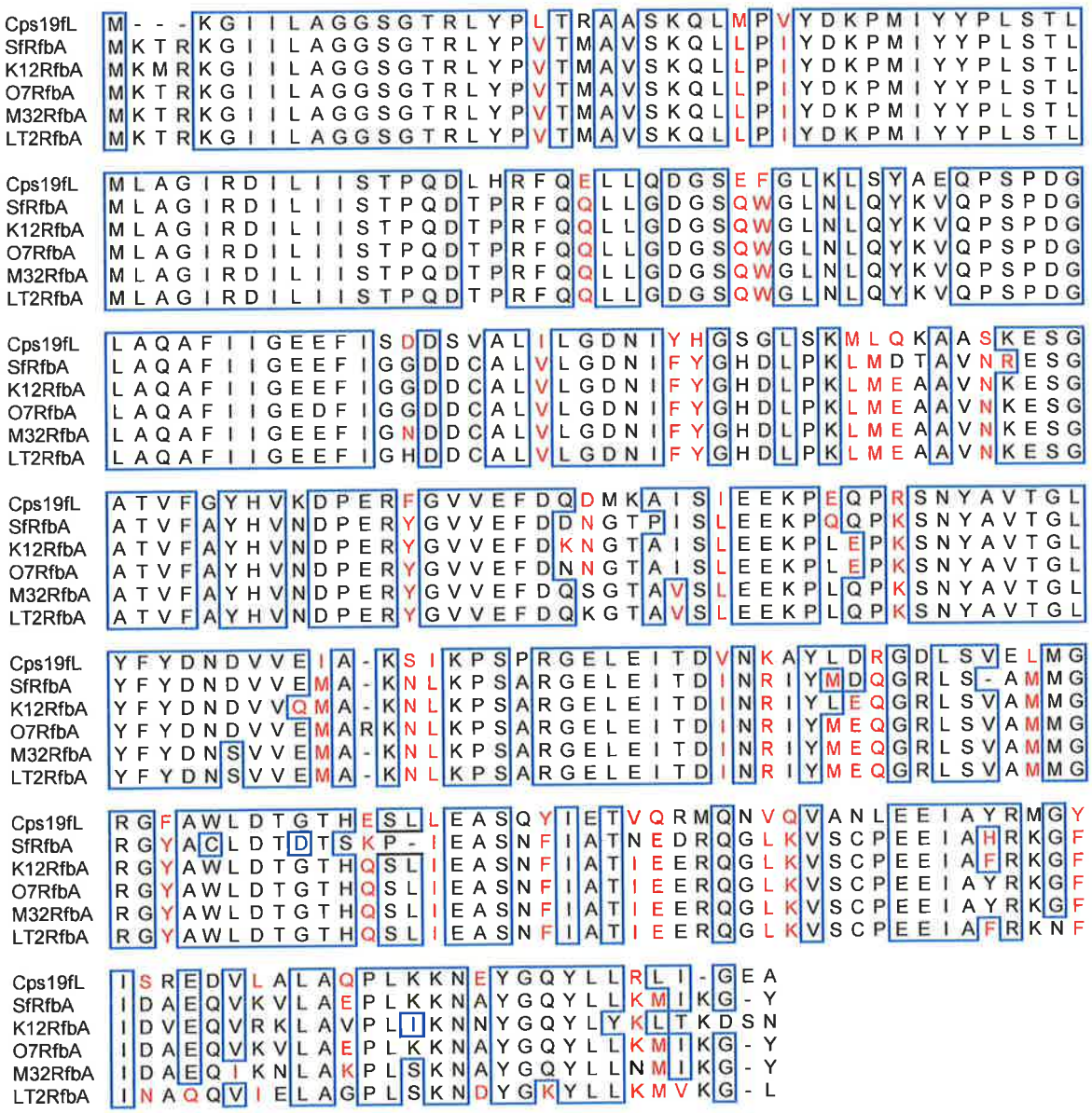


Fig. 3.8. Alignment of Cps19fL. Alignment of the amino acid sequences of Cps19fL, *S. flexneri* RfbA (SfRfbA) (Macpherson *et al.*, 1994), *E. coli* K12 RfbA (K12RfbA) (Stevenson *et al.*, 1994), *E. coli* O7 RfbA (O7RfbA) (Marolda and Valvano, 1995), *S. enterica* E1 RfbA (M32RfbA) (Wang *et al.*, 1992) and *S. enterica* serovar typhimurium LT2 RfbA (LT2RfbA) (Jiang *et al.*, 1991), as determined using the default settings of the program CLUSTAL (Higgins and Sharp, 1988). Residues identical to those in Cps19fL are boxed; similar residues are shown in red; - indicates absence of a residue.

Table 3.6. Similarity of Cps19fM to other proteins.

	% Identity[a]								
	XcRfbC	YeRfbF	M32RfbC	LT2RfbC	NmRfbC	SfRfbC	07RfbC	K12RfbC	SgStrM
Cps19fM[b]	34.6	32.3	31.6	30.9	31.5	31.8	24.3	29.3	29.1
	[179]	[161]	[177]	[178]	[178]	[157]	[177]	[184]	[141]
XcRfbC[c]	100	52.7	56.7	58	46.5	51.7	50.9	52.8	33.3
		[150]	[157]	[157]	[187]	[178]	[171]	[159]	[156]
YeRfbF[d]		100	52.9	55.4	52	54	48	49.7	39.4
			[166]	[175]	[173]	[174]	[177]	[181]	[175]
M32RfbC[e]			100	70.4	57.7	83.3	71	66.3	36.3
				[189]	[175]	[168]	[176]	[169]	[168]
LT2RfbC[f]				100	58.7	71.9	65.7	66.1	35.3
					[179]	[178]	[175]	[177]	[167]
NmRfbC[g]					100	59.1	51.2	55.6	37.9
						[181]	[172]	[169]	[161]
SfRfbC[h]						100	73.6	66.9	35.5
							[178]	[169]	[172]
07RfbC[i]							100	58.6	37.9
								[181]	[174]
K12RfbC[j]								100	34.8
									[178]

a. Percentage of identical amino acids determined with FASTA as implemented in PROSIS. Numbers in parentheses indicate the number of amino acids over which the % identity occurs.
b. *S. pneumoniae* Cps19fM
c. *X. campestris* RfbC (Koplin *et al.*, 1993)
d. *Y. enterocolitica* RfbF (Zhang *et al.*, 1993)
e. *S. enterica* E1 RfbC (Wang *et al.*, 1992)
f. *S. enterica* serovar typhimurium strain LT2 RfbC (Jiang *et al.*, 1991)
g. *N. meningitidis* RfbC (Hammerschmidt *et al.*, 1994)
h. *S. flexneri* RfbC (Macpherson *et al.*, 1994)
i. *E. coli* O7 RfbC (Marolda and Valvano, 1995)
j. *E. coli* K12 RfbC (Stevenson *et al.*, 1994)
k. *S. griseus* StrM (Distler *et al.*, 1987)

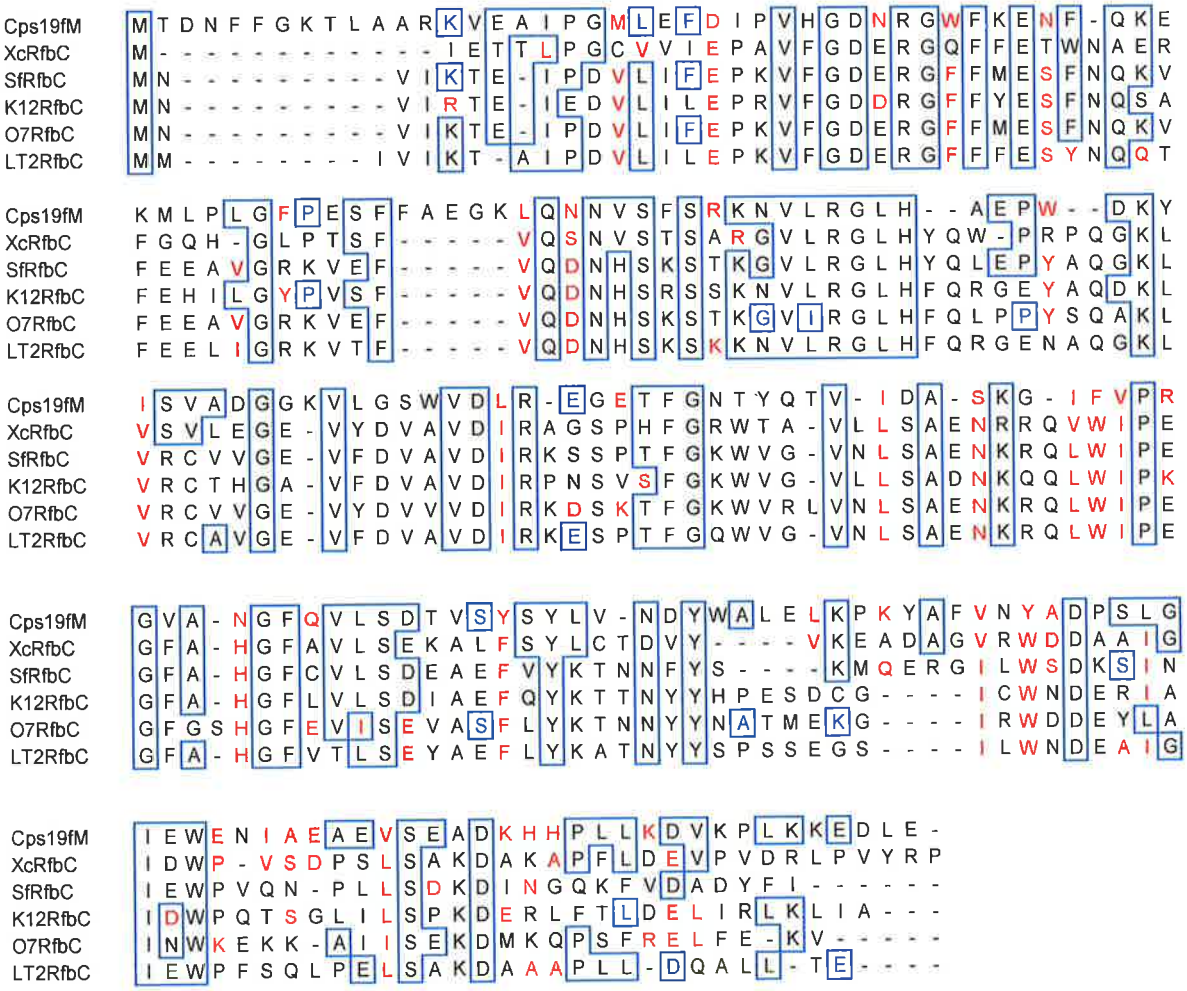


Fig. 3.9. Alignment of Cps19fM. Alignment of the amino acid sequences of Cps19fM, *X. campestris* RfbC (XcRfbC) (Koplin *et al.*, 1993), *S. flexneri* RfbC (SfRfbC) (Macpherson *et al.*, 1994), *E. coli* K12 RfbC (K12RfbC) (Stevenson *et al.*, 1994), *E. coli* O7 RfbC (O7RfbC) (Marolda and Valvano, 1995) and *S. enterica* serovar typhimurium LT2 RfbC (LT2RfbC) (Jiang *et al.*, 1991), as determined using the default settings of the program CLUSTAL (Higgins and Sharp, 1988). Residues identical to those in Cps19fM are boxed; similar residues are shown in red; - indicates absence of a residue.

Table 3.7. Similarity of Cps19fN to other proteins.

	% Identity[a]													
	Nm RfbB	Sy RfbB	Xc RfbB	Se Gdh	Bs SpsJ	Sv GraE	Ec RffG	LT2 RfbB	Sf RfbB	Stf TylB	K12 RfbB	Sg StrE	Ng RfbB	O7 RfbB
Cps19fN[b]	46.3	42	44.1	44.7	45.8	43.6	44.8	45.5	44	43.2	45.8	42.4	43.2	46.1
	[350]	[338]	[333]	[338]	[325]	[330]	[355]	[325]	[343]	[333]	[343]	[330]	[340]	[343]
NmRfbB[c]	100	57.1	60.5	46.1	48.3	42.2	75.5	71.1	69.3	45.2	71.9	45.8	91.9	71.3
		[343]	[349]	[343]	[333]	[339]	[351]	[352]	[352]	[343]	[352]	[345]	[345]	[352]
SyRfbB[d]		100	62.1	44.9	47.9	47	58.7	54.3	54	47.8	55.1	46.5	55.4	54.8
			[338]	[339]	[334]	[338]	[339]	[341]	[341]	[339]	[341]	[340]	[341]	[341]
XcRfbB[e]			100	46.9	49.5	45.8	58.9	56.7	56.6	47.2	56.7	45.3	57.9	57.9
				[331]	[331]	[336]	[348]	[349]	[328]	[335]	[349]	[333]	[335]	[349]
SeGdh[f]				100	49.7	57.2	44.3	41	35.7	66.1	44.7	61.9	44	44.4
					[314]	[318]	[348]	[284]	[350]	[327]	[284]	[323]	[341]	[284]
BsSpsJ[g]					100	45.9	47.6	47	45.2	46.5	44.6	44.5	47.9	45.9
						[314]	[334]	[270]	[330]	[314]	[336]	[317]	[333]	[270]
SvGraE[h]						100	45	27.8	43.5	57.9	41.9	60.9	43.8	43.8
							[342]	[224]	[283]	[318]	[344]	[317]	[336]	[283]
EcRffG[i]							100	72.3	70.6	42.2	72.3	43.1	73.2	71.8
								[354]	[354]	[348]	[354]	[341]	[329]	[354]
LT2RfbB[j]								100	85	42	87.8	43.5	69.1	87.5
									[361]	[281]	[361]	[271]	[346]	[361]
SfRfbB[k]									100	37.7	90.3	46.5	66.8	90.6
										[332]	[361]	[269]	[346]	[361]
StfTylB[l]										100	39.2	67.9	44.9	37.3
											[302]	[320]	[341]	[332]
K12RfbB[m]											100	33.1	68.8	91.1
												[311]	[346]	[361]
SgStrE[n]												100	45.3	34.1
													[340]	[311]
NgRfbB[o]													100	68.8
														[346]
O7RfbB[p]														100

a. Percentage of identical amino acids determined with FASTA as implemented in PROSIS. Numbers in parentheses indicate the number of amino acids over which the % identity occurs.
b. *S. pneumoniae* Cps19fN
c. *N. meningitidis* RfbB (Hammerschmidt *et al.*, 1994)
d. *Synechocystis* sp. RfbB (GenBank accession no. D64003)
e. *X. campestris* RfbB (Koplin *et al.*, 1993)
f. *Saccharopolyspora erythraea* Gdh (Linton *et al.*, 1995)
g. *B. subtilis* SpsJ (GenBank accession no. Z99123)
h. *S. violaceoruber* GraE (Bechthold *et al.*, 1995)
i. *E. coli* RffG (Daniels *et al.*, 1992)
j. *S. enterica* serovar typhimurium strain LT2 RfbB (Jiang *et al.*, 1991)
k. *S. flexneri* RfbB (Macpherson *et al.*, 1994)
l. *S. fradiae* TylB (Merson-Davies and Cundliffe, 1994)
m. *E. coli* K12 RfbB (Stevenson *et al.*, 1994)
n. *S. griseus* StrE (Distler *et al.*, 1987)
o. *N. gonorrhoeae* RfbB (Robertson *et al.*, 1994)
p. *E. coli* O7 RfbB (Marolda and Valvano, 1995)

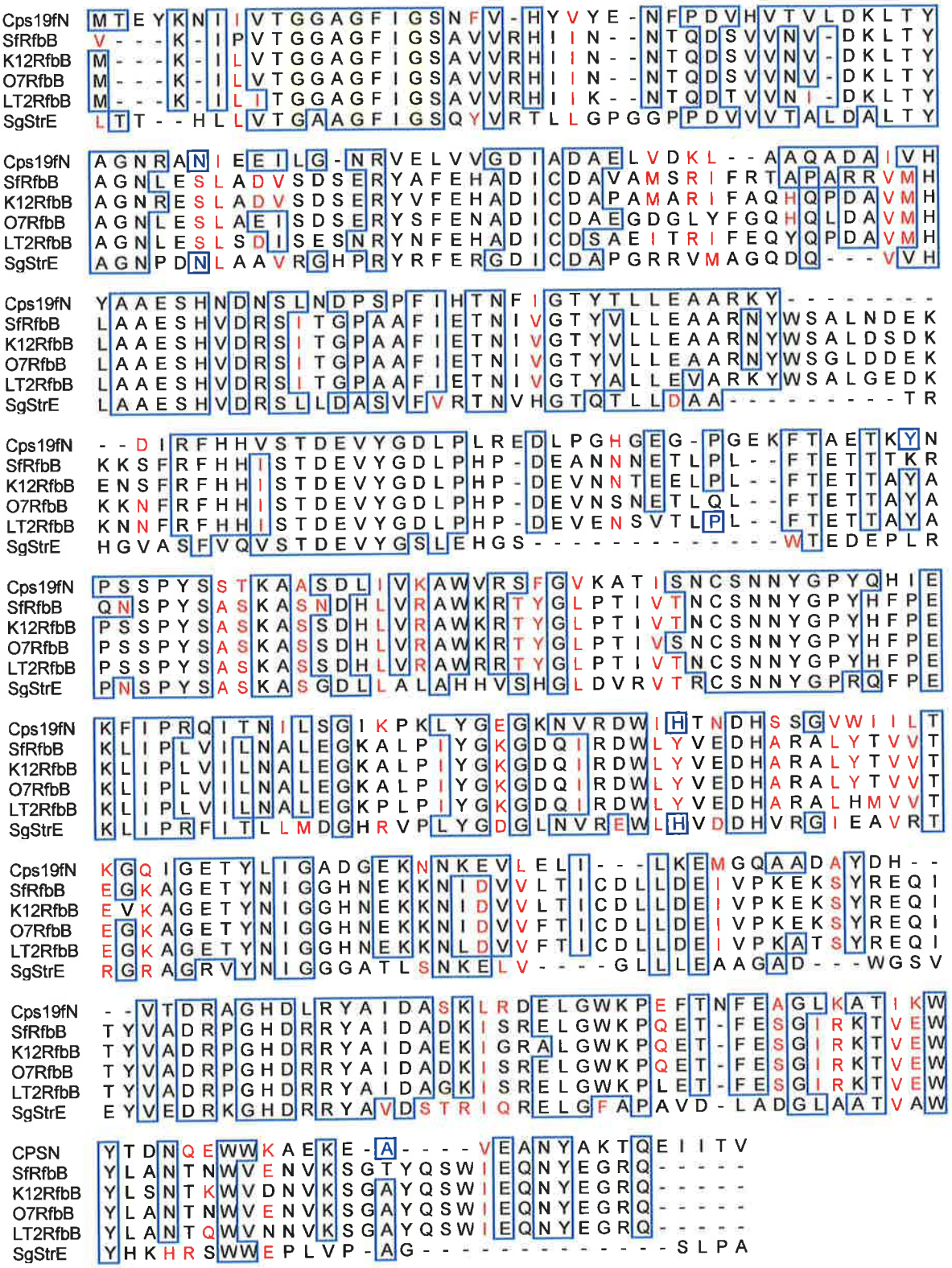


Fig. 3.10. Alignment of Cps19fN. Alignment of the amino acid sequences of Cps19fN, *S. flexneri* RfbB (SfRfbB) (Macpherson *et al.*, 1994), *E. coli* K12 RfbB (K12RfbB) (Stevenson *et al.*, 1994), *E. coli* O7 RfbB (O7RfbB) (Marolda and Valvano, 1995), *S. enterica* serovar typhimurium LT2 RfbB (LT2RfbB) (Jiang *et al.*, 1991) and *S. griseus* StrE (SgStrE) (Distler *et al.*, 1992), as determined using the default settings of the program CLUSTAL (Higgins and Sharp, 1988). Residues identical to those in Cps19fN are boxed; similar residues are shown in red; - indicates absence of a residue. The yellow shaded region corresponds to a NAD-binding motif (Macpherson *et al.*, 1994).

Table 3.8. Similarity of Cps19fO to other proteins.

	% Identity[a]								
	SgStrL	LT2RfbD	K12RfbD	BsSpsK	O7RfbD	XcRfbD	YeRfbG	SvGraD	SfRfbD
Cps19fO[b]	38.6	35	32.4	39.7	31.9	27.4	27	37.4	30.6
	[285]	[286]	[296]	[257]	[298]	[252]	[285]	[297]	[297]
SgStrL[c]	100	35.4	34.4	39.9	33.2	32	32.8	49.7	32.5
		[291]	[291]	[258]	[292]	[297]	[195]	[292]	[292]
LT2RfbD[d]		100	84.6	35.7	84.4	29.1	30.7	37.9	80.7
			[299]	[244]	[301]	[299]	[231]	[290]	[300]
K12RfbD[e]			100	35.2	91	39.5	31.9	35.7	87.7
				[273]	[301]	[299]	[229]	[294]	[300]
BsSpsK[f]				100	441	21.3	29.5	39.2	33.5
					[276]	[227]	[251]	[278]	[236]
O7RfbD[g]					100	37.9	30.2	34.8	87.7
						[301]	[232]	[296]	[302]
XcRfbD[h]						100	22.8	36.4	36.2
							[263]	[305]	[301]
YeRfbG[i]							100	29.6	30.4
								[206]	[230]
SvGraD								100	33
									[294]
SfRfbD[k]									100

a. Percentage of identical amino acids determined with FASTA as implemented in PROSIS. Numbers in parentheses indicate the number of amino acids over which the % identity occurs.
b. *S. pneumoniae* Cps19fO
c. *S. griseus* StrL (Distler *et al.*, 1987)
d. *S. enterica* serovar typhimurium LT2 RfbD (Jiang *et al.*, 1991)
e. *E. coli* K12 RfbD (Stevenson *et al.*, 1994)
f. *B. subtilis* SpsK (GenBank accession no. Z99123)
g. *E. coli* O7 RfbD (Marolda and Valvano, 1995)
h. *X. campestris* RfbD (Koplin *et al.*, 1993)
i. *Y. enterocolitica* RfbG (Zhang *et al.*, 1993)
j. *S. violaceoruber* GraD (Bechthold *et al.*, 1995)
k. *S. flexneri* RfbD (Macpherson *et al.*, 1994)

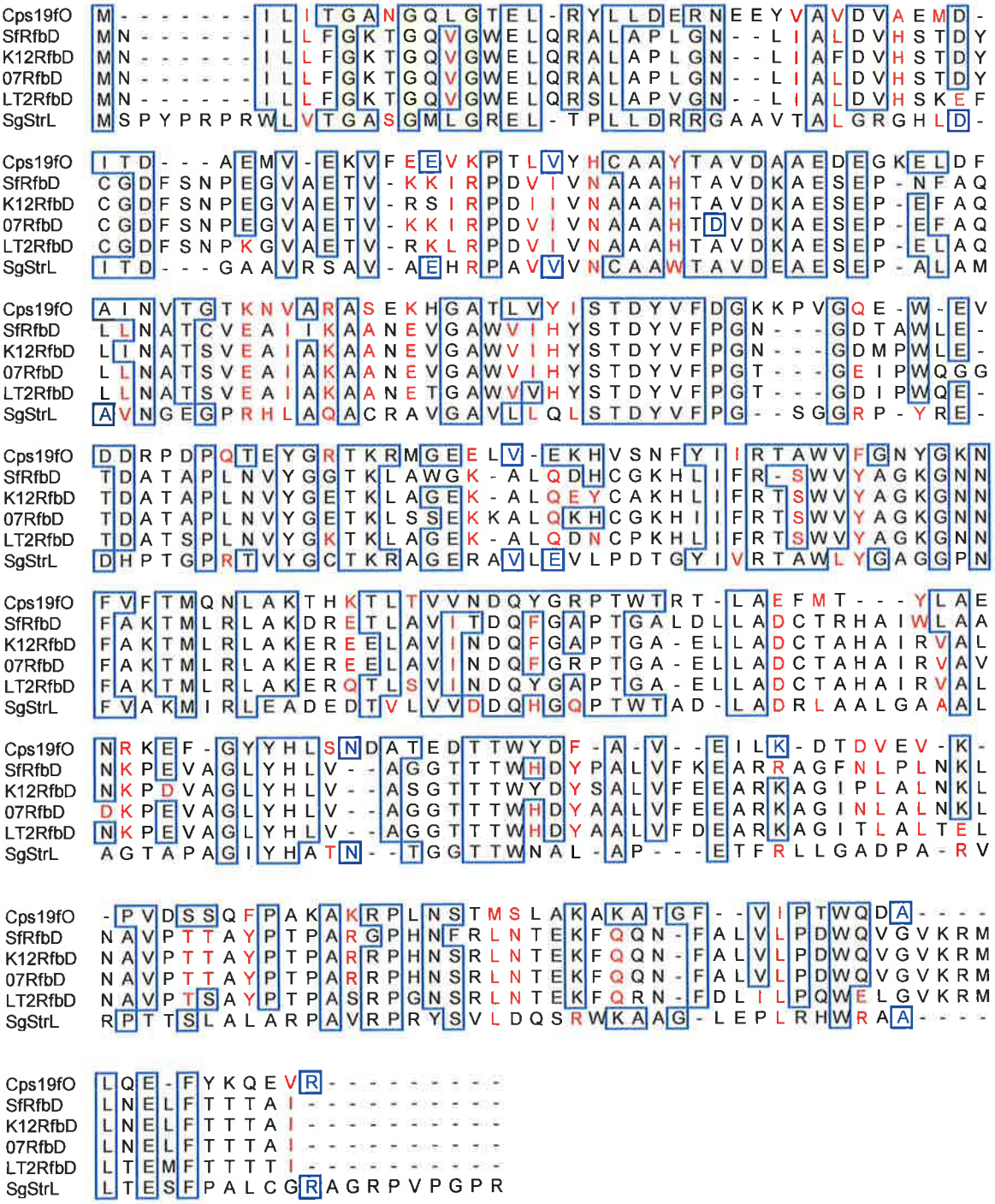


Fig. 3.11. Alignment of Cps19fO. Alignment of the amino acid sequences of Cps19fO, *S. flexneri* RfbD (SfRfbD) (Macpherson *et al.*, 1994), *E. coli* K12 RfbD (K12RfbD) (Stevenson *et al.*, 1994), *E. coli* O7 RfbD (O7RfbD) (Marolda and Valvano, 1995), *S. enterica* serovar *typhimurium* LT2 RfbD (LT2RfbD) (Jiang *et al.*, 1991) and *S. griseus* StrL (SgStrL) (Distler *et al.*, 1987), as determined using the default settings of the program CLUSTAL (Higgins and Sharp, 1988). Residues identical to those in Cps19fO are boxed; similar residues are shown in red; - indicates absence of a residue. The yellow shaded region corresponds to a NAD-binding motif (Macpherson *et al.*, 1994).

3.2.4 Insertion-duplication mutagenesis of *cps19fG-O* genes

In order to confirm the involvement of the ORFs designated *cps19fG* to *cps19fO* in type 19F CPS production, the chromosomal copies of the respective genes in the encapsulated strain Rx1-19F were individually disrupted by insertion-duplication mutagenesis (see section 2.10). To achieve this, a small internal segment of each ORF was obtained either by band isolation of an appropriate restriction fragment or by PCR amplification, and then cloned into the plasmid pVA891 (**Table 3.9**). Recombinant plasmids were then transformed into Rx1-19F (section 2.4.4). As described in section 2.10, Ery-resistant transformants are the result of a homologous recombination event directed by the cloned fragment of pneumococcal DNA that leads to the integration of the pVA891 plasmid into the host chromosome, such that the cloned fragment is duplicated, and consequent disruption of the gene of interest, as illustrated in **Fig. 2.1**.

Ery-resistant *S. pneumoniae* Rx1-19F transformants were obtained for all except the *cps19fM*- and *cps19fO*-containing pVA891 derivatives. As these two genes are part of the same biochemical pathway as *cps19fL* and *cps19fN*, no further attempts were made to disrupt these genes. Correct disruption of the respective chromosomal ORF was confirmed

Table 3.9. Summary of *cps19f* gene fragments inserted into pVA891

ORF	Nucleotides	Fragment cloned into pVA891
cps19fG	6,317-6,603	*Hin*dIII-*Sau*3A fragment from pJCP464
cps19fH	7,054-7,247	*Eco*RV-*Sau*3A fragment from pJCP464
cps19fI	8,179-8,440	*Hae*III-*Sau*3A fragment from pJCP465
cps19fJ	9,651-10,165	*Eco*RI- *Eco*RV fragment from pJCP466
cps19fK	10,708-11,177	PCR product using primers J17 and J18 (**Table 2.4**)
cps19fL	11,600-11,958	*Sma*I-*Bam*HI fragment from pJCP467
cps19fM	12,549-12,893	PCR product using primers J28 and J29 (**Table 2.4**)
cps19fN	13,148-13,961	PCR product using primers J30 and J31 (**Table 2.4**)
cps19fO	14,270-14,679	PCR product using primers J24 and J25 (**Table 2.4**)

by Southern hybridisation analysis of each insertion-duplication mutant generated (**Fig 3.12**). CPS gene-specific probes hybridised with two separate DNA fragments in the respective mutants but with only one fragment in the Rx1-19F control, indicating that the gene is disrupted in the mutants. As expected, the pVA891-specific probe hybridised only with the mutants. For Rx1-19F-J, Rx1-19F-L and Rx1-19F-N, the hybridisation pattern is different for the two probes because the enzyme used to restrict the chromosomal DNA was also used to clone the appropriate fragment into pVA891. In such cases, the interrupted *cps* gene and pVA891-related sequences will be on distinct chromosomal restriction fragments. The *cps19fG, cps19fH, cps19fI, cps19fJ, cps19fK, cps19fL* and *cps19fN* mutants all exhibited a rough phenotype and did not produce a type 19F capsule, as judged by quellung reaction, confirming that all are part of the *cps19f* locus. On the

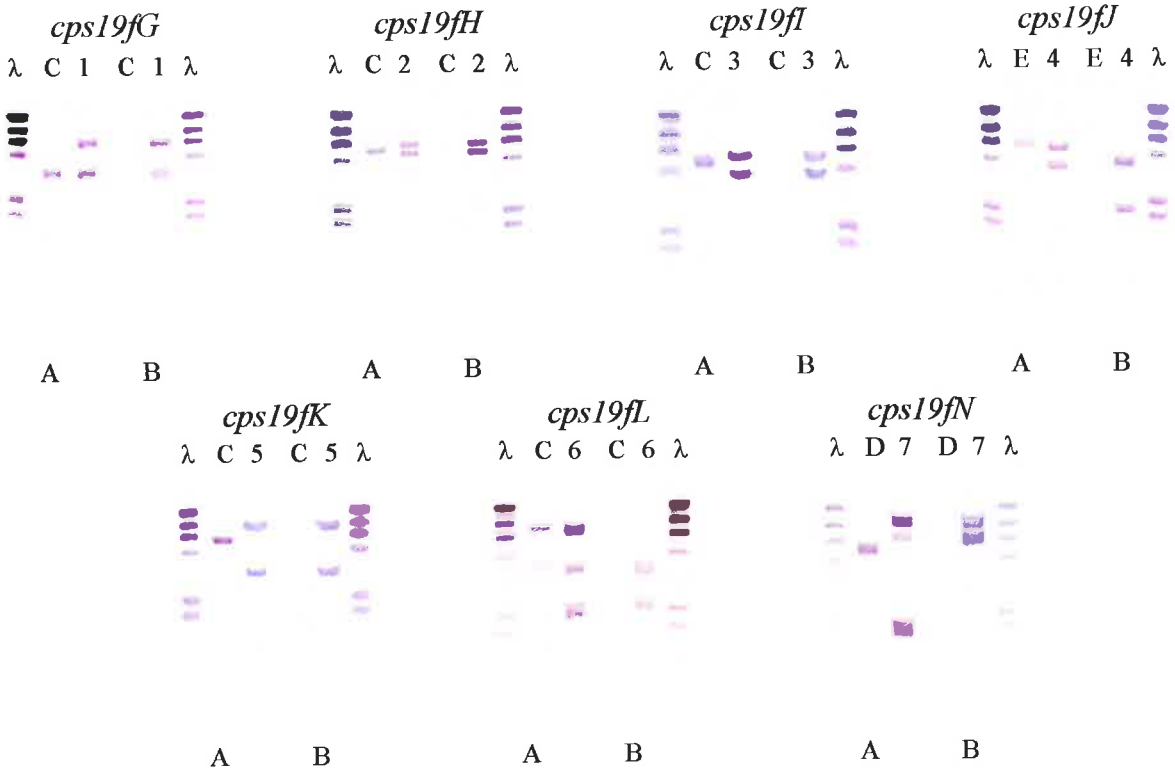


Fig. 3.12. Southern hybridisation analysis of insertion-duplication mutants. Southern blots of *Cla*I-restricted chromosomal DNA from Rx1-19F (C), Rx1-19F-G (1), Rx1-19F-H (2), Rx1-19F-I (3), Rx1-19F-K (5), and Rx1-19F-L (6); *Eco*RI-restricted chromosomal DNA from Rx1-19F (E) and Rx1-19F-J (4); *Bam*HI-restricted chromosomal DNA from Rx1-19F (D) and Rx1-19F-N (7) were probed with the gene specific probes indicated above each blot (**A**) and with a pVA891 specific probe (**B**). The sizes of the DIG-labelled lambda (λ) markers are as follows: 23 kb, 9.4 kb, 6.6 kb, 4.4 kb, 2.3 kb and 2.0 kb.

other hand, insertion of pVA891 into the apparent non-coding region downstream of *cps19fO* (nucleotides 15,149-15,444 were used as the target) did not interfere with type 19F capsule expression in *S. pneumoniae* Rx1-19F.

3.2.5 T7 expression of several Cps19f proteins

Expression of *cps19fH*, *cps19fI*, *cps19fJ*, *cps19fK*, *cps19fL*, *cps19fM cps19fN* and *cps19fO* cloned in *E. coli* DH5 (pGP1-2), in which the T7 RNA polymerase gene is under the control of lambda P_L/cI_{857} (Tabor and Richardson, 1985), was investigated. Expression of these genes from the vector T7 promoter was performed as described in section 2.12.1 and analysed on SDS-PAGE gels stained with Coomassie Brilliant Blue (section 2.12.3).

To examine the expression of *cps19fH* in *E. coli*, one of the nested deletion derivatives of pJCP464, containing *cps19fH* (nucleotides 6,578-7,846) downstream from the vector T7 promoter, was transformed into *E. coli* DH5(pGP1-2). Induction of cultures at 42°C for 30 min followed by 37°C for 2 hours (section 2.12.1) resulted in the expression of low levels of a 34-35 kDa protein in cells containing the pJCP464 derivative, which was not seen in uninduced cultures (see arrow, **Fig. 3.13A**). Both the size of the translation product and the low level of expression of *cps19fH* is consistent with initiation at the TTG codon as indicated in **Appendix I**.

The *cps19fI* and *cps19fJ* genes were PCR amplified using primer pairs J11/J14 and J15/J16, respectively (primers are described in **Table 2.4**) and cloned into pGEM-7Zf(+) such that the ORFs were downstream of the T7 promoter. These plasmids were transformed into *E. coli* DH5(pGP1-2). However, induction of cultures at 42°C did not result in increased production of any polypeptides with sizes corresponding to these ORFs, as judged by SDS-PAGE analysis (result not shown). However, this was not unexpected given previous observations that very hydrophobic proteins such as Cps19fI and Cps19fJ

are poorly expressed in this system (Morona *et al.*, 1994b).

To determine whether *cps19fK* was expressed in *E. coli*, a deletion derivative of pJCP466 (designated pJCP470), which placed nucleotides 10,304-11,720 downstream from the vector T7 promoter, was transformed into *E. coli* DH5(pGP1-2). Induction of the culture resulted in the expression of a polypeptide of approximately 41 kDa, which is consistent with the size of Cps19fK predicted from the DNA sequence (**Fig. 3.13B**).

Plasmids for T7 expression of *cps19fL, cps19fM, cps19fN* and *cps19fO* were constructed by cloning an appropriate PCR product into pBluescript SK+ or KS+ as shown in **Table 3.10** and transforming these constructs into *E. coli* DH5(pGP1-2). Induction of the genes under T7 control (section 2.12.1) resulted in the expression of polypeptides of approximately 32, 22 and 39 kDa for *cps19fL, M* and *N*, respectively, as predicted from the DNA sequence (**Fig. 3.13B**). However, no protein product with a size corresponding to that predicted for Cps19fO was detected when cells containing the *cps19fO* construct were induced (result not shown) and this was not investigated further.

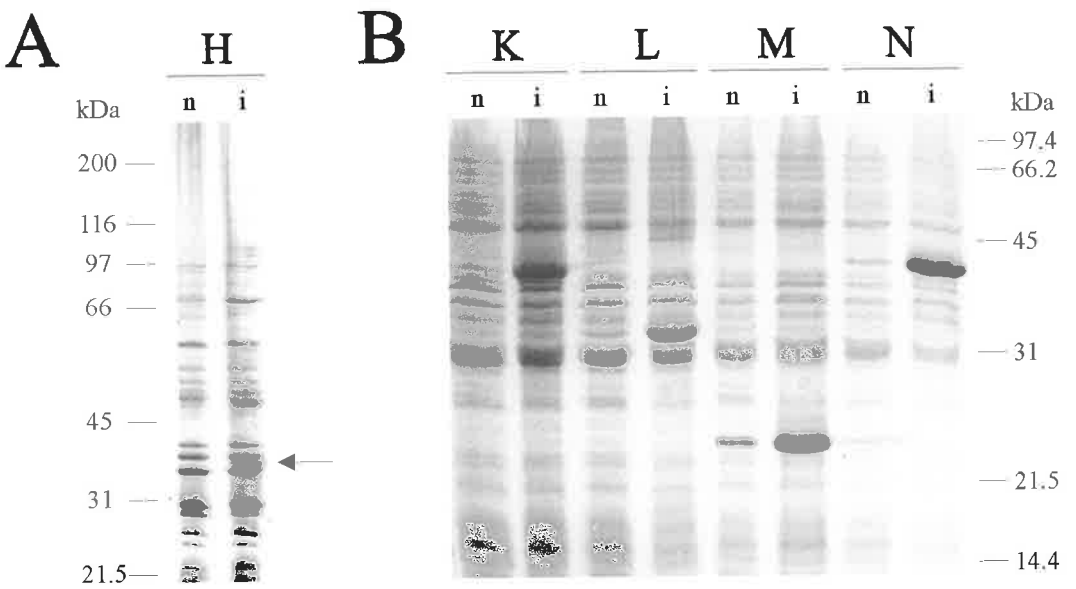


Fig. 3.13. T7-expression of *cps19f* genes in *E. coli*. *E. coli* DH5(pGP1-2) was transformed with various plasmids containing individual *cps19fH* (H), *cps19fK* (K), *cps19fL* (L), *cps19fM* (M) or *cps19fN* (N) ORFs. Lysates of cultures incubated at 42°C for 2 h to induce expression of T7 RNA polymerase) (**i**), and non-induced cultures (**n**) were separated by SDS-PAGE and stained with Coomassie Brilliant Blue. The mobilities of molecular size markers are shown separately for panels A and B.

Table 3.10. Plasmid constructs for T7 expression of *cps19fL-O*

Gene	Primers[a]	Restriction sites	*cps19f* sequence[b]	Vector
cps19fL	J21-J29	*Hin*dIII-*Eco*RV	11,347-12,487	pBluescript SK+
cps19fM	J12-J31	*Bam*HI-*Hin*dII	11,968-13,134	pBluescript KS+
cps19fN	J25-J28	*Hin*dIII-*Xba*I	12,547-14,129	pBluescript SK+
cps19fO	J37-J38	*Hin*dIII-*Bam*HI	14,009-15,562	pBluescript SK+

a. Primers sequences are as described in **Table 2.4**
b. The corresponding nucleotide numbers of the *cps19f* sequence (**Appendix I**)

3.2.6 Complementation of an *E. coli rffE* mutant with *cps19fK*

Whilst the precise mechanism is not understood, *E. coli rffE* (*nfrC*) mutants are known to be resistant to infection with bacteriophage N4 (Kiino *et al.*, 1993). In order to determine whether Cps19fK and *E. coli* RffE are functional homologues, the *rffE*-negative mutant *E. coli* KI8828 (Kiino *et al.*, 1993) was transformed with pJCP470 (which contains only the complete *cps19fK* ORF), or with pGEM-7Zf(+). The various strains, including the *rffE*-positive wild type parent *E. coli* MC4100 (Silhavy *et al.*, 1984), were then examined for susceptibility to infection with N4 phage (**Fig. 3.14**). Transformation with pJCP470, but not pGEM-7Zf(+), clearly conferred susceptibility to N4 on *E. coli* KI8828. Thus, Cps19fK is a functional RffE homologue and is therefore likely to be a UDP-Glc*N*Ac-2-epimerase.

3.2.7 Complementation of *S. flexneri rfbBDAC* with *cps19fLMNO*

As mentioned in section 3.2.3.6, database searches indicated that the *cps19fL-O* region had significant similarity to a portion of the *S. flexneri rfb* region (*rfbBDAC*) responsible for biosynthesis dTDP-Rha. To confirm functional homology, a portion of the *S. pneumoniae* Rx1-19F chromosome containing *cps19fL-O* (equivalent to nucleotides 11,351-15,449 in **Appendix I**) was PCR amplified using primers J21 and J36 (**Table 2.4**) and cloned into the *Hin*dIII and *Eco*RI sites of pK194. The recombinant plasmid

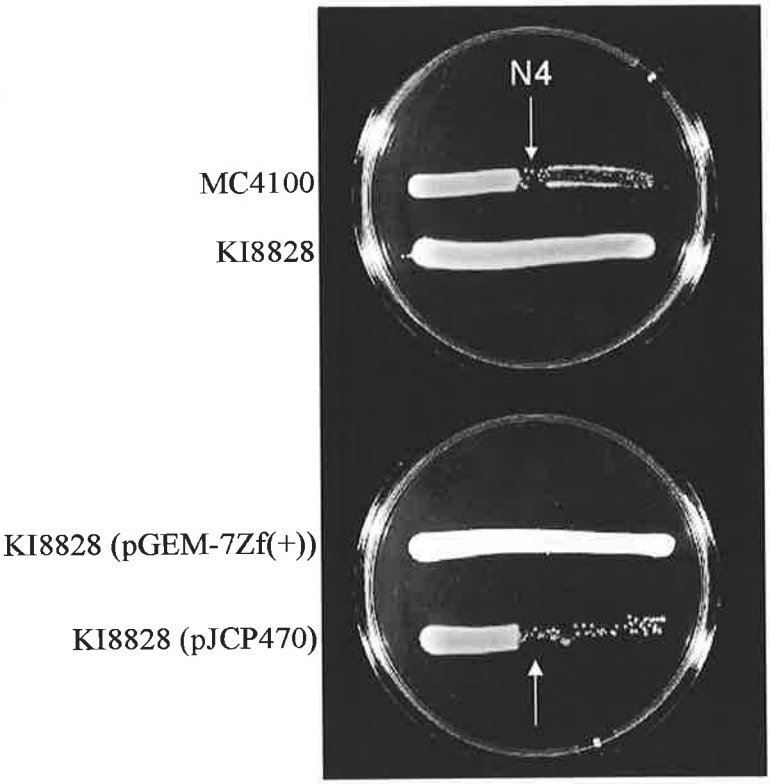


Fig. 3.14. N4 bacteriophage susceptibility. A suspension of bacteriophage N4 (10^9 pfu/ml) was streaked vertically on both plates as indicated by the arrows. The indicated *E. coli* strain was then streaked from left to right across the phage streak and incubated for 18 h at 37°C.

(designated pJCP471), or pK194, was then transformed into *E. coli* SØ874 containing pPM2716. The latter plasmid, from which *rfbBDAC* has been deleted, is a derivative of pPM2213 which contains the complete *S. flexneri rfb* region and directs the expression of *S. flexneri* serotype 4 O-antigen in *E. coli* K12 (Macpherson *et al.*, 1994). Lysates of *E. coli* SØ874 containing pPM2213, pPM2716, pPM2716 + pJCP470, or pPM2716 + pK194, were then subjected to Western blot analysis (section 2.12.4) using a rabbit antiserum raised against *S. flexneri* serotype 4 O-antigen (**Fig. 3.15**). Immunoreactive O-antigen can be seen in both the pPM2213 and the pPM2716 + pJCP470 tracks, indicating that *cps19fLMNO* can complement the *S. flexneri* serotype 4 *rfbBDAC* deletion in *E. coli*. Thus, Cps19fL is a Glc-1-phosphate thymidylyl transferase, Cps19fM is a dTDP-4-keto-6-deoxyglucose-3,5- epimerase, Cps19fN is a dTDP-Glc-4,6-dehydratase and Cps19fO is a dTDP-L-Rha synthase.

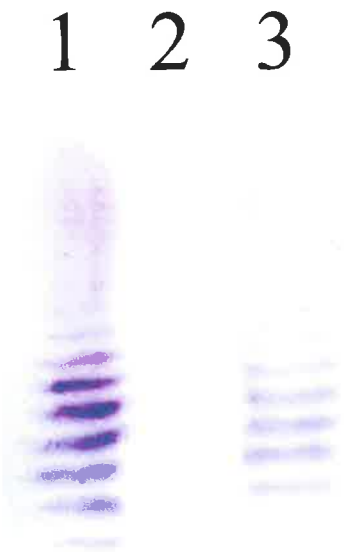


Fig. 3.15. Complementation of *S. flexneri rfbBDAC* by *cps19fLMNO*. *E. coli* lysates (prepared as described in section 2.12.2) were separated by SDS-PAGE, electroblotted onto nitrocellulose, probed with anti-*S. flexneri* type 4 serum, and developed as described in sections 2.12.3 and 2.12.4. Lanes: 1, *E. coli* SØ874 containing pPM2213; 2, *E. coli* SØ874 containing pPM2716; 3, *E. coli* SØ874 containing pPM2716 + pJCP470; 4, *E. coli* SØ874 containing pPM2716 + pK194.

3.2.8 Putative biosynthetic pathway for *S. pneumoniae* type 19F CPS

Functions have been previously proposed for the majority of the first six *cps19f* gene products on the basis of database comparisons (Guidolin *et al.*, 1994). Cps19fE is believed to catalyse the first step in CPS biosynthesis, namely transfer of Glc-1-phosphate to a lipid carrier (presumably und-P). Cps19fF is a putative Man*N*Ac transferase, while Cps19fC and Cps19fD are believed to be involved with chain length regulation and export of CPS. Collectively, this information and the findings of the current study indicate that the trisaccharide biological repeat unit of type 19F CPS is: →2)-α-L-Rha$_p$-(1-PO_4^-→4)-β-D-Man$_p$NAc-(1→4)-α-D-Glc$_p$-(1→, i.e. the first sugar in the repeat unit is Glc, not Rha as shown in previously published structures (Lee and Fraser, 1980; Ohno *et al.*, 1980). A putative biosynthetic pathway for type 19F CPS is shown in **Fig. 3.16**.

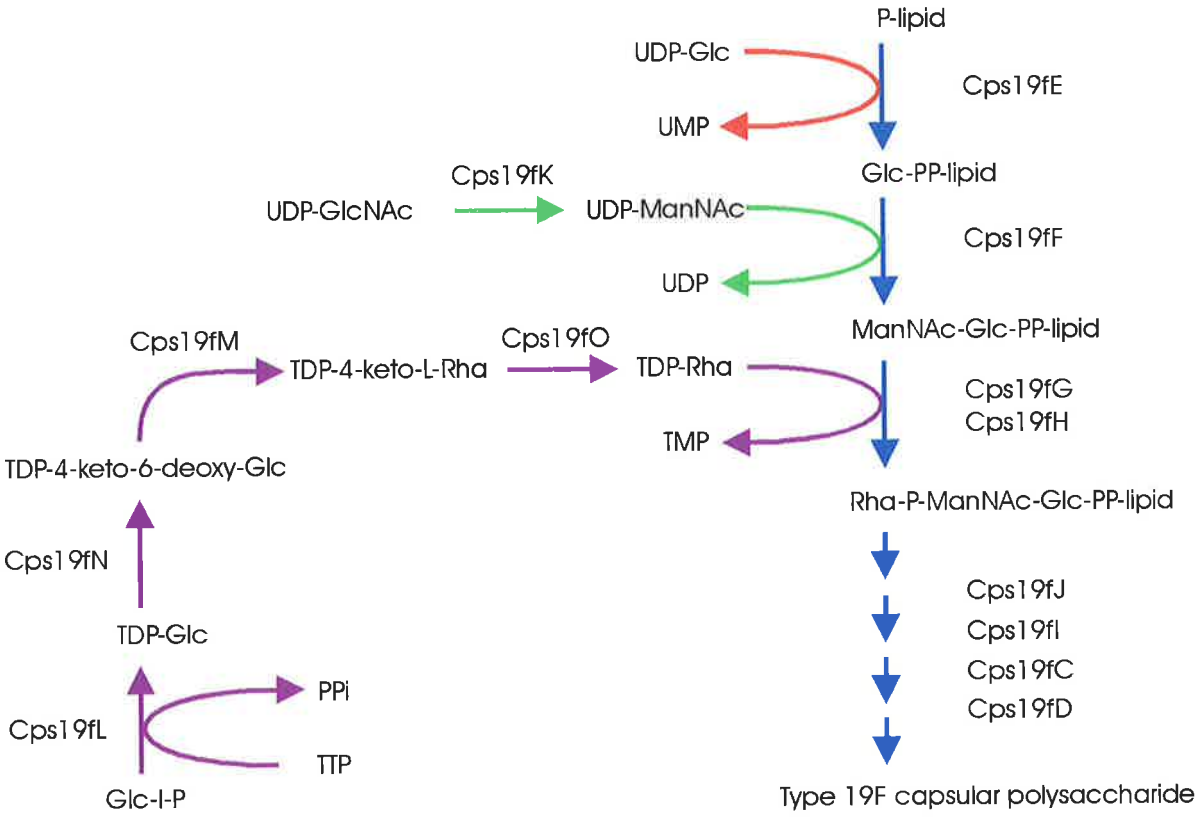


Fig. 3.16. Putative biosynthetic pathway for *S. pneumoniae* type 19F CPS.

Biosynthesis of type 19F CPS probably occurs via a mechanism analogous to that proposed for Rol/Cld- and Rfc-dependent O-antigen assembly in *Salmonella enterica* serogroups B and E (section 1.8.2; Whitfield, 1995). The initial step, catalysed by Cps19fE, involves transfer of Glc-1-phosphate from UDP-Glc to a lipid carrier on the cytoplasmic face of the cell membrane. Cps19fE has several large hydrophobic domains in its N-terminal portion, which would anchor it to the membrane and facilitate interaction with the lipid carrier (Guidolin *et al.*, 1994). Cps19fF, G and H then catalyse the sequential transfer of the other component monosaccharide precursors (synthesised in the cytoplasm by Cps19fK, L, M, N, and O) to form the trisaccharide repeat unit. These lipid-linked repeat units are then translocated from the cytoplasmic to the extracellular side of the cell membrane and polymerised in a blockwise fashion, extending the polysaccharide at the reducing terminus. These two steps are catalysed by Cps19fJ and I, respectively, both

of which are integral membrane proteins. The process of translocation and polymerisation may be closely linked and the two proteins possibly form a complex in the membrane.

In *Salmonella* and *Shigella* O-antigen assembly, Rol/Cld has been proposed to regulate chain length by modulating interaction between the lipid-linked nascent O-antigen and either the polymerase Rfc, or RfaL, a ligase responsible for transfer of O-antigen to the lipid A core molecule (Morona *et al.*, 1995; Whitfield, 1997). Cps19fC and D may perform a similar function in *S. pneumoniae* type 19F. Pneumococcal CPS is believed to be covalently linked to the cell wall peptidoglycan (Sørensen *et al.*, 1990), but the precise nature of this linkage and the enzyme responsible are unknown.

3.3 Conclusions

Sequential rounds of InPCR and plasmid insertion/rescue were used to isolate the region of the *S. pneumoniae* type 19F chromosome responsible for CPS biosynthesis. The data presented in this chapter, combined with that which was described previously (Morona *et al.*, 1994a; Guidolin *et al.*, 1994) indicates that the *cps19f* locus consists of 15 genes, which are tightly clustered on the chromosome. The *cps19f* locus is flanked by *dexB* and IS*1202* at the 5' end and by *aliA* at the 3' end. Arrecubieta *et al* (1995) have also shown that the *cps* locus of *S. pneumoniae* type 3 is flanked at the 5' end by *dexB*. Dillard *et al* (1995) have reported that the sequences flanking the 3' end of the *cps* locus of a different type 3 *S. pneumoniae* strain were homologous to *plpA*. Examination of the published nucleotide sequences of *plpA* (Pearce *et al.*, 1994) and *aliA* (Alloing *et al.*, 1994) indicates that these sequences describe the same gene and so the chromosomal location reported for the *cps3* locus is identical to that for *cps19f*. However, the *aliA* gene is not functional in *S.*

pneumoniae type 3, as it contains a large deletion at the 5' end of the gene (Dillard *et al.*, 1995).

Clues as to the likely function of the *cps19f* gene products have been provided by comparisons with known proteins whose sequences have been deposited with databases (as described above). Moreover, for *cps19fK*, *cps19fL*, *cps19fM*, *cps19fN* and *cps19fO*, the function of gene products has been confirmed by complementation of mutations in *E. coli*. This information has been used to propose a biosynthetic pathway for type 19F CPS. However, experimental confirmation of the function of the remaining proteins encoded by *cps19f* will require characterisation of the phenotypic impact of mutagenesis of the respective ORFs. Interpretation of phenotypic data obtained with the insertion-duplication mutants generated in the present study and previously (Guidolin *et al.*, 1994) is complicated by the likelihood of polar effects. Hence, further phenotypic analysis of the function of individual *cps19f* genes awaits the construction of in-frame deletion mutants, in which transcription of the remainder of the type 19F *cps* locus would not be expected to be affected.

Chapter 4

ANALYSIS OF CAPSULE LOCI FROM VARIOUS PNEUMOCOCCAL SEROTYPES

4.1 Introduction

Analysis of the *cps19f* locus identified genes encoding a range of functions, including those involved in biosynthesis of activated sugar precursors, various glycosyl transferases, a polysaccharide polymerase, and proteins possibly involved in export functions. Some of these specific functions might be expected to be required by all pneumococci, regardless of capsular serotype; others might be expected to be present in various subsets of capsular types (depending on which sugars are present in their repeat units), and some genes might be expected to be serotype-specific. To examine the specificity of individual *cps19f* genes, their distribution amongst diverse *S. pneumoniae* serotypes was examined by Southern hybridisation analysis.

This provided information on the organisation of pneumococcal CPS loci which was used to develop a LR-PCR protocol for isolation of large fragments of the *cps* loci from other *S. pneumoniae* serotypes.

4.2 Results

4.2.1 Serotype specificity of the *cps19f* genes and flanking DNA sequences

Southern hybridisation analysis (described in section 2.7.2) to determine the serotype specificity of the individual *cps19f* genes was undertaken. DNA fragments corresponding to each of the *cps19f* genes and the flanking regions were isolated from appropriate clones either by restriction enzyme digestion and subsequent purification after agarose gel electrophoresis, or by PCR amplification using specific primers. They were then labelled with DIG (section 2.7.1) and used to probe *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, 22, 23F, 24, and the other members of serogroup 19 (19A, 19B and 19C). Hybridisation and washing were performed under high stringency conditions (as described in section 2.7.2.2). The results of these 76 Southern hybridisation reactions are summarised in **Table 4.1**.

Large variations in the hybridisation patterns were obtained with the different gene-specific probes. Probes specific for sequences flanking *cps19f* (*dexB*, the 5' intergenic region, the 3' intergenic region and *aliA*) hybridised with all serotypes tested (see section 4.2.1.3). However, of the genes within the *cps* locus, high stringency homologues of only *cps19fA* and *cps19fB* were present in all serotypes tested.

4.2.1.1 Presence of *cps19f* homologues in other members of serogroup 19

Serogroup 19 consists of four immunologically cross-reactive serotypes 19F, 19A, 19B and 19C. The results in **Table 4.1** suggest that the *cps* loci of types 19B and 19C are very closely related to 19F. They hybridised with all the *cps19f* probes except *cps19fI* and *cps19fJ*. Hybridisation with probes specific for the two genes flanking *cps19fI* and *cps19fJ* (*cps19fH* and *cps19fK*) was weaker suggesting a lesser degree of similarity in this region.

Table 4.1. Hybridization of type 19F *cps* genes and neighbouring sequences with chromosomal DNA from other pneumococcal serotypes.

	DIG labelled DNA probes[a]																		Presence of sugar in capsule[b]		
			cps19fA-O																		
Type/Group	*dexB*	IG 5'	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	IG 3'	*aliA*	Glu	ManNAc	Rha
19F	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	*	*	*
19A	+	+	+	+	-	-	-	-	±	±	-	±	-	-	+	+	+	+	+	*	*	*
19B	+	+	+	+	+	+	+	+	+	±	-	-	±	+	+	+	+	+	+	*	*	*
19C	+	+	+	+	+	+	+	+	+	±	-	-	±	+	+	+	+	+	+	*	*	*
2	+	+	+	+	-	-	-	-	-	-	-	-	-	±	+	+	+	+	+	*		*
3	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	±	+	*		
4	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	+	+		*	
6A	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	*		*
6B	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	*		*
7F	+	+	+	+	+	+	+	-	-	-	-	-	-	+	+	+	+	+	+	*		*
8	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	*		
9N	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	*	*	
9V	+	+	+	+	-	-	-	+	-	-	-	-	-	-	-	-	-	+	+	*	*	
12	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	*	*	
14	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	+	+	*		
16	+	+	+	+	+	+	+	-	-	-	-	-	-	+	+	+	+	+	+	*		*
17	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	*		*
18C	+	+	+	+	+	+	+	-	-	-	-	-	-	+	+	+	+	+	+	*		*
22	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	*		*
23F	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	*		*
24	+	+	+	+	+	+	+	-	-	-	-	-	-	+	+	+	+	+	+	*		*

a. The following DNA fragments were labelled with DIG and used as probes at high stringency: 0.5 kb *Nco* I-*Eco* RI restriction fragment for *dexB* and 0.9 kb *Cla* I-*Nco* I restriction fragment for the 5' intergenic region (IG 5') from previously published sequence (Morona *et al*., 1994); nucleotides 336-1,468, 1,571-2,380, 2,380-2,998, 3,126-3,739, 3,682-4,979, 5,225-5,725, 6,015-6,630, 6,674-7,731, 7,789-8,965, 9,013-10,278, 10,530-11,539, 11,539-12,493, 12,456-13,139, 13,139-14,134 and 14,134-14,955 for *cpsA-O* genes, respectively; nucleotides 14,955-15,449 for the 3' intergenic region (IG 3') and a 1.2 kb *Hin* dIII restriction fragment for *aliA*. Strong, weak and no hybridization are indicated by +, ± and - respectively.

b. The presence of the sugars glucose (Glu), *N*-acetyl mannosamine (Man*N*Ac) and rhamnose (Rha) in the capsule of each pneumococcal type is indicated by an asterisk.

However, type 19A did not appear to have sequences capable of high-stringency hybridisation with the probes specific for *cps19fC*, *cps19fD*, *cps19fE*, *cps19fF*, *cps19fI*, *cps19fK*, and *cps19fL*, and hybridised only weakly with probes specific for *cps19fG*, *cps19fH* and *cps19fJ*. The apparent dissimilarity of the type 19A and type 19F *cps* loci is intriguing given the structural similarity between the two polysaccharides as shown in **Table 1.3**. To further examine the relationship between *cps19f* and *cps19a*, low stringency Southern hybridisation (section 2.7.2.3) of type 19A chromosomal DNA with the *cps19fK* and *cps19fL* gene probes was undertaken. Under these conditions, both these probes hybridised with type 19A indicating that 19A contained homologues of both *cps19fK* and *cps19fL*.

4.2.1.2 Presence of *cps19f* homologues in other serotypes

Outside of serogroup 19, types 7F, 16, 18C and 24 were the most similar to type 19F, hybridising at high stringency to 9 of the 15 *cps19f* gene probes. These four serotypes all had sequences closely related to *cps19fA-E* and *cps19fL-O*, as shown in **Table 4.1**. The least similar were types 8 and 12, which hybridised only with *cps19fA* and *cps19fB*. Sequences closely related to *cps19fL*, *cps19fM*, *cps19fN* and *cps19fO*, the genes required for dTDP-L-Rha biosynthesis (section 3.2.3.6), were detected in all serotypes tested whose CPS contains L-Rha.

The CPSs of four of the serotypes that were tested contain Man*N*Ac (types 4, 9N, 9V and 12). However, only the two group 9 serotypes contained a gene hybridising to *cps19fF*, which encodes the putative UDP-Man*N*Ac transferase (section 1.9.3). This is consistent with the fact that only group 9 CPS has the same β-D-Man$_p$NAc-(1→4)-α-D-Glc$_p$ linkage seen in group 19 CPS. It was surprising, however, that none of these serotypes contained high or low stringency homologues to *cps19fK*, which encodes a UDP-Glc*N*Ac-2-epimerase (section 3.2.3.5). This enzyme activity is likely to be essential for

synthesis of UDP-Man*N*Ac, and is therefore presumably encoded by a divergent gene in the other three serotypes. Interestingly, the sequence of the genome of *S. pneumoniae* type 4 became available during the course of this project (available from the TIGR Microbial Database [http://www.tigr.org/pub/ data/s_pneumoniae/]). Examination of the *cps4* locus revealed the presence of a gene (designated *cps4L*, see section 8.3.4) which has 65.9% identity to *cps19fK* and presumably encodes the UDP-Glc*N*Ac-2-epimerase required for type 4 CPS biosynthesis.

Of the various serotypes tested, the CPS of all but type 4 contains Glc (**Table 4.1**). However, types 2, 3, 6A, 6B, 8, 9V, 12, 17, 19A, 22 and 23F also lack sequences which hybridise to *cps19fE*, encoding the type 19F putative Glc-1-phosphate transferase (section 1.9.3). It is possible that Glc may not be the start of the biological repeat unit in some or all of these serotypes. However, glucosyl transferase activity which adds Glc-1-phosphate to a lipid carrier has recently been observed in serotypes 2, 6A, 9V, 12F, 17F, 22F and 19A (Kolkman *et al.*, 1997a), suggesting the presence of a divergent or unrelated *cpsE* gene in these serotypes. As expected, both serotypes 3 and 4 lacked Glc-1-phosphate transferase activity (Kolkman *et al.*, 1997a). Type 4 CPS lacks Glc. The type 3 CPS does contain this sugar, but the capsule is synthesised by a single processive transferase (Dillard *et al.*, 1995) and so does not require a *cps19fE* homologue, as will be discussed in detail in chapter 8.

4.2.1.3 Analysis of the intergenic regions

Southern hybridisation analysis with probes specific for *dexB*, *cps19fA-O* and *aliA* were very specific with the individual probes hybridising with one, or at the most two, DNA fragments for each serotype. This is illustrated in **Fig. 4.1A** in which restricted chromosomal DNA from various *S. pneumoniae* serotypes was probed with *cps19fA*. However, the probes specific for either the 5' or the 3' intergenic regions hybridised with multiple DNA fragments (**Fig. 4.1B** and **C**). Database searches with these intergenic

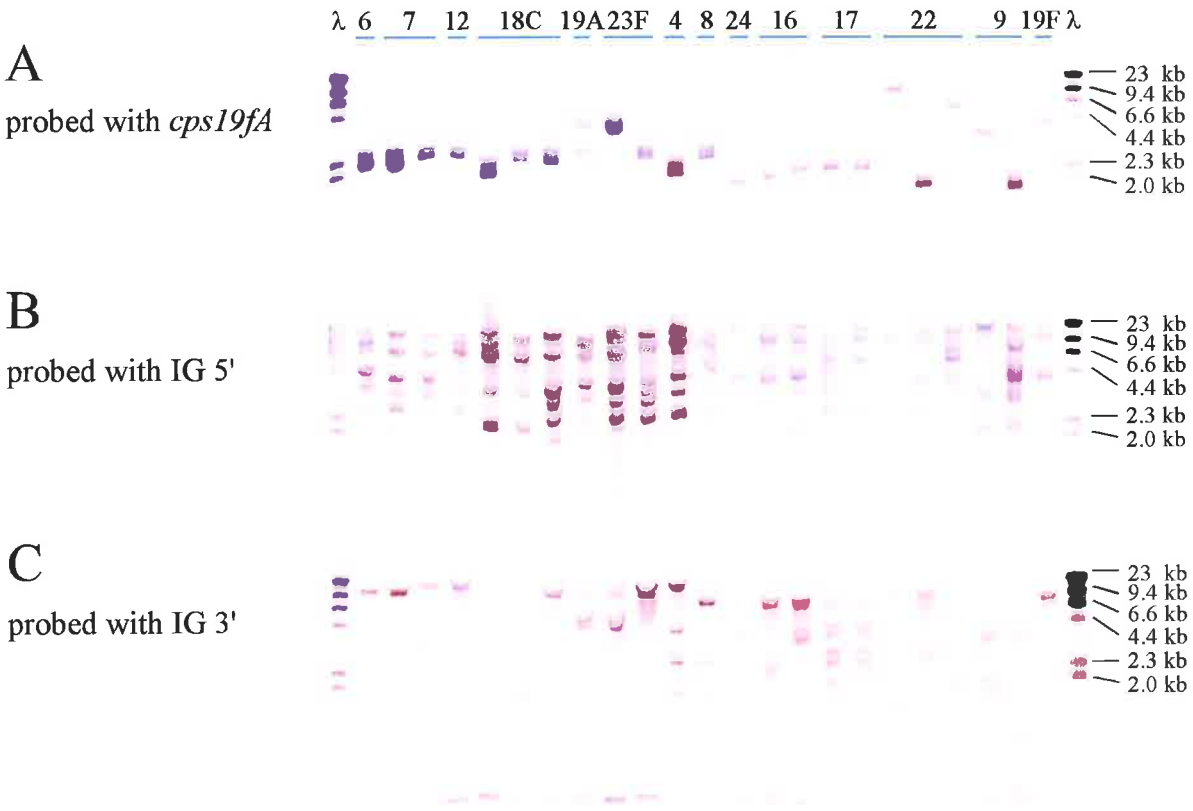


Fig. 4.1 Southern hybridisation of pneumococcal chromosomal DNA. Southern blots of *Cla*I-restricted chromosomal DNA of the indicated *S. pneumoniae* serotypes were probed with DIG-labelled probes specific for *cps19fA* (A), IG 5' (B) and IG 3' (C). The molecular size standards are shown on the right-hand side of the figure and correspond to DIG-labelled *Hin*dIII-digested λ phage DNA.

regions against the TIGR *S. pneumoniae* type 4 genomic DNA sequence showed that sections of these regions were, in fact, repeated. A region of 500 nucleotides in the 5' intergenic region, which lies 300 nucleotides downstream from the 3' end of *dexB*, is repeated six times and an adjoining region of 100 nucleotides is repeated at least 20 times. Large sections of the 3' intergenic region are also repeated at least six times in the partially sequenced *S. pneumoniae* type 4 chromosome.

4.2.2 Amplification of capsule loci by LR-PCR

Southern hybridisation analysis indicated that high stringency homologues of *dexB*,

cps19fA, *cps19fB* and *aliA* are present in all the pneumococcal serotypes investigated. This suggested that, at least in theory, complete capsule loci of these serotypes could be amplified using LR-PCR with primers based on sequences in *dexB* and *aliA*. However, the distance between the 3' end of *dexB* and the 5' end of *aliA* in type 19F is large (19.1 kb), and the *cps19fA-O* operon itself is 14.8 kb in size (**Fig. 4.2**). Moreover, the expected size for many *cps* loci would be even larger than that for *cps19f* as their CPS repeat units contain up to 7 sugars, compared with 3 for type 19F (van Dam *et al.*, 1990), and so would require extra genes to encode additional biosynthetic enzymes. Thus the *dexB-aliA* region of these types would be expected to exceed 20-25 kb, and the efficiency of LR-PCR amplification might be low. To increase the likelihood of successfully amplifying LR-PCR products from larger *cps* loci the 5' primer (J39, described in **Table 2.4**) was based on the sequence at the 3' end of *cps19fB* reducing the size of the potential LR-PCR products by up to 5 kb (the distance between the 3' end of *dexB* and the 3' end of *cps19fB*, **Fig. 4.2**). Such smaller LR-PCR products are sufficient to characterise other *cps* loci as they would contain all of the type-specific portions of the loci, missing only the highly conserved *cpsA* and *B* genes. The other primer (J36, described in **Table 2.4**) was based on sequence at the 5' end of *aliA* (**Fig. 4.2**) as there are no genes which are conserved in all serotypes at the 3' end of the *cps* locus. The primers were not based on the apparently conserved 3' intergenic region because of the theoretically greater possibility of minor sequence divergence in non-coding DNA. LR-PCR amplification, using these primers was undertaken as described in section 2.9.2.

All pneumococcal serotypes tested in **Table 4.1**, except type 3 (the 5' end of the *aliA* gene has been deleted in all type 3 strains tested [Caimano *et al.*, 1998]), and a type 20 isolate, were used as templates for LR-PCR. The PCR reactions were analysed by electrophoresis on a 1% agarose gel. 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

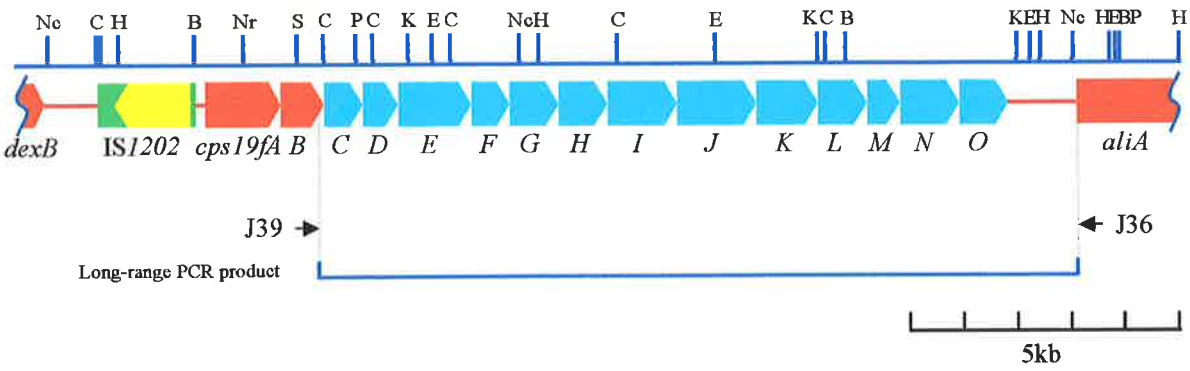


Fig. 4.2. The *dexB-aliA* region of the type 19F chromosome. The conserved regions including those within the *cps19f* locus are indicated in red. The position of the two primers used for long-range PCR, J39 and J36 (**Table 2.4**), are indicated with arrows.

not from types 7F, 9V, 12, 16, 17, 19C, 22 and 24. Analysis of the DNA fragments reveals that the LR-PCR products ranged in size from 15-20 kb as shown in **Fig. 4.3A**. The PCR products were also digested with *Cla*I and electrophoresed on a 1% agarose gel (**Fig. 4.3B**). 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 LR-PCR products from different type 19F strains (**Fig. 4.3B**).

4.2.3 Southern hybridisation analysis of LR-PCR products

In order to confirm that they contained *cps*-related sequences, the LR-PCR products from the various *S. pneumoniae* serotypes were restricted with *Cla*I, and subjected to Southern hybridisation analysis using probes specific for *cps19fC* and *cps19fL* (**Fig. 4.3C** and **D**).

The *cps19fC* probe hybridised at high stringency with a 0.9-kb DNA fragment in types 4, 9N, 14, 18C, 19F, 19B and 20. Both the hybridisation pattern and the size of the DNA fragment which hybridised with the *cps19fC* probe are consistent with the Southern hybridisation data obtained when probing *Cla*I-restricted chromosomal DNA with the *cps19fC* probe (**Table 4.2**). Chromosomal hybridisation data were not available for the

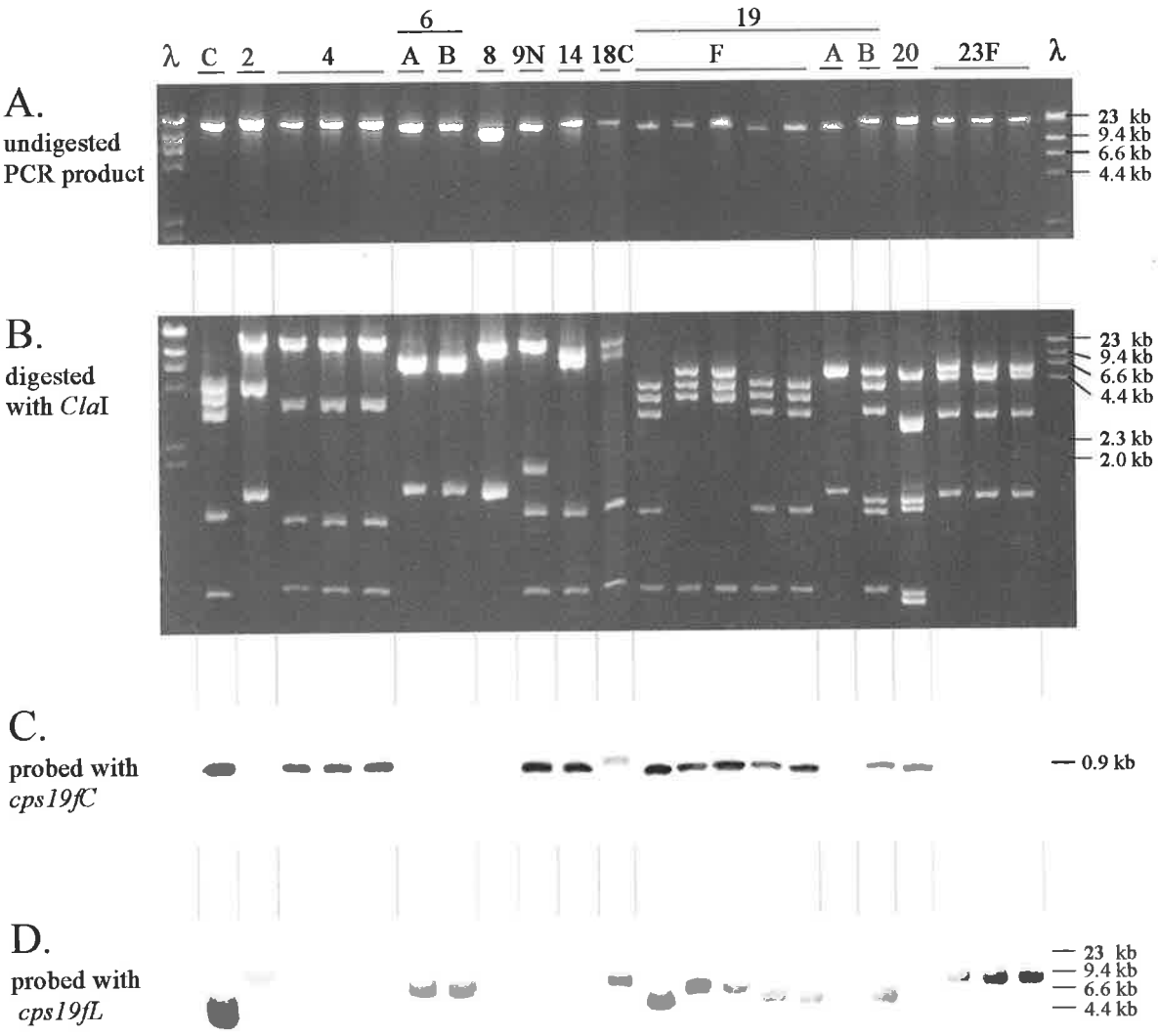


Fig. 4.3. LR-PCR products. PCR product, undigested (**A**), or digested with endonuclease *Cla*I (**B**), was electrophoresed on a 1% agarose gel in the presence of ethidium bromide. *Cla*I restricted PCR product was subjected to Southern hybridisation analysis using DIG-labelled probes specific for *cps19fC* (**C**) or *cps19fL* (**D**). The molecular size standards are shown on the right-hand side of the figure and correspond to *Hin*dIII-digested λ phage DNA.

type 20 isolate, from which the LR-PCR product was amplified, as it had not been included in the previous Southern hybridisation study (**Table 4.1**).

The *cps19fL* probe hybridised with DNA fragments ranging in size from 6 to 10 kb in the *Cla*I-restricted PCR products from types 2, 6A, 6B, 18C, 19F, 19B and 23F. Hybridisation was consistent with that obtained from Southern hybridisation with *Cla*I-restricted chromosomal DNA from these isolates, although, the sizes of the restriction fragments differ (**Table 4.2**). 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.

As expected, neither the *cps19fC* nor the *cps19fL* probes hybridised at high stringency with the LR-PCR products obtained from types 8 and 19A. Previous Southern hybridisation data (**Table 4.1**) indicated that *S. pneumoniae* types 8 and 19A do not contain DNA sequences highly homologous to *cps19fC* or *cps19fL*.

Table 4.2. Sizes of *Cla*I-restricted fragments which hybridise to *cps19fC* and *cps19fL* probes.

	Sizes[a] of *Cla*I-restricted fragments which hybridise to:			
	cps19fC[b]		*cps19fL*[b]	
Serotype	LR-PCR product	Chrom. DNA	LR-PCR product	Chrom. DNA
2	-	-	10 kb	15 kb
4	0.9 kb	0.9 kb	-	-
6A	-	-	8 kb	10 kb
6B	-	-	8 kb	10 kb
8	-	-	-	-
9N	0.9 kb	0.9 kb	-	-
14	0.9 kb	0.9 kb	-	-
18C	0.9 kb	0.9 kb	10 kb	15 kb
19F	0.9 kb	0.9 kb	6-9 kb	10-12 kb
19A	-	-	-	-
19B	0.9 kb	0.9 kb	6 kb	10 kb
20	0.9 kb	NA[c]	-	NA
23F	-	-	10 kb	10 kb

a. Size is approximate.
b. DIG labelled probes specific for *cps19fC* and *cps19fL* as described in **Table 4.1**.
c. Not analysed.

4.2.4 Partial DNA sequencing of the LR-PCR products

As additional confirmation that the LR-PCR products contain *cps* related genes, they were subjected to one round of sequence analysis using dye-terminator sequencing reactions with the J39 primer (located at the 5' end of the PCR product). This provided sequence data for the 3' end of *cpsB* and the 5' portion of *cpsC* for all the LR-PCR products except type 18C. No sequence data were obtained using this latter template, presumably due to the low yield of the PCR product obtained. Analysis of the various sequences and those published for types 1, 3 and 14 (Muñoz *et al.*, 1997; Arrecubieta *et al.*, 1995; Kolkman *et al.*, 1996) showed that there are two distinct *cpsC* genes in these loci, designated class I and class II. Types 1, 3, 9N, 14, 19F, 19B and 20 have class I *cpsC* genes which exhibit >95% identity to *cps19fC*, whereas types 2, 6A, 6B, 8, 19A and 23F have class II *cpsC* genes which exhibit 72-74% identity to *cps19fC*, but >95% identity to each other (**Fig. 4.4**). **Table 4.3** shows the sequence similarity between the various *cpsC* genes. The sequences obtained from the LR-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. 4.4**).

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 cross-over point between nucleotides 345 and 354 of the type 4 sequence (**Fig. 4.4**). Comparison of the TIGR type 4 sequence data (section 8.3.4) with the 19F *cps* sequence showed another point of divergence within the *cpsB* gene. The comparison of the sequences of *cps19fB* and *C*, and *cps4B* and *C*, is shown in **Fig. 4.5**. A region of 852 nucleotides, including most of *cpsB* and part of *cpsC*, shares approximately 74% similarity between *cps19f* and *cps4*, whereas flanking regions exhibit greater than 95% identity. This may have arisen as a consequence of recombination between a class I *cps* locus and a DNA fragment

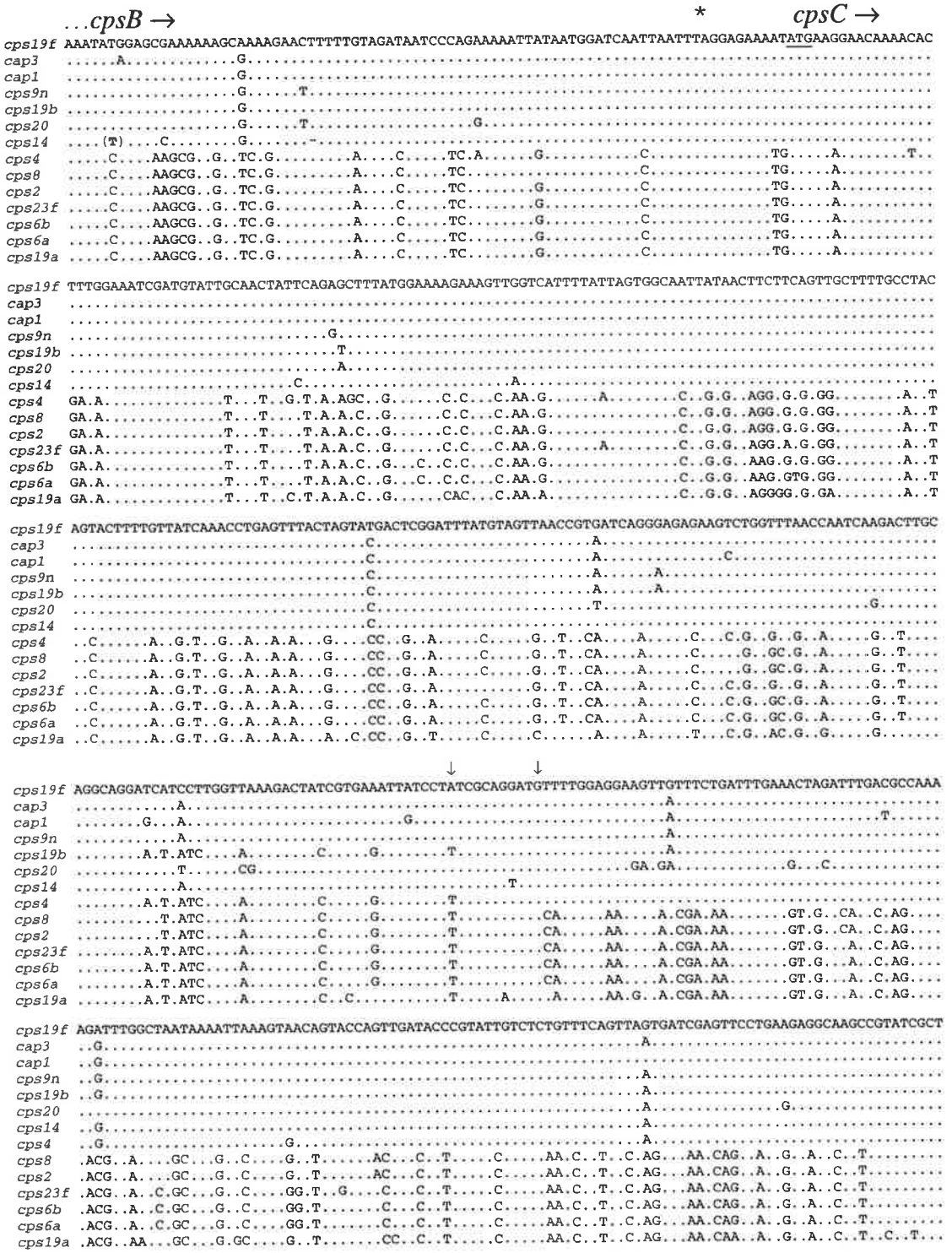


Fig. 4.4. Comparison of class I and class II *cps* sequences. The 500 nucleotides of sequence shown (100 nucleotides per line) corresponds to nucleotides 2,273-2,772 of the *cps19f* sequence. The following sequences are available under the GenBank accession numbers as indicated: *cps19f*, U09239; *cap3*, Z47210; *cap1*, Z83335; *cps14*, X85785; *cps23f*, AF030373; *cps19a*, AF094575. The *cps4* sequence is available from the TIGR Microbial database. 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 arrows indicate the region where the cross-over between class I and class II sequences has occurred in *cps4*.

(approximately 852 nucleotides long) from a class II *cps* locus, resulting in a mosaic *cpsB-C* region. Analysis of the available type 23F sequence data (Coffey *et al.*, 1998a) indicated that the *cps23f* locus (which has a class II *cpsC* gene) also diverges from *cps19f* (which has a class I *cpsC* gene) within the *cpsB* gene, but 98 nucleotides further downstream from the point of divergence for *cps4*. This suggests that the point of sequence divergence from class I to class II within *cpsB* may vary between different serotypes.

Table 4.3. Similarity between *cpsC* sequences from various pneumococcal serotypes.

	% Identity[a]												
	cps19f	*cps3*	*cap1*	*cps9n*	*cps19b*	*cps20*	*cps14*	*cps4*	*cps8*	*cps2*	*cps23f*	*cps6b*	*cps6a*	*cps19a*
cps19f	100	98.5	97.9	97.9	96.7	96.5	97.9	82.4	73.9	73.7	73.3	73.1	72.9	72.2
cps3		100	98.6	99	97.9	96.9	98.3	83	74.3	74.1	73.7	73.5	73.3	72.7
cap1			100	98.1	97.1	95.9	97.3	82.6	73.7	73.5	73.7	73.5	73.3	72.7
cps9n				100	97.7	96.7	97.7	82.4	74.1	73.9	73.5	73.3	73.1	72.5
cps19b					100	95.6	96.5	84.5	75.4	75.2	75.2	75	74.9	73.9
cps20						100	95.8	81.2	73.7	73.5	73.1	72.9	72.7	72.5
cps14							100	82.6	73.9	73.7	73.4	73.2	73	72.4
cps4								100	88.6	88.8	89	88.8	88.6	85.3
cps8									100	99.8	97.7	98.3	98.1	94
cps2										100	97.9	98.5	98.3	94.2
cps23f											100	98.6	98.5	93.6
cps6b												100	99.8	94.2
cps6a													100	94
cps19a														100

a. Percentage of DNA identity over 517 nucleotides, equivalent to nucleotides 2,273-2,789 of the *cps19f* sequence (GenBank accession no U09239), as determined in DNASIS.

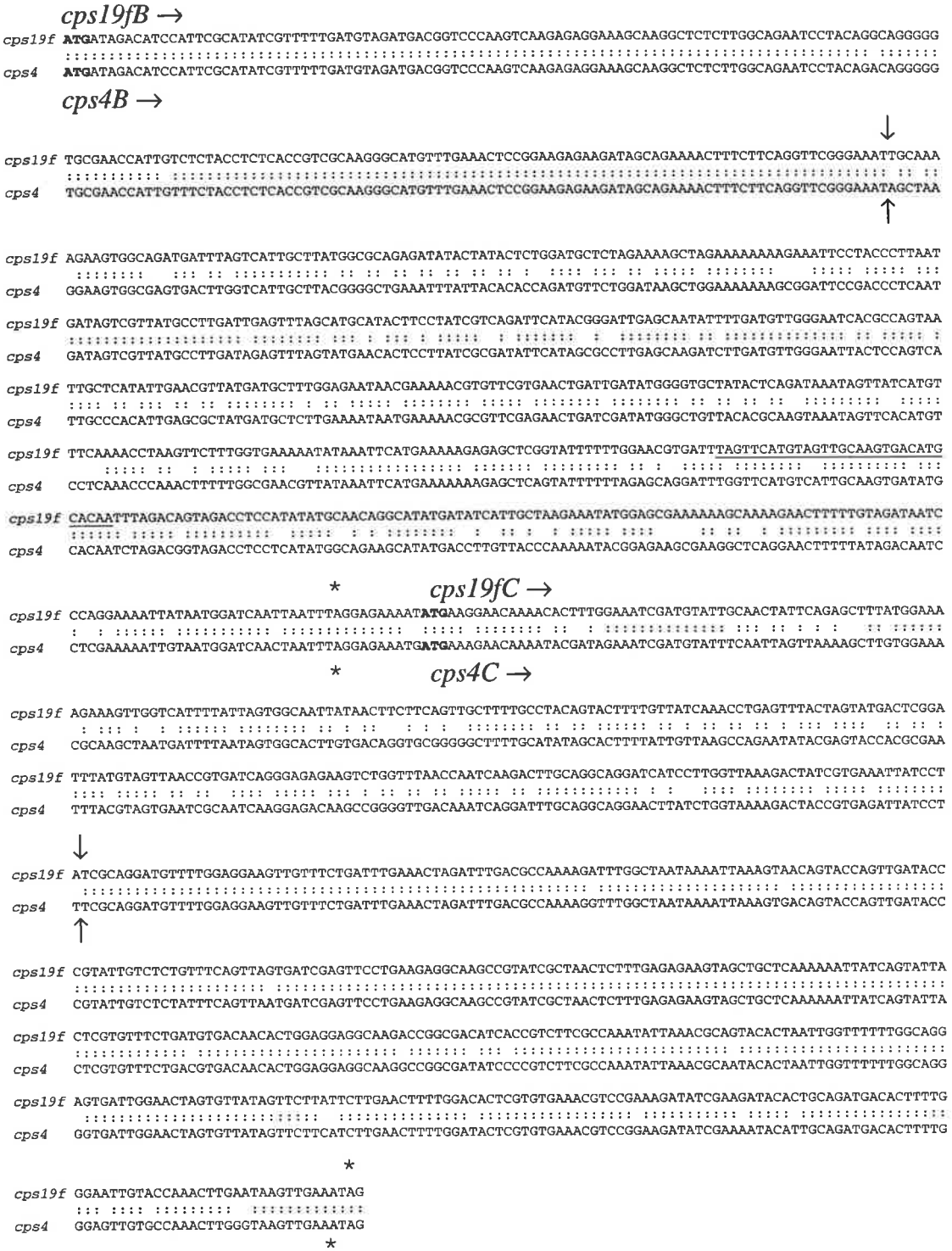


Fig. 4.5. Comparison of the *cps19fB* and *C* genes with the *cps4B* and *C* genes. The first nucleotide of the *cps19fB* sequence corresponds to nucleotide 1,616 of the type 19F sequence (Guidolin *et al.*, 1994). The ATG start codons of *cpsB* and *C* are shown in bold and the stop codons are indicated by an asterisk. The primer J39 sequence is underlined. The arrows indicate the approximate points of sequence divergence.

4.2.5 Phylogenetic analysis of *cpsC* sequences

To further investigate the differences between the two classes of *cpsC* sequences, their phylogenetic relationship was investigated. An alignment of the partial *cpsC* sequences was generated using CLUSTAL W (Thompson *et al.*, 1994; data not shown) and this alignment was used to generate a phylogenetic tree using the Neighbour-Joining method, and the distance measure of Tamura and Nei (1993), as implemented in the program MEGA (Kumar *et al.*, 1994). The tree in **Fig. 4.6** shows two highly significant clusters of *cpsC* sequences, 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 cross-over point near nucleotide 345 (as shown in **Fig. 4.4**) 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 409-444 in **Fig. 4.4**) which is presumably the result of a recombination event.

4.3 Conclusions

The Southern hybridisation data described in this chapter indicate that the DNA flanking the *cps* locus is highly conserved among pneumococci of diverse serotypes. Using LR-PCR with primers specific for common regions, it was possible to amplify the major portion of the capsule loci from several different pneumococcal serotypes. Southern hybridisation analysis confirmed that the large PCR products obtained did indeed contain *cps*-related DNA. Moreover, direct sequencing of the PCR products identified two

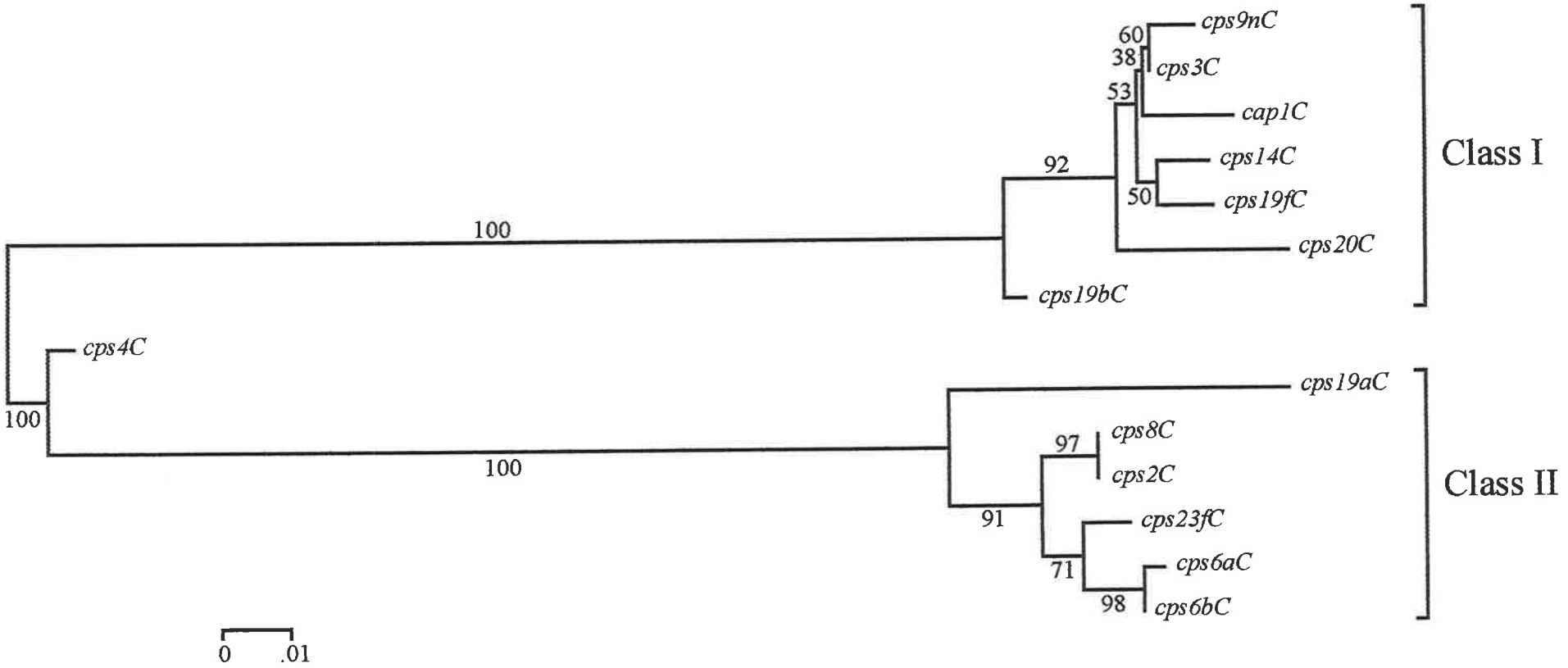


Fig. 4.6. Phylogenetic tree of *cpsC* sequences. The *cpsC* gene sequences were aligned in CLUSTAL W (Thompson *et al*, 1994), and the phylogenetic tree generated using MEGA (Kumar *et al*, 1994), as described in the text. The numbers associated with the branches are Bootstrapping Confidence Limits, resulting from 500 replications, as defined in MEGA. The scale represents the number of nucleotide substitutions per site.

apparent classes of the *cpsC* gene. This was confirmed by phylogenetic analysis of the sequence data. The presence of the *cpsC* gene in all *cps* loci examined is consistent with the important proposed role of CpsC as a chain length regulator in pneumococcal CPS production (section 1.9.3). At this stage, it is not possible to determine whether the differences between class I and class II *cpsC* genes is 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 have previously been shown to impact on the modal chain length of the LPS O-antigen (Klee *et al.*, 1997).

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 (Coffey *et al.*, 1998a; Nesin *et al.*, 1998). 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.

A comparison between the *cps19f* and *cps4* sequences revealed that the distinction between class I and class II sequences can be extended to include the 3' region of *cpsB*, where the J39 primer sequence is located (**Fig. 4.5**). The 3' region of *cps19fB* exhibits only 75% similarity with the 3' region of *cps4B*. It is possible that the *cps* loci from some serotypes may be even more variable in this region, such that the J39 primer is non-functional, and could explain why a LR-PCR product was not obtained in some of the serotypes tested. This suggests that the efficiency of the LR-PCR could be significantly

improved by using a primer sequence from the 5' region of *cps19fB*, which appears to be more conserved, in lieu of J39 (**Fig. 4.5**).

Although LR-PCR successfully obtained PCR products from 13 serotypes, it was unsuccessful from 8 others. This may be simply due to the size of *cpsB-aliA* regions from these serotypes. Indeed, the CPS of the serotypes for which LR-PCR products were not obtained all have large repeat units containing 5 to 7 sugars. No attempts were made to modify the LR-PCR amplification conditions used, but it is likely that such modifications could improve the amplification efficiency of the larger *cps* loci. The possibility of gene rearrangements within the *cps* loci of these strains also cannot be excluded.

The LR-PCR products that have so far been obtained provide a ready source of DNA to study the as yet uncharacterised *cps* genes required for CPS production from serotypes 6A, 6B, 8, 9N, 18C and 20. Additionally, with improvements, the LR-PCR protocol has the potential to amplify the *cps* loci of other serotypes. This would extend its utility for the epidemiological and clonal analysis of clinical isolates.

Chapter 5

CHARACTERISATION OF THE *S. PNEUMONIAE* TYPE 19B CAPSULE LOCUS

5.1 Introduction

The structure of the type 19B CPS is more complex than that of type 19F. The oligosaccharide repeat unit of type 19B CPS has a fourth sugar (Man*N*Ac) in its backbone, with a disaccharide side-chain, as shown in **Fig. 5.1**. Thus the type 19B capsule locus (*cps19b*) would be predicted to contain extra genes required for biosynthesis of the more complicated repeat unit; one would also predict that a distinct polysaccharide repeat unit transporter and polysaccharide polymerase would be required.

In chapter 4, Southern hybridisation was used to demonstrate that all but 2 of the 15 genes in the *S. pneumoniae* type 19F capsule locus hybridised to the DNA of a type 19B *S. pneumoniae* strain. These 2 genes, *cps19fI* and *cps19fJ*, encode the polysaccharide polymerase and polysaccharide repeat unit transporter, respectively, and are located together near the middle of the *cps19f* locus. DNA from the 2 flanking genes, *cps19fH* and *cps19fK*, hybridised only weakly to 19B DNA, whereas all other *cps19f* genes hybridised strongly at high stringency. These data suggested that the type 19B capsule locus does contain different polysaccharide repeat unit transporter and polysaccharide

polymerase genes. It was not obvious, however, where the postulated additional genes required for type 19B CPS biosynthesis were located and hence characterisation of the *cps19b* locus was undertaken.

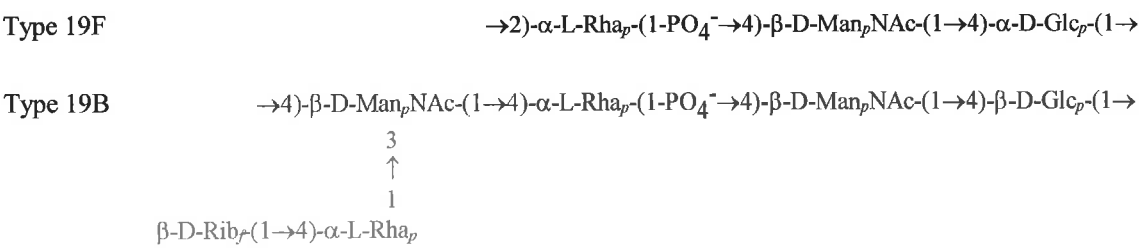


Fig. 5.1. Biological repeat units of pneumococcal type 19F and type 19B capsular polysaccharide. The order of the sugars in the repeat units have been altered compared to the published chemical structures for 19F (Ohno *et al.*, 1980) and 19B (Beynon *et al.*, 1991) as it has been determined that glucose is the first sugar in the biological repeat unit. D-Glu$_p$, glucose; D-Man$_p$NAc, *N*-acetyl mannosamine; L-Rha$_p$, rhamnose; D-Rib$_f$, ribose; PO_4^-, phosphate;.

5.2 Results

5.2.1 Isolation of the type 19B-specific *cps* genes

In order to isolate the type-specific portion of the *cps19b* locus, primers J5 which is homologous to *cps19fF* and J27 which is complementary to *cps19fL* (**Table 2.4**) were used to amplify the intervening region of the 19B chromosome using LR-PCR with an extension time of 12 min per cycle (see section 2.9.2). The LR-PCR product (from the 3' end of *cps19bB* to *aliA*) which was described in chapter 4 was not used because it contains a large amount of DNA (*cps19bC-cps19bG* and *cps19bL-aliA*) which was predicted (on the basis of hybridisation data) to be greater than 95% identical to *cps19f*. The J5/J27-directed PCR product was approximately 10.5 kb in size. This is 4 kb larger than the equivalent region of *cps19f*, indicating that extra ORFs may be present in this region of the *cps19b* locus. A

map of the 10.5-kb PCR product was generated, using the restriction enzymes *Bam*HI, *Cla*I, *Hin*dIII and *Eco*RI, and appropriate fragments of the PCR product were then cloned into pBluescript KS+, generating four recombinant plasmids with inserts as shown in **Fig. 5.2**.

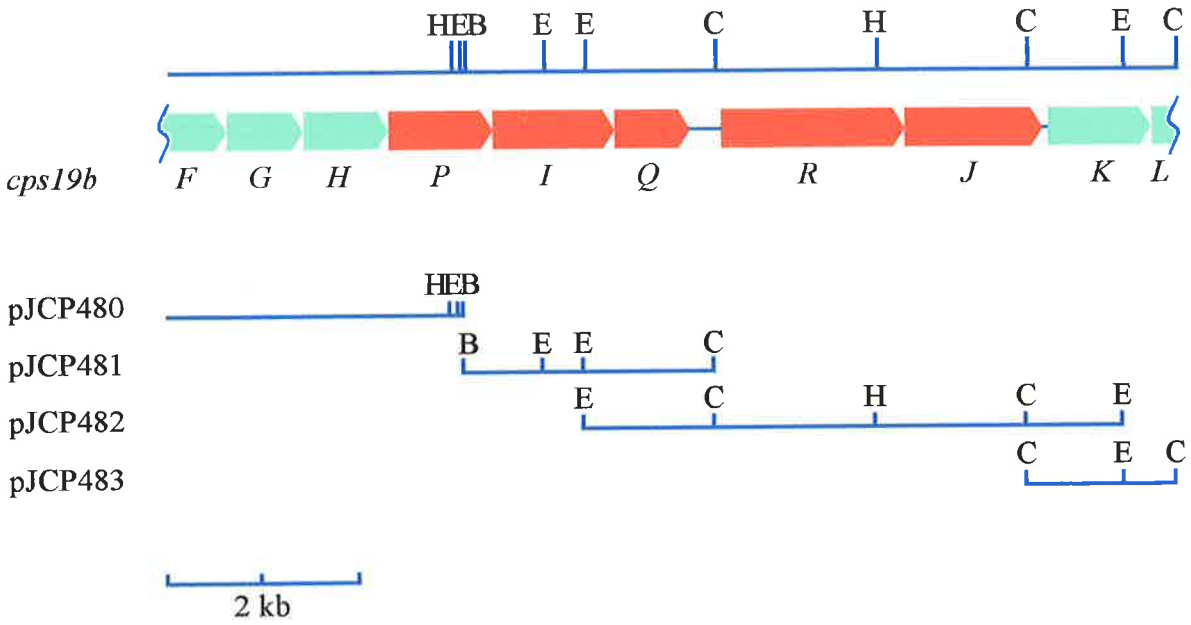


Fig. 5.2. Physical map of part of the *S. pneumoniae* type 19B capsule locus (*cps19b*). Green arrows represent ORFs also present in *cps19f* and red arrows represent type 19B-specific ORFs. Gene designations are indicated below the map; *cps19bF-L* are abbreviated to *F-L*, respectively. Restriction sites are as follows; B, *Bam*HI; C, *Cla*I; E, *Eco*RI; H, *Hin*dIII. The regions of DNA subcloned into various recombinant plasmids are shown below the map.

Both strands of the pneumococcal DNA inserts of each of the above plasmids (or nested deletion derivatives thereof) were subjected to sequence analysis (section 2.8) in order to compile the sequence of this portion of the *cps19b* locus, as shown in **Appendix II**. The sequence across the *Bam*HI site at the junction of the inserts of pJCP480 and pJCP481 was obtained by dye terminator sequencing (section 2.8.3) of the PCR product with the primer J44 (**Table 2.4**). Examination of the compiled sequence revealed, as

expected, that the first 2.3 kb of sequence at the 5' end has a high degree of similarity to the *cps19f* sequence. This region contains the homologues to *cps19fG* and *cps19fH* (*cps19bG* and *cps19bH*) which exhibited 98.3% and 91.1% identity, respectively. The sequence then diverges suddenly (in the vicinity of nucleotide 2,319 of the *cps19b* sequence) just prior to the end of *cps19bH*, truncating the *cps19bH* gene product by two amino acids with respect to Cps19fH, as shown in **Fig. 5.3A**. There are five new potential

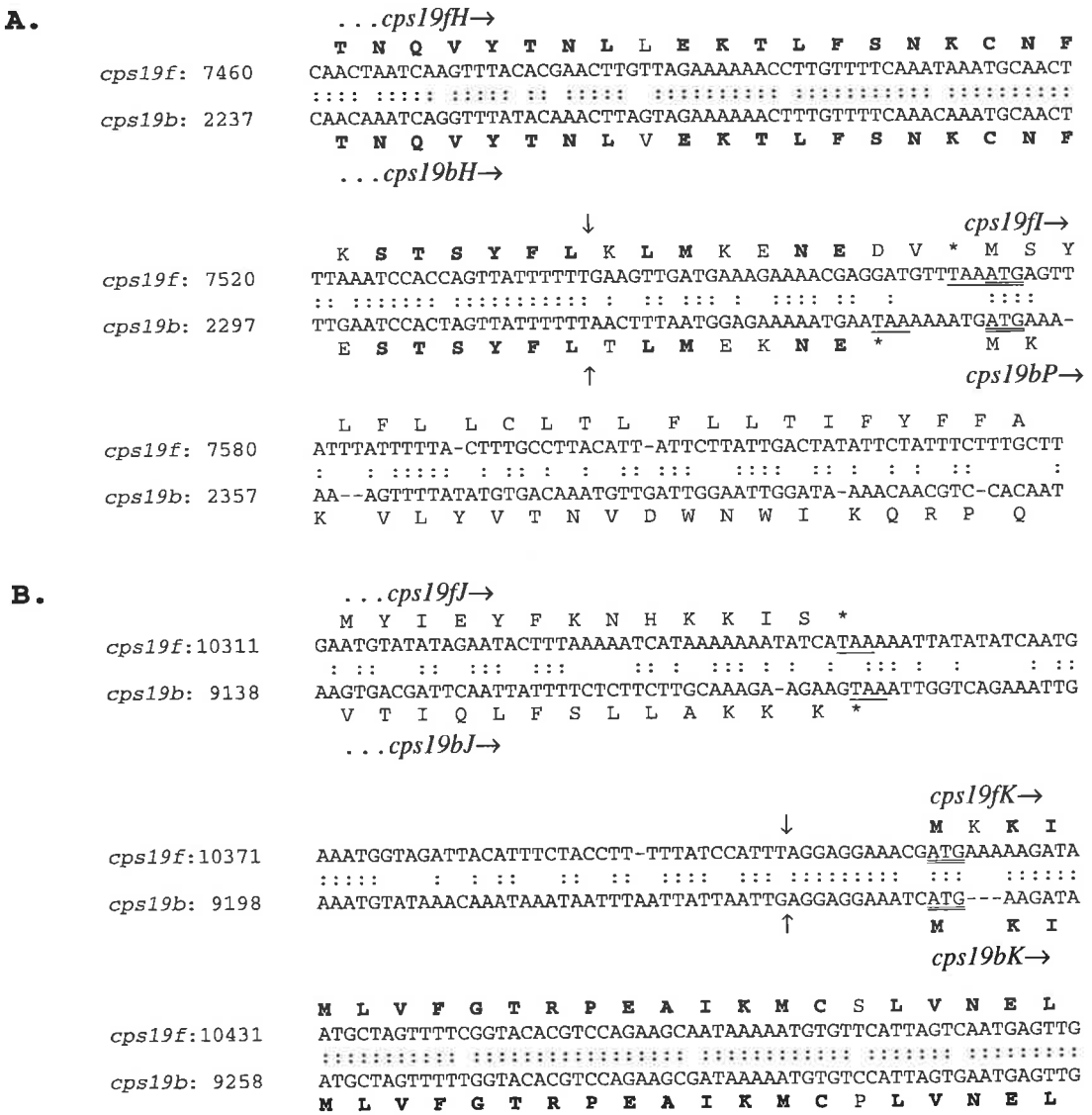


Fig. 5.3. Homology between *cps19f* and *cps19b*. Location of the approximate points of sequence divergence, at the end of *cps19H* (**A**), and immediately upstream of *cps19K* (**B**). The amino acid translation for the depicted ORFs are shown above the sequence for *cps19f* and below the sequence for *cps19b*. The stop codons are underlined and the ATG start codons are double underlined. Arrows indicate the approximate points of divergence. Identical amino acids are shown in bold.

ORFs, between *cps19bH* and *cps19bK*, which have been designated *cps19bP, I, Q, R* and *J*. Each ORF is preceded by a ribosome binding site and the majority are very closely linked. The only potentially significant intergenic gap of 204 nucleotides occurs between *cps19bQ* and *cps19bR*. However, no potential stemmed-loop structures or obvious promoter sequences were found in this region. As predicted, the sequence at the 3' end of the 10.5 kb *cps19b* PCR product again shows similarity to the *cps19f* sequence, starting in the vicinity of nucleotide 9,233 (**Fig. 5.3B**); this is immediately before the start of the *cps19bK* gene, which has 93% identity to *cps19fK*.

5.2.2 Characterisation of the *cps19b* genes

The locations and several properties of each of the type 19B-specific ORFs, *cps19bP, I, Q, R* and *J*, are summarised in **Table 5.1**. Significant similarities with other known proteins, revealed by comparison with sequence databases, are described below.

Table 5.1. Summary of ORFs *cps19bP-J*.

ORF	Location in sequence[a]	Predicted MW	No. amino acids	Hydrophobicity index[b]	Predicted pI	% G+C content[c]
cps19bP	2,350-3,432	43,334	361	-0.41	8.69	29.5
cps19bI	3,451-4,695	48,667	414	0.77	9.73	27.2
cps19bQ	4,703-5,605	34,876	300	-0.26	8.20	29.7
cps19bR	5,809-7,746	76,348	645	-0.30	8.65	27.2
cps19bJ	7,736-9,181	53,851	481	0.91	9.89	29.5

a. Nucleotide numbers correspond to *cps19b* sequence as shown in Appendix II.
b. According to Kyte and Doolittle (1982), as implemented in PROSIS.
c. Percent guanine plus cytosine (G+C) of coding region.

5.2.2.1 *cps19bP*

The *cps19bP* gene encodes a putative 43.3 kDa protein. Database searches with Cps19bP did not reveal any significant similarity to any other proteins.

5.2.2.2 *cps19bI*

The *cps19bI* gene encodes a putative 48.7 kDa protein. Database searches with Cps19bI also failed to identify any significant similarity to any other proteins. However, Cps19bI is a very hydrophobic protein and the hydropathy plot (**Fig. 5.4**) illustrates the marked similarity between Cps19bI and Cps19fI, each having 10-12 hydrophobic, potentially membrane-spanning domains. This is a typical hydropathy profile for Rfc-like proteins and suggests that Cps19bI, like Cps19fI, may be a polysaccharide polymerase.

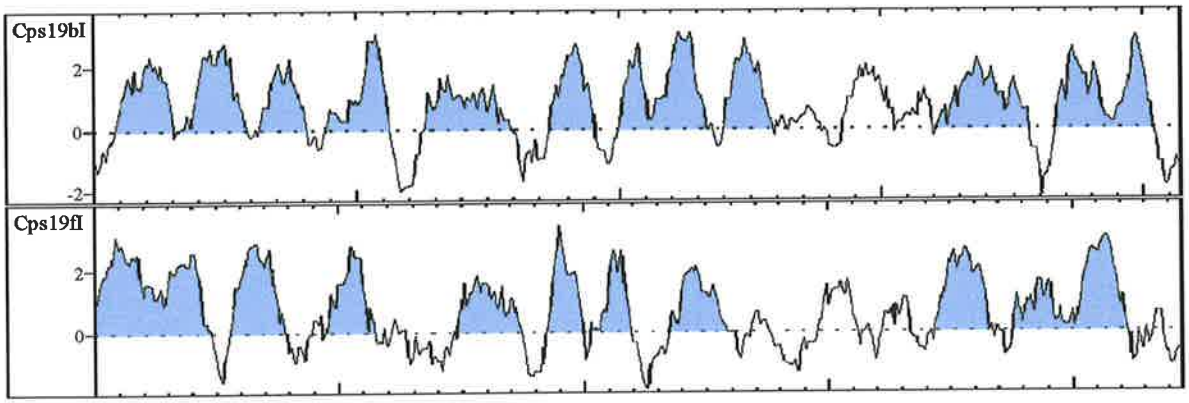


Fig. 5.4. Hydropathy plots of Cps19bI and Cps19fI. The hydropathy plots were generated by the method of Kyte and Doolittle (1982) and aligned using PROFILEGRAPH (Hofmann and Stöffel, 1989). Positive numbers on the Y-axis indicate hydrophobic regions. The position of every 10th amino acid is marked on each X-axis.

5.2.2.3 *cps19bQ*

The *cps19bQ* gene encodes a putative 34.9 kDa protein, which has similarity with rhamnosyl transferases from *Leptospira interrogans, Shigella dysenteriae*, and *S. flexneri*,

and a 6-deoxyaltrosyl transferase from *Y. enterocolitica* as shown in **Table 5.2.** The alignment of Cps19bQ with the other proteins shows several regions of similarity, including a motif previously identified in rhamnosyl and other 6-deoxyhexosyl transferases (section 3.2.3.2; Morona *et al.*, 1995), as shown in **Fig. 5.5.** This motif contains conserved aspartate residues and is reminiscent of the catalytic sites identified in $RfbA_{O:54}$ of *S. enterica* serovar borreze (Keenleyside and Whitfield, 1996). A second conserved motif was also identified (**Fig. 5.5**, yellow shaded amino acids), suggesting that these proteins form a closely related sub-group of this type of transferase. Cps19bQ is the second putative rhamnosyl transferase in the *cps19b* locus. Cps19bH, which has 91.1% identity with Cps19fH, is also proposed to be a rhamnosyl transferase, adding Rha to the repeat unit backbone (section 3.2.3.2). Thus Cps19bQ is predicted to add Rha (the first sugar of the disaccharide side-chain [see **Fig. 5.1**]) to the distal Man*N*Ac.

Table 5.2. Similarity of Cps19bQ to other proteins.

	% Identity[a]				
	Cps19bQ[b]	SdRfbQ[c]	LiRfbF[d]	YeRfbC[e]	SfRfbF[f]
Cps19bQ	100	24	22.9	22.1	20.2
		[267]	[245]	[213]	[257]
SdRfbQ		100	31.3	30	30.2
			[297]	[300]	[298]
LiRfbF			100	28.9	30.6
				[308]	[284]
YeRfbC				100	29.5
					[298]
SfRfbF					100

a. Percentage of identical amino acids determined with FASTA as implemented in PROSIS. Numbers in parentheses indicate the number of amino acids over which the % identity occurs.
b. *S. pneumoniae* Cps19bQ.
c. *S. dysenteriae* RfbQ (Klena and Schnaitman, 1993).
d. *L. interrogans* RfbF (Mitchison *et al.*, 1997).
e. *Y. enterocolitica* RfbC (Zhang *et al.*, 1993).
f. *S. flexneri* RfbF (Macpherson *et al.*, 1994).

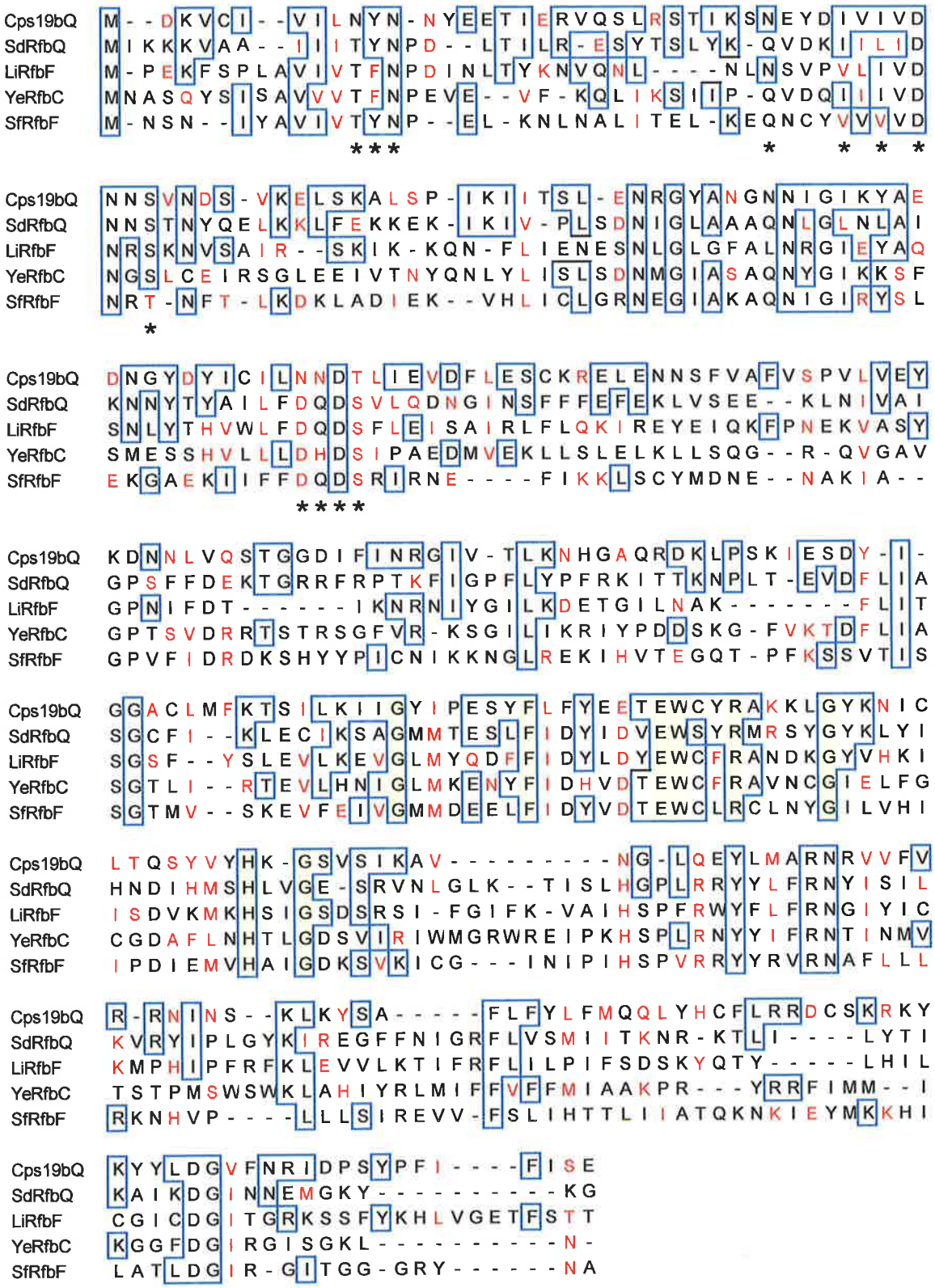


Fig. 5.5. Alignment of Cps19bQ with other proteins. Alignment of Cps19bQ with *S. dysenteriae* RfbQ, (SdRfbQ) (Klena and Schnaitman, 1993), *L. interrogans* RfbF, (LiRfbF) (Mitchison *et al.*, 1997), *Y. enterocolitica* RfbC, (YeRfbC) (Zhang *et al.*, 1993) and *S. flexneri* RfbF, (SfRfbF) (Morona *et al.*, 1995) using the default settings of the program CLUSTAL (Higgins and Sharp, 1988) and enhanced by manual adjustment. Residues identical to Cps19bQ are boxed; similar residues are shown in red; - indicates absence of a residue. The asterisks correspond to a motif found in a variety of rhamnosyl and 6-deoxy-hexosyl transferases (section 3.2.3.2; Morona *et al.*, 1995). The yellow shaded region corresponds to an additional motif shared by this sub-group of transferases.

5.2.2.4 *cps19bR*

The *cps19bR* gene encodes a putative 76.3 kDa protein. Cps19bR has similarity (21.4% identity, 40.3% similarity) to the central portion (amino acids 366-1071) of hypothetical protein 3 from the capsule locus of *H. influenzae* type b (Van Eldere *et al.*, 1995) as shown in **Fig. 5.6**. Although the function of hypothetical protein 3 is not known, it is located in the serotype-specific Region II of the capsule locus. The only sugar in the *H. influenzae* type b capsule also found in the pneumococcal type 19B CPS is ribose (Rib). Thus it is possible that Cps19bR and protein 3 could function in the synthesis of an activated Rib precursor. However, the nature of the activated Rib precursor that is used in CPS biosynthesis in bacteria is not known (Van Eldere *et al.*, 1995), and the possibility that both proteins are ribosyl transferases cannot be excluded.

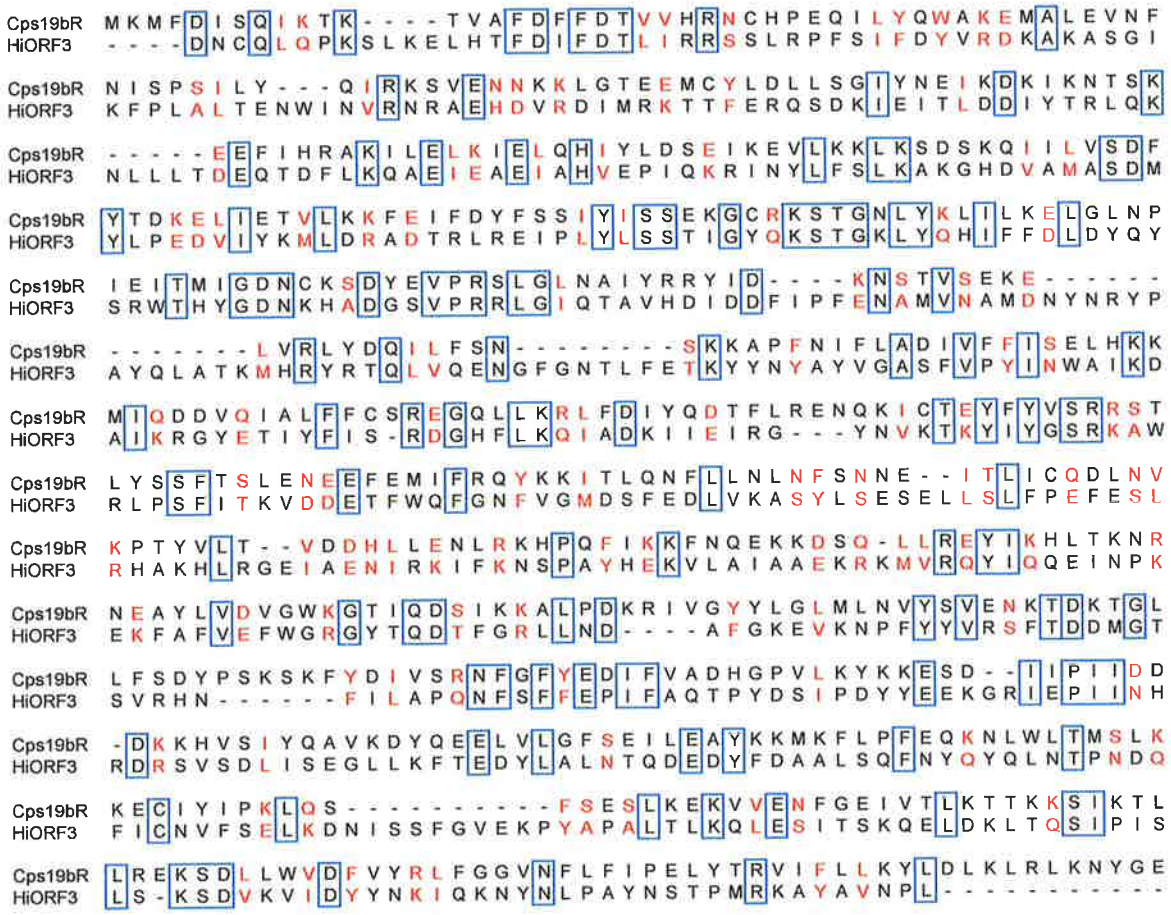


Fig. 5.6. Alignment of Cps19bR with *H. influenzae* ORF 3 (HiORF3). Alignment was performed using the default settings of the program CLUSTAL (Higgins and Sharp, 1988). Residues identical to Cps19bR are boxed; similar residues are shown in red; - indicates absence of a residue.

5.2.2.5 *cps19bJ*

The *cps19bJ* gene encodes a putative 53.9 kDa protein with low level similarity to RfbX proteins from *E. coli, S. dysenteriae*, *Y. enterocolitica*, to the CapF protein of *S. aureus* and to Cps19fJ, as shown in **Table 5.3**. The RfbX proteins are known to be involved in export of O-antigen repeat units (Liu *et al.*, 1996; Macpherson *et al.*, 1995). The hydropathy plots for RfbX-like proteins are all very similar, with 10-12 hydrophobic, membrane-spanning domains, and the similarity between Cps19bJ and Cps19fJ is shown in **Fig. 5.7**. Thus, Cps19bJ is likely to be the polysaccharide repeat unit transporter.

Table 5.3. Similarity of Cps19bJ to other proteins.

	% Identity[a]						
	Cps19bJ[b]	K12RfbX[c]	YeTrsA[d]	SdRfbX[e]	SaCapF[f]	YeRfbX[g]	Cps19fJ[h]
Cps19bJ	100	23.3	21.6	22.3	21	19	18.2
		[404]	[402]	[394]	[395]	[420]	[406]
K12RfbX		100	28.3	31.4	21.5	16.8	18.1
			[406]	[401]	[395]	[386]	[414]
YeTrsA			100	28.4	20.4	17.5	19.2
				[401]	[401]	[406]	[416]
SdRfbX				100	21.1	16.7	19.4
					[393]	[377]	[402]
SaCapF					100	20	16.4
						[404]	[397]
YeRfbX						100	18.1
							[425]
Cps19fJ							100

a. Percentage of identical amino acids determined with FASTA as implemented in PROSIS. Numbers in parentheses indicate the number of amino acids over which the % identity occurs.
b. *S. pneumoniae* Cps19bJ.
c. *E. coli* K12 RfbX (Stevenson *et al.*, 1994)
d. *Y. enterocolitica* TrsA (Skurnik *et al.*, 1995)
e. *S. dysenteriae* RfbX (Klena and Schnaitman, 1993).
f. *S. aureus* CapF (Lin *et al.*, 1994).
g. *Y. enterocolitica* RfbX (Zhang *et al.*, 1997).
h. *S. pneumoniae* Cps19fJ (section 3.2.3.4)

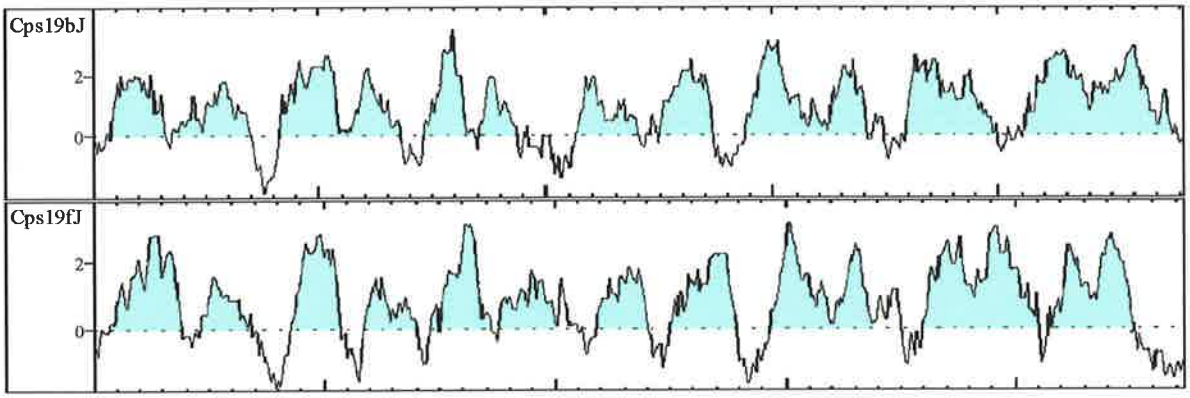


Fig. 5.7. Hydropathy plots of Cps19bJ and Cps19fJ. Hydropathy plots were generated by the method of Kyte and Doolittle (1982) and aligned using PROFILEGRAPH (Hofmann and Stöffel, 1989). Positive numbers on the Y-axis indicate hydrophobic regions. The position of every 10th amino acid is marked on each X-axis.

5.2.3 Serotype specificity of the *cps19b* genes

To examine the relationship between *cps19b* and encapsulation loci of other *S. pneumoniae* serotypes, DNA fragments corresponding to the individual *cps19b* genes (described in **Table 5.4**) were labelled with DIG and used to probe (at high stringency) Southern blots of restricted chromosomal DNA from representative pneumococci belonging to the following types/groups: 2, 3, 4, 6, 7F, 7B, 8, 9, 12, 14, 16, 17, 18, 19F, 19A, 19C, 22, 23 and 24. None of the probes hybridised to DNA from any serotype tested, except to the closely related type 19C (**Table 1.3**), which has high stringency homologues for all 5 genes (result not shown).

Table 5.4 DNA fragments used as gene-specific probes

Probe	Fragment used as probe:
cps19bP	*Spe*I-*Bam*HI fragment from pJCP480 (nucleotides 2,305-3,188 of *cps19b* sequence)
cps19bI	*Pvu*II-*Hin*dII fragment from pJCP481 (nucleotides 3,412-4,545 of *cps19b* sequence)
cps19bQ	*Hin*dII-*Cla*I fragment from pJCP481 (nucleotides 4,540-5,718 of *cps19b* sequence)
cps19bR	*Cla*I-*Nsi*I fragment from pJCP482 (nucleotides 5,713-7,766 of *cps19b* sequence)
cps19bJ	*Nsi*I-*Pvu*II fragment from pJCP482 (nucleotides 7,760-9,046 of *cps19b* sequence)

5.2.4 Capsule type switching by transformation

To determine if the genes sequenced were sufficient for type 19B CPS biosynthesis, the 10.5-kb PCR product was transformed into an unencapsulated, Ery-resistant, derivative of Rx1-19F, designated Rx1-19F-I, in which the *cps19fI* gene had been disrupted by insertion-duplication mutagenesis using pVA891, as described in section 3.2.4. Several smooth transformants were checked for Ery sensitivity, indicating loss of the pVA891 sequence. Southern hybridisation was used to confirm the absence of both pVA891 and the *cps19fI* gene, and the presence of each of the *cps19bP*, *I*, *Q*, *R* and *J*, genes in one transformant, designated Rx1-19B. The hybridisation results, using probes specific for *cps19fI* and *cps19bI*, are shown in **Fig. 5.8**; other results not shown. The production of a type 19B capsule by Rx1-19B was then confirmed by quellung reaction. This shows that it is possible to alter capsule production from type 19F to type 19B by replacing part of the *cps* locus, and that the region of *cps19b* described in this study determines the 19B serotype.

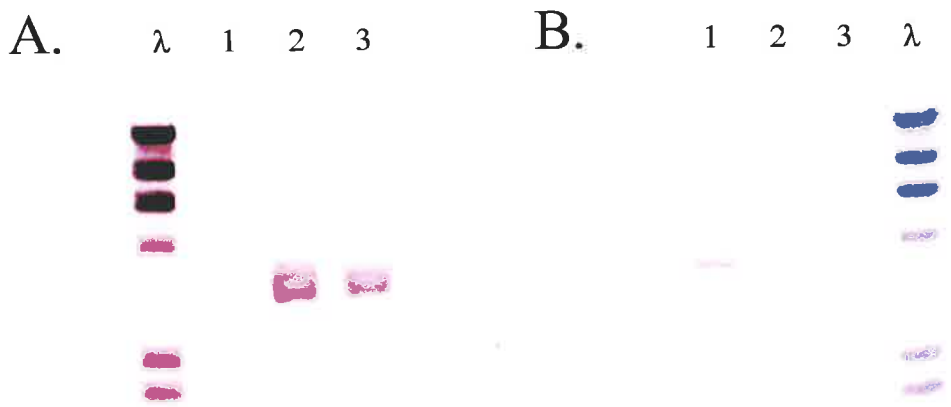


Fig. 5.8. Southern hybridisation of Rx1-19B. *Cla*I-restricted chromosomal DNA of *S. pneumoniae* strains Rx1-19F (1), Rx1-19B (2), and 19B (3) was probed with DIG labelled probes specific for *cps19bI* (**A**) and *cps19fI* (**B**). The sizes of the DIG-labelled lambda (λ) markers are as follows: 23 kb, 9.4 kb, 6.6 kb, 4.4 kb, 2.3 kb and 2.0 kb.

5.3 Conclusions

When the presence of the 15 *cps19f* genes in other pneumococcal serotypes was examined, the hybridisation patterns showed blocks of *cps19f* genes which hybridised to at least one other serotype flanking blocks of genes which did not (**Table 3.1**). This suggests that different serotypes may have evolved as a consequence of the replacement of one cluster of genes within the capsule locus with an alternative gene cluster. The work described in this chapter demonstrates that the five additional type-specific genes in the *cps19b* locus are indeed grouped together. Transformation of Rx1-19F-I with the 10.5-kb *cps19b* PCR product yielded a transformant, Rx1-19B, which expresses type 19B capsule. Thus, this cluster of five genes (*cps19bPIQRJ*) is sufficient to encode all of the additional and/or distinct functions required for production of type 19B rather than type 19F CPS.

The chemical structure of the *S. pneumoniae* type 19B capsule is considerably more complex than that for type 19F. The type 19B backbone has an additional Man*N*Ac which also carries a (1→3) linked β-D-Rib$_f$-(1→4)-α-L-Rha$_p$ side-chain (**Fig. 5.1**). Therefore, biosynthesis of type 19B CPS would be predicted to require several additional and/or different enzymes; 3 transferases for the addition of the 3 extra sugars, at least one enzyme for the synthesis of the activated Rib precursor, a distinct polysaccharide repeat unit transporter and a distinct polysaccharide polymerase.

Analysis of the predicted protein products from the type 19B ORFs identified candidates for several of these enzymes. These are the rhamnosyl transferase (Cps19bQ), needed for the addition of Rha to the distal Man*N*Ac, the polysaccharide repeat unit transporter (Cps19bJ), and the polysaccharide polymerase (Cps19bI). Cps19bR is the most likely candidate for the enzyme required for the synthesis of the activated Rib precursor. However, the two transferases needed for the addition of the distal Man*N*Ac and Rib

remain unidentified. It is possible that Cps19bP functions as one of these transferases. However, there are no other ORFs in the type-specific region of *cps19b* which could encode the other transferase.

One possible explanation for the absence of an ORF encoding a third transferase is as follows. In type 19B CPS both Man*N*Ac sugars are (1→4) linked. Cps19bF, which is almost identical to Cps19fF, a putative Man*N*Ac transferase (section 1.9.3), could be responsible for the addition of both Man*N*Ac residues. The ability of a single transferase to transfer the same sugar to what appears to be different acceptors has been proposed previously. For example, RfbG in *Shigella flexneri* is thought to add Rha, via the same linkage, at two separate positions within the O-antigen repeat unit (Morona *et al.*, 1995). Interestingly, type 19F and type 19A, have closely related *cps19J* (repeat unit transporter) genes (section 4.2.1) and their capsules are similar, as both have a trisaccharide backbone containing only one Man*N*Ac (**Table 1.3**). On the other hand, both type 19B and type 19C, which also have highly homologous *cps19J* genes, have a tetrasaccharide backbone containing two Man*N*Ac residues (**Table 1.3**). Given that Cps19fJ and Cps19bJ are quite distinct, the difference in the CPS backbone might be explained by the specificity of these polysaccharide repeat unit transporters. Cps19fJ may have absolute specificity for the trisaccharide repeat unit and may thus prevent Cps19fF from adding an additional Man*N*Ac sugar to the backbone. On the other hand, Cps19bJ may be specific for the tetrasaccharide backbone and may only transport the repeat unit after the addition of both the distal Man*N*Ac by Cps19bF and the disaccharide side-chain by Cps19bQ and Cps19bP. This hypothesis could be investigated by the construction of an Rx1-19F in-frame *cps19fJ* deletion mutant. Biochemical analysis of this strain should determine the presence of either a trisaccharide or a tetrasaccharide repeat unit linked to the lipid carrier. The presence of a tetrasaccharide would support the above hypothesis.

Alternatively, Cps19bR could be the ribosyl transferase, while Cps19bP could be the second Man*N*Ac transferase. However, the activated Rib precursor is unlikely to be ubiquitous to the pneumococcus, as Rib is not a component of any essential cell polysaccharide, such as teichoic acid. Cps19bR is a relatively large protein and could be bi-functional, capable of both synthesising the activated Rib precursor and transferring it to the repeat unit. Construction and characterisation of defined in-frame deletion mutants in all of the type 19B-specific genes described in this chapter are required to determine their precise functions.

Chapter 6

ANALYSIS OF THE *S. PNEUMONIAE* TYPE 19A *CPS* LOCUS

6.1 Introduction

Analysis of purified type 19A CPS has yielded two distinct putative structures (**Fig. 6.1**). One is the same as type 19F except for a 1→3 linkage (rather than 1→2) between Glc and Rha (Katzenellenbogen and Jennings, 1983). This difference would necessitate an alteration only in the specificity of the polysaccharide polymerase (Cps19fI). The alternative structure involves the same trisaccharide backbone as type 19F, but with additional β-D-Glc$_p$NAc-(1→3)-β-D-Gal$_p$-(1-PO_4^-→2) and α-L-Fuc$_p$-(1-PO_4^-→3) side chains attached to the Glc and Rha, respectively (Lee and Fraser, 1980). This would necessitate a number of additional enzyme activities not found in the *cps* locus of type 19F strains. Interestingly, individual type 19A strains were subsequently reported to be capable of producing either structural type, depending on the growth conditions (Lee *et al.*, 1987). Sequence analysis of the type 19A *cps* locus was undertaken in an attempt to provide a molecular explanation for these findings.

Type 19F →2)-α-L-Rha$_p$-(1-PO_4^-→4)-β-D-Man$_p$NAc-(1→4)-α-D-Glc$_p$-(1→

Type 19A (1) →3)-α-L-Rha$_p$-(1-PO_4^-→4)-β-D-Man$_p$NAc-(1→4)-α-D-Glc$_p$-(1→

(2) →2)-α-L-Rha$_p$-(1-PO_4^-→4)-β-D-Man$_p$*N*Ac-(1→4)-α-D-Glc$_p$-(1→
3 2
↑ ↑
α-L-Fuc$_p$-(1→PO_4^- β-D-Glc$_p$*N*Ac-(1→3)-β-D-Gal$_p$-(1- PO_4^-

Fig. 6.1. Biological repeat units of pneumococcal type 19F and type 19A capsular polysaccharide. The order of the sugars in the repeat units have been altered compared to the published chemical structures for 19F (Ohno *et al*, 1980) and 19A (Lee and Fraser, 1980; Lee *et al.*, 1987) reflecting the fact that glucose is the first sugar in the biological repeat unit. D-Glu$_p$, glucose; D-Man$_p$*N*Ac, *N*-acetyl mannosamine; L-Rha$_p$, rhamnose; D-Gal$_p$, galactose; D-Glc$_p$*N*Ac, *N*-acetyl glucosamine; L-Fuc$_p$, fucose; PO_4^-, phosphate.

6.2 Results

6.2.1 PCR amplification and sequencing of the type 19A locus

Southern hybridisation data (section 4.2.1) suggested that the *cps* loci of type 19F and 19A are significantly different. Only 7 of the 15 *cps19f* gene-specific probes (*cps19fA, B, G, H, M, N* and *O*) hybridised at high stringency to 19A chromosomal DNA. It was assumed that the arrangement of these conserved genes within the two loci would be similar. Thus, a series of over-lapping DNA fragments containing type 19A-specific genes flanked by conserved sequences were generated by LR-PCR (section 2.9.2) using primers based on the *cps19f* sequence. The primers used are described in **Table 2.4** and a map of the PCR products spanning the entire *cps19a* locus is shown in **Fig. 6.2**. DNA from two different type 19A clinical isolates (19A1, obtained from J. Henrichsen, and 19A2, obtained from C. J. Lee [**Table 2.2**]) was used as template. Interestingly, the PCR products amplified from regions between *cps19A* and *cps19J* (using primer pairs CPS5'/J22 and CPSXI/J11 as described in **Table 2.4**) were identical in size for both type 19A isolates and type 19F, but the PCR products obtained from the 5' intergenic regions (using primers

DEXB and CPSA2, as described in **Table 2.4**) varied between all three strains (**Fig. 6.3**). The 19A2 isolate also differed in the 3' region of the *cps19a* locus; the PCR products obtained from the *cps19J* to *cps19O* region (using primer pairs CPSXI/J29, J9/J25 and J9/J36, as described in **Table 2.4**) were either smaller or absent from 19A2 (**Fig. 6.3**). This suggests that part of this region of the *cps19a* locus of this strain may have been deleted. The PCR products obtained from this region of the *cps19a* locus (using primers J88 and J36, described in **Table 2.4**) from six Australian type 19A isolates were either identical in size or larger (probably indicating the presence of an IS element in the 3' intergenic region) than that from 19A1 and 19F (**Fig. 6.4**), suggesting that this part of the *cps19a* locus in 19A2 is atypical.

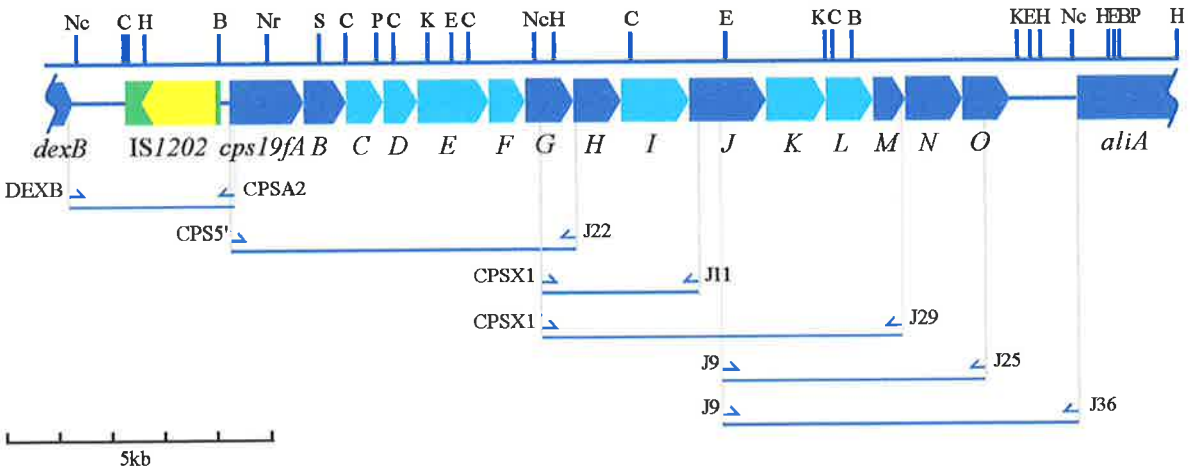


Fig. 6.2. Schematic representation of the PCR products amplified using *cps19f* primers. The *cps19f* genes which hybridised (at high stringency) to type 19A chromosomal DNA are shown in dark blue and the remainder of the locus is shown in pale blue. The primers used are described in **Table 2.4**.

6.2.2 Analysis of the *cps19a* locus

The sequence of the PCR products from 19A1 were determined using specifically designed primers, as described in section 2.8. Analysis of the compiled sequence (**Appendix III**) revealed that the *cps19f* and *cps19a* loci are very closely related. The

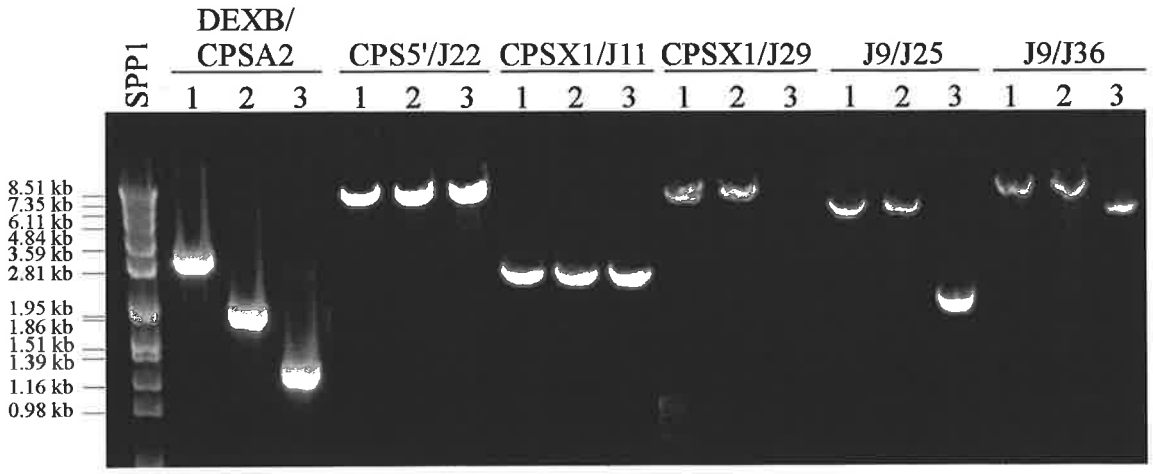


Fig. 6.3. Comparison of the PCR products obtained from *S. pneumoniae* Rx1-19F, 19A1 and 19A2. The PCR products from Rx1-19F (1), 19A1 (2) and 19A2 (3) were electrophoresed in a 0.8% agarose gel and stained with ethidium bromide. *Eco*RI-digested SPP1 bacteriophage DNA was used as a size marker (section 2.6.2); the approximate sizes of the PCR products are indicated on the left side of the figure.

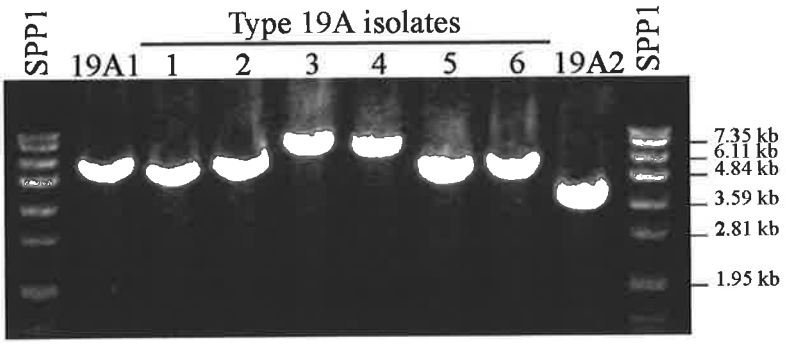


Fig. 6.4. Comparison of the PCR products obtained from the 3' region of the *cps19a* locus from various type 19A isolates. The PCR products were electrophoresed in a 0.8% agarose gel and stained with ethidium bromide. *Eco*RI-digested SPP1 bacteriophage DNA was used as a size marker (section 2.6.2) and the approximate sizes of the fragments are indicated on the right side of the figure.

cps19a locus has the same number of ORFs organised in identical order to those in *cps19f* with homologies to the *cps19f* genes ranging from 70.1% to 99.4% identity. The sizes, G+C content and % identity of the *cps19a* and *cps19f* protein products are shown in **Table 6.1**.

Not withstanding the overall similarity between the *cps19a* and *cps19f* loci, several interesting differences between the two loci were noted. Whereas the start codon for both *cps19aG* and *cps19aH* is TTG, only *cps19fH* has a TTG start codon in the *cps19f* locus.

The intergenic gaps between the *cps19a* genes and the *cps19f* genes are all similar, except between *cps19aK* and *cps19aL* which is much larger (152 nucleotides) compared to that between *cps19fK* and *cps19fL* (38 nucleotides). The largest variation between the *cps19a* and *cps19f* loci occurs in the 5' intergenic region. This region in the 19A1 strain has several deletions compared to the same region in type 19F, but the 3' intergenic regions of types 19F and 19A1 are almost identical (96.7% identity). The differences in the 5' intergenic region will be discussed in section 6.2.5.

Table 6.1. Comparison of the *cps19a* and *cps19f* ORFs.

cps19a	Predicted size			*cps19f*	Predicted size			% Identity	
ORFs	Da	no. aa	%G+C	ORFs	Da	no. aa	%G+C	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
Cps19aH	34,455	292	32.2	Cps19fH	34,474	292	30.3	90.8	95.2
Cps19aI	51,604	444	32.7	Cps19fI	51,734	445	29.7	78.5	80.7
Cps19aJ	54,650	474	33	Cps19fJ	55,055	473	29.7	82.3	83.3
Cps19aK	40,749	362	36.9	Cps19fK	40,950	362	35.2	85.2	92.8
Cps19aL	32,242	289	43.3	Cps19fL	32,215	289	42.3	79.9	92.4
Cps19aM	22,408	198	41.2	Cps19fM	22,379	198	41.5	87.6	94.4
Cps19aN	39,086	349	42.4	Cps19fN	39,053	349	42.1	98.2	99.1
Cps19aO	32,330	283	41.3	Cps19fO	32,330	283	41.5	99.4	99.3

The *cps19a* and *cps19f* sequences were examined to identify potential cross-over points where recombination between the *cps* loci may have occurred. A cross-over point was identified within the *cps19M* gene; the first 348 nucleotides of *cps19aM* have 80.3%

identity to *cps19fM* whereas the remainder of *cps19aM* is 98% identical to *cps19fM*, as shown in **Fig. 6.5B**. However, there is no clearly identifiable cross-over point at the 5' end of the loci; 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. 6.5A**. This suggests that the *cps19a* locus and the type 19A serotype may be the result of several recombination events between the ancestral *cps* locus (possibly type 19F) 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 expressed serotype. The *cps19aB* gene is almost identical (except for the first 42 nucleotides) to the *cps4B* gene which was described in chapter 4, and shows the same point of sequence divergence from the *cps19fB* gene. A small region of *cps19aB* (nucleotides 3,221-3,374) has 100% identity to *cps19fB*. This region presumably accounts for the high stringency hybridisation of the *cps19aB* DNA to a *cps19fB* probe (**Table 4.1**) as there is only 76.7% identity between the remainder of the *cps19aB* and *cps19fB* genes. The highly conserved region may either encode a functionally important domain in the *cps19B* gene product or may simply be the result of a recombination event.

The overall identity between *cps19fJ* and *cps19aJ* is only 82%, which is insufficient for the *cps19fJ* probe to hybridise to the *cps19aJ* gene under high stringency conditions. However, on closer examination of the sequences, two small regions (nucleotides 10,605-10,784 and 10,910-11,116) at the 5' end of *cps19aJ* have >90% DNA sequence identity (97.6% and 93.2%, respectively) to *cps19fJ* (**Fig. 6.6**), which presumably accounts for the Southern hybridisation data obtained previously (**Table 4.1**). Similarly, Cps19aJ and Cps19fJ have only 83.3% amino acid identity, but when their hydropathy profiles are compared they are almost indistinguishable (**Fig. 6.7**). This suggests that the divergence

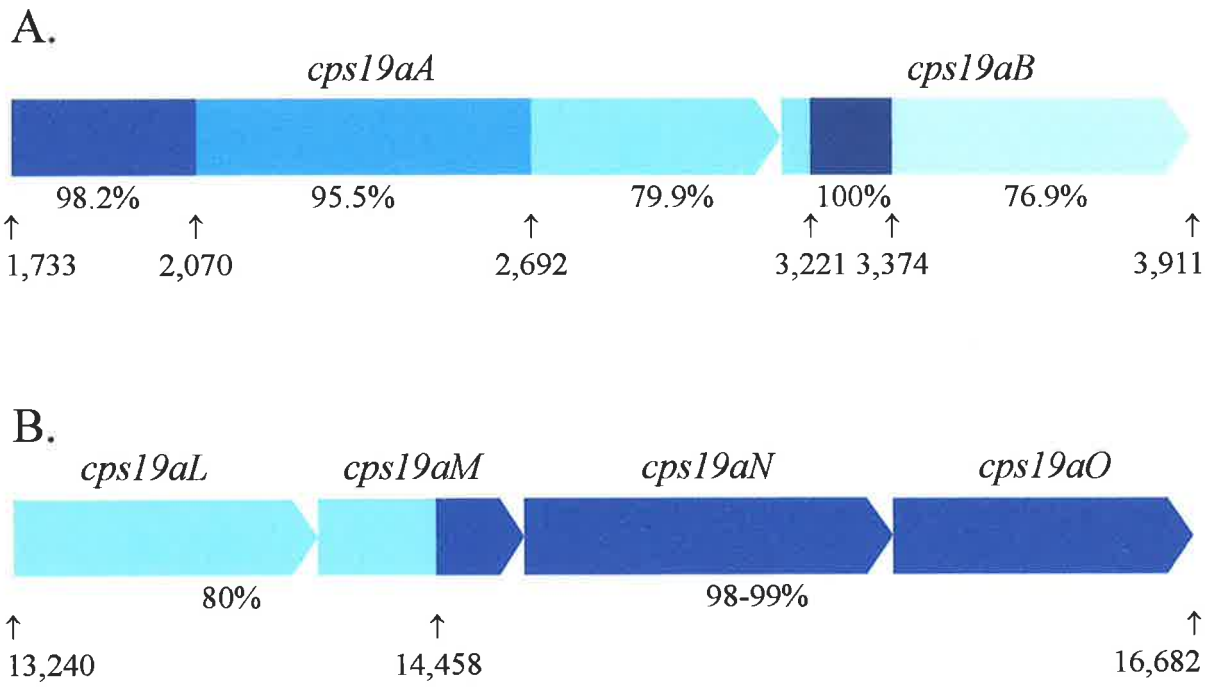


Fig. 6.5. Diagrammatic representation of the similarity of the *cps19a* locus to *cps19f* in the regions where recombination between the loci may have occurred. A. The *cps19aAB* region. There are several possible recombination points in this region of the locus. **B. The *cps19aL-O* region.** There is a single cross-over point within *cps19aM*. Increasing similarity is represented by progressively darker shades of blue and the % identity is shown under the individual coloured regions. The arrows indicate the points of divergence and the number below the arrow corresponds to the nucleotide number in the *cps19a* sequence as shown in Appendix III.

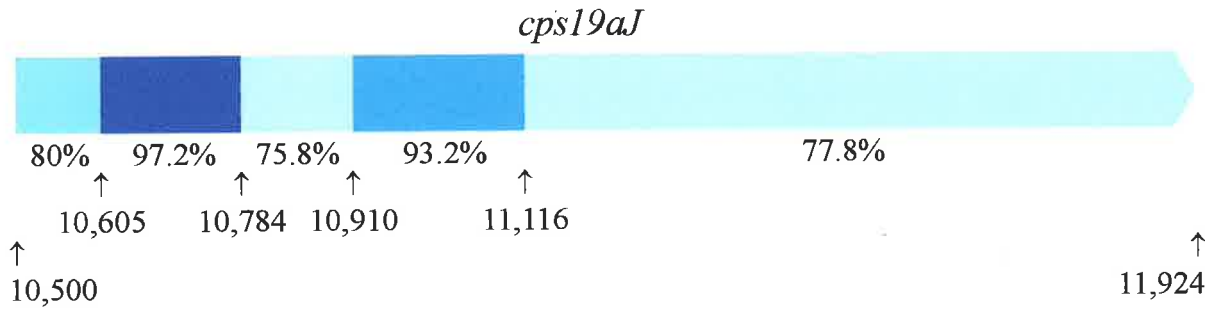


Fig. 6.6. Diagrammatic representation of the similarity between *cps19aJ* and *cps19fJ* sequences. Increasing similarity of *cps19aJ* to *cps19fJ* is represented by progressively darker shades of blue and the % identity is shown under the individual coloured regions. The arrows indicate the points of divergence and the number below the arrow corresponds to the nucleotide number of the *cps19a* sequence as shown in Appendix III.

between these two sequences has resulted in conservative amino acid substitutions, which do not impact upon the function of Cps19aJ or Cps19fJ, both of which transport the same

trisaccharide repeat unit across the membrane. It is also tempting to speculate that the highly conserved regions (with >90% identity) may be important for the function of these proteins. The first conserved region incorporates the second membrane-spanning domain and the following hydrophilic domain; and the second conserved region includes a small hydrophilic domain as well as the fifth and sixth membrane-spanning domains (**Fig. 6.7**).

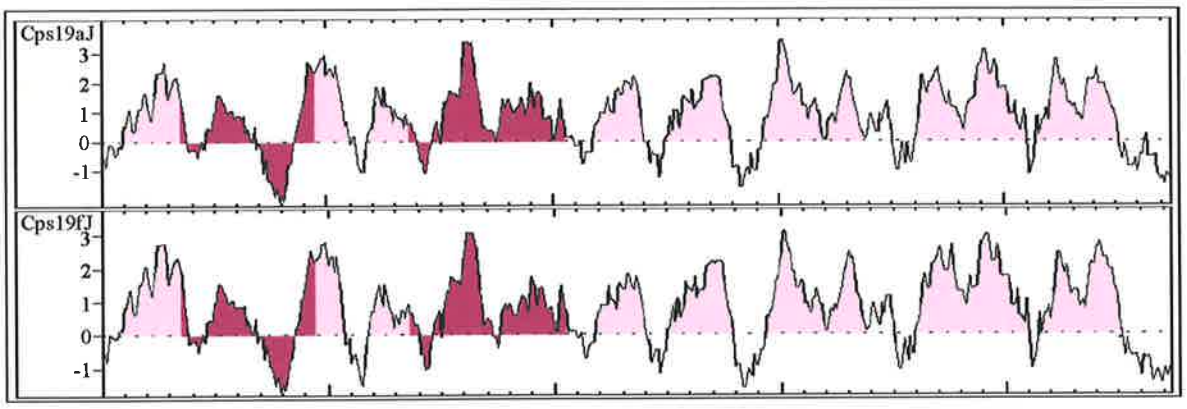


Fig. 6.7. Hydropathy profiles of Cps19aJ and Cps19fJ showing the conserved regions. The hydrophobic domains are shown in pink except in the highly conserved regions which are shown in maroon. Hydropathy plots were generated by the method of Kyte and Doolittle (1982) and aligned using PROFILEGRAPH (Hofmann and Stöffel, 1989). Positive numbers on the Y-axis indicate hydrophobic regions. The position of every 10th amino acid is marked on each X-axis.

The putative polysaccharide polymerases, Cps19aI and Cps19fI, are predicted to form different glycosidic linkages in type 19A ($\alpha1\rightarrow3$) and type 19F ($\alpha1\rightarrow2$) CPS, respectively. These two proteins are 80.7% identical and their amino acid sequences were examined to identify any potentially significant differences between them. Their hydropathy profiles are almost identical as shown in **Fig. 6.8**. A cluster of non-conservative amino acid substitutions were located in the region between amino acids 290 and 320 of Cps19fI and Cps19aI (**Fig. 6.8**). No such clustering of non-conservative amino acid substitutions was observed when comparing either Cps19fH/Cps19aH or Cps19fJ/Cps19aJ (results not shown). This region is predicted to be on the outer surface of the cytoplasmic membrane, based on the topology of the O-antigen polymerase (Rfc) from

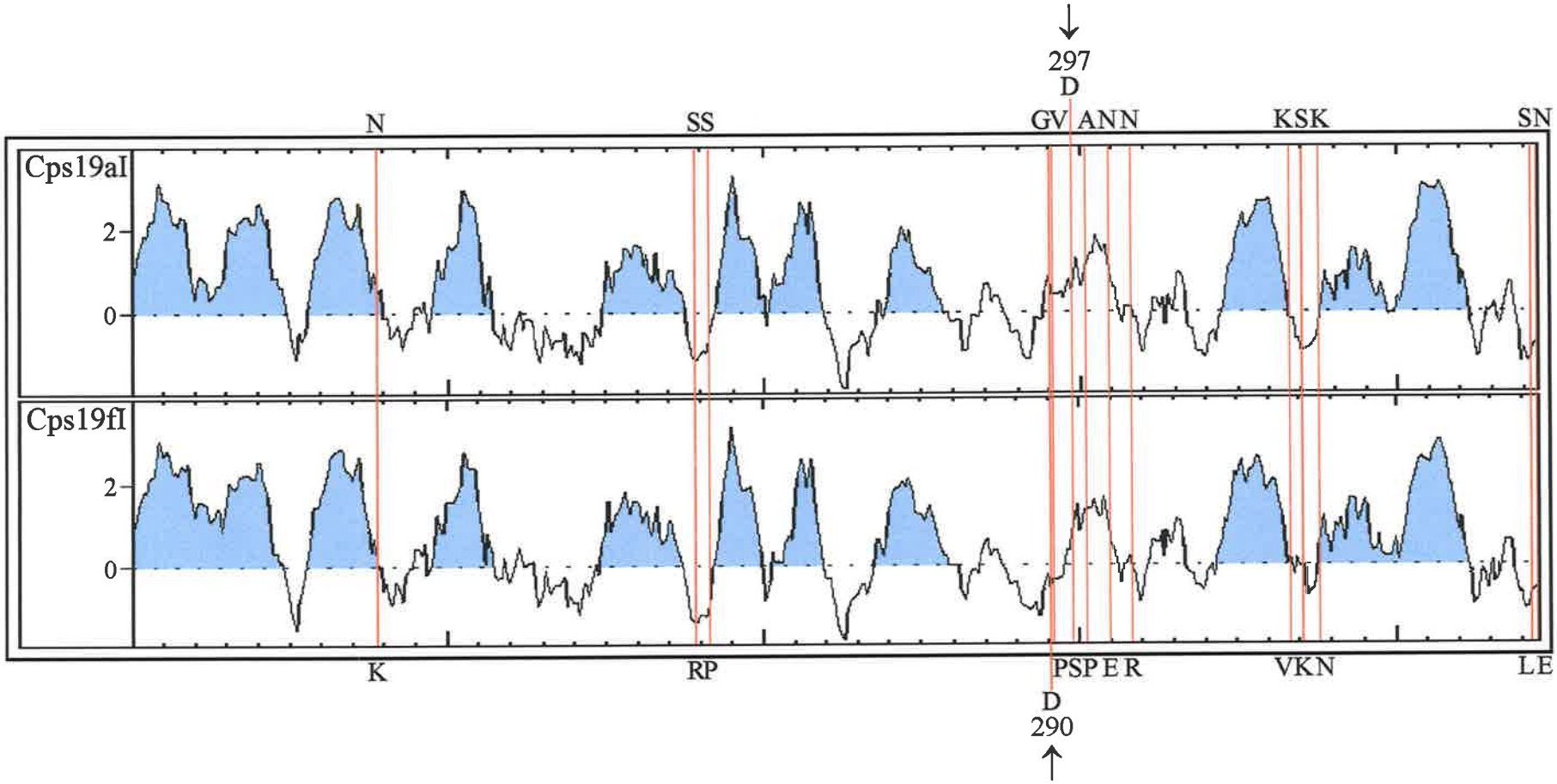


Fig. 6.8. Comparison of the hydropobicity profiles of Cps19aI and Cps19fI. The hydropathy profiles of Cps19aI and Cps19fI were generated by the method of Kyte and Doolittle (1982) and aligned using PROFILEGRAPH (Hofmann and Stöffel, 1989). Positive numbers on the Y-axis indicate hydrophobic regions and putative membrane-spanning domains are shaded. The position of every 10th amino acid is marked on each X-axis. The red lines indicate the position of non-conserved amino acid changes with the amino acid indicated in single letter code above the line for Cps19aI and below the line for Cps19fI. The position of the Aspartate (D) residue is indicated by arrows.

S. flexneri (Daniels *et al.*, 1998). A comparison of the hydropathy profiles of Cps19fI and *S. flexneri* Rfc was shown previously (**Fig. 3.5**). The CPS repeat units are predicted to be transported across the cytoplasmic membrane prior to polymerisation (section 1.8.2). Thus, the external location of the non-conserved regions in Cps19aI and Cps19fI is consistent with that of the putative catalytic site in these proteins. Of particular interest is the single aspartate (D) residue present in this region of both proteins, although its position in not conserved. Aspartate residues are predicted to be required for the formation of glycosidic linkages, and have been identified in the catalytic sites of β–glycosyl transferases (Saxena *et al.*, 1995). As shown in Fig. **6.8**, the position of the aspartate residue varies by seven amino acids between the two proteins. This is equivalent to two turns of an α-helix and would position the two aspartate residues on the same side of an α-helix in both proteins, but in a different spacial position, perhaps facilitating formation of the altered glycosidic linkage formed by these two proteins.

6.2.3 Comparison of the *cps19a* loci from *S. pneumoniae* strains 19A1 and 19A2

The 19A2 PCR product obtained using primers J9 and J36, which amplified the 3' region of the *cps19a* locus encoding the dTDP-Rha biosynthesis genes (*cps19L-O*), was smaller in size to that from both Rx1-19F and 19A1 (**Fig. 6.3**). To identify the deletion present in 19A2, this PCR product was sequenced and this region of the locus (designated $cps19a_2$) is shown in **Appendix IV**.

Analysis of the sequence identified a gene rearrangement in the 3' region of the $cps19a_2$ locus, as well as deletion of 1.4 kb of DNA between the end of *cps19aO* and the start of *aliA*, as shown in **Fig. 6.9**. The first 3,347 nucleotides of the $cps19a_2$ sequence have 99.8% identity to *cps19a*, followed by 1,185 nucleotides with 80% identity to *cps19a*

and 84% identity to *cps19f*. The remainder of the sequence then diverges until the final 94 nucleotides which are 90% identical to the same region in *cps19a*. The conserved regions contain the genes *cps19a$_2$JKLMN*, with an approximate recombination point 120 nucleotides from the end of *cps19a$_2$M*. The next 1.1 kb of DNA contains an inverted copy of *cps19a$_2$O* (**Fig. 6.9**) with 76.4% identity to *cps19aO* and *cps19fO*. A potential promoter was identified upstream of *cps19a$_2$O* in the same region (but on the opposite DNA strand) as that for *aliA* (**Fig 6.10**). There are 61 nucleotides between the stop codons of *cps19a$_2$N* and *cps19a$_2$O* and a stemmed-loop structure which could be a transcription terminator (ΔG=-30.5 kcal/mole) was identified in this region (**Fig 6.10**).

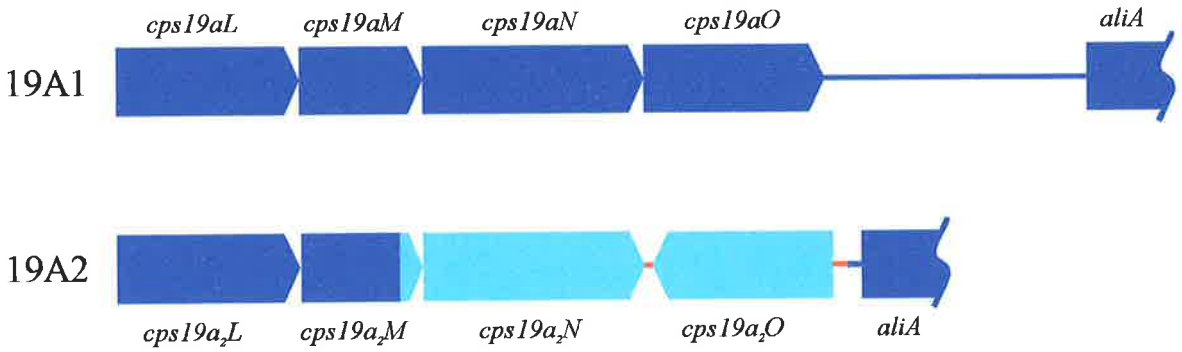


Fig. 6.9. Diagrammatic representation of the *cps19aL-aliA* region of *cps19a* and *cps19a$_2$*. The dark blue region of *cps19a$_2$* is >90% identical to the same region in *cps19a* and the pale blue region exhibits 75-80% identity to the equivalent regions in *cps19a*. The red line indicates small non-coding regions in *cps19a$_2$* with no similarity to *cps19a*.

6.2.4 Serotype specificity of the *cps19a* genes

The relationship between the *cps19aC*, *D*, and *E* genes and *cps* loci of other *S. pneumoniae* serotypes was examined by Southern hybridisation. The distribution of *cps19aA* and *B* genes was not investigated because previous Southern hybridisation data

```
                 ←aliA                                    ←P aliA
          K   S   S   K   M                     -10                  -35
           CAAAACTTGAAAAGTAGTAGAAGAGGAAATTTTGAAAGAGAGATAATATCTGAGAAAAGTCTCTCTACAGTTATATTAAAAGACT
cps19a2    GTTTTGAACTTTTCATCATCTTCTCCTTTAAAACTTTCTCTCTATTATAGACTCTTTTCAGAGAGATGTCAATATAATTTTCTGA
                                                            -35                  -10
                                                                          P cps19a2O→

           TAAAAAATT
cps19a2    ATTTTTTAATTAATTTGGTAAATTCTGAAAAATACTTTTAAGGATATTCTAAAGTAATAACTATTCGGCTCTCTTGCGCTTCACT

cps19a2    AACAAAGTTTATACCAAAAAAGGGTTCACTAAAAACCGATTTTTCTATATTTTTGATATAATTAAGTATAAGAACTATCAAGGAG

               cps19a2O→
               M  I  L  I  T  G  A  N  G  Q  L  G  T  E  L  R  Y  L  L  N  E  R  N  V  D  Y  V
cps19a2    AAAAATATGATCTTAATTACAGGTGCAAATGGTCAACTTGGTACCGAACTACGTTATTTACTAAATGAGCGAAATGTGGACTATG

             A  V  D  V  T  E  M  D  I  T  N  S  E  M  V  E  K  V  F  A  E  V  K  P  T  L  V  Y
cps19a2    TTGCAGTGGATGTGACCGAAATGGATATCACTAATTCTGAAATGGTTGAAAAAGTCTTTGCTGAGGTTAAGCCAACTCTAGTCTA

            H  C  A  A  Y  T  A  V  D  A  A  E  D  E  G  K  E  L  D  F  A  I  N  V  T  G  T  E
cps19a2    TCATTGTGCAGCCTACACTGCTGTTGATGCAGCCGAAGATGAAGGGAAAGAACTGGATTTTGCCATCAACGTAACTGGAACTGAA

           N  V  A  K  G  S  E  K  Y  G  A  I  L  V  Y  I  S  T  D  Y  V  F  D  G  K  K  P  V  G
cps19a2    AATGTAGCCAAGGGCTCTGAGAAATACGGAGCTATTCTGGTCTATATCTCAACAGACTACGTCTTTGATGGAAAGAAACCAGTAG

             Q  E  W  E  V  D  D  I  P  D  P  Q  T  E  Y  G  R  T  K  R  M  G  E  E  L  V  E  N
cps19a2    GACAAGAGTGGGAAGTCGATGACATACCTGATCCGCAAACAGAGTACGGACGTACCAAGCGAATGGGAGAAGAACTTGTTGAGAA

            L  T  S  Q  H  Y  I  I  R  T  A  W  V  F  G  N  Y  G  R  N  F  V  F  T  M  Q  S  L
cps19a2    CCTTACGTCACAGCATTACATCATTCGTACTGCTTGGGTTTTTGGAAATTATGGAAGGAACTTTGTCTTTACCATGCAAAGCCTA

           A  E  T  H  K  T  L  T  V  V  S  D  Q  Y  G  R  P  T  W  T  R  T  L  A  E  F  M  T  Y
cps19a2    GCCGAAACCCATAAAACGCTCACAGTGGTCAGCGACCAATACGGTCGCCCAACTTGGACACGGACCTTGGCCGAGTTCATGACCT

             V  T  E  N  Q  K  E  Y  G  Y  Y  H  L  S  N  D  A  A  E  D  T  T  W  Y  D  F  A  S
cps19a2    ATGTGACTGAAAATCAAAAAGAGTATGGCTACTATCACTTATCGAATGACGCAGCTGAGGACACCACTTGGTATGACTTTGCTAG

            E  I  L  K  E  S  D  V  E  L  L  P  V  D  S  S  K  F  P  A  K  A  K  R  P  L  N  S
cps19a2    CGAGATTCTTAAGGAGAGTGATGTTGAACTATTGCCAGTAGACTCTAGCAAGTTCCCCGCTAAGGCTAAACGACCTCTCAACTCA

           T  M  S  L  A  K  A  K  A  T  G  F  V  I  P  T  W  Q  D  A  L  K  E  F  Y  K  Q  E  V
cps19a2    ACTATGAGTTTGGCAAAGGCAAAGGCAACAGGTTTCGTCATCCCAACCTGGCAAGATGCCCTTAAAGAGTTTTATAAACAAGAAG

             R  K  *
cps19a2    TAAGAAAATAAAAGCTTACTACATACTAGAAAAGCAAGAGACCTCGAGGTCTCTTGCTTTTTTAGTATCTTATTATTTAATCACT
                                 ———— ——————————→←—————————— ————                  *  K  I  V

cps19a2    TGTTGCGTTTTAGCATAGTTAGCCTCTACTGCTTCTTTTTCAGATTTCCACCAGTCTTGGTTATCTGTGTACCACTTGATGGTCT
              ACGCAAAATCGTATCAATCGGAGATGACGAAGAAAAAGTCTAAAGGTGGTCAGAACCAATAGACACATGGTGAACTACCAGA
            Q  Q  T  K  A  Y  N  A  E  V  A  E  K  E  S  K  W  W  D  Q  N  D  T  Y  W  K  I  T

cps19a2    CTTTGAGGCCTGCTTCAAAGTTGGTAAACTCTGGCTTCCACCCCAACTCATCACGGAGCTTGCTAGCATCAATCGCATAACGCAA
           GAAACTCCGGACGAAGTTTCAACCATTTGAGACCGAAGGTGGGGTTGAGTAGTGCCTCGAACGATCGTAGTTAGCGTATTGCGTT
           E  K  L  G  A  E  F  N  T  F  E  P  K  W  G  L  E  D  R  L  K  S  A  D  I  A  Y  R  L

cps19a2    GTCGTGACCAGCTCGATCGGTCACATGGTCATAAGCGTCAGCGGGTTGCCCCATCTCCTTGAGAATGAGCTCTAACACTTCCTTG
           CAGCACTGGTCGAGCTAGCCAGTGTACCAGTATTCGCAGTCGCCCAACGGGGTAGAGGAACTCTTACTCGAGATTGTGAAGGAAC
             D  H  G  A  R  D  T  V  H  D  Y  A  D  A  P  Q  G  M  E  K  L  I  L  E  L  V  E  K

cps19a2    TTGTTCTTCTCACCGTCAGCACCAATCA
           AACAAGAAGAGTGGCAGTCGTGGTTAGT
            N  N  K  E  G  D  A  G  I
                   ←cps19a2N...
```

Fig. 6.10. The sequence of *cps19a₂*, from the 3' end of *cps19a₂N* to the start of aliA. The sequence shown is the complementary strand of nucleotides 4,360-5,832 of the *cps19a₂* sequence (**Appendix IV**). The double stranded sequence at the 5' and 3' ends represents the regions with homology to *cps19a*. The amino acids are shown above or below the first letter of the codon and the direction of translation is as indicated. The stemmed-loop structure is indicated by arrows below the sequence. The putative promoter regions (-10 and -35 sequences) for *aliA* and *cps19a₂O* are also shown.

(section 4.2.1) indicated that homologues of the closely related *cps19fA* and *B* genes were present in all the serotypes tested. DNA fragments corresponding to the individual genes were isolated by PCR amplification using specific primers and restriction enzyme digestion (when necessary) and subsequent purification after agarose gel electrophoresis (**Table 6.2**). They were then labelled with DIG (section 2.7.1) and used to probe *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. Hybridisation and washing conditions were at high stringency (as described in section 2.7.2.2).

Table 6.2 DNA fragments used as gene-specific probes

Probe	Fragment used as probe:
cps19aC	400 bp *Pst*I fragment from J92-J95 PCR product (nt 3,932-4,356 of *cps19a* sequence)
cps19aD	500 bp *Pst*I fragment from J93-J94 PCR product (nt 4,356-4,820 of *cps19a* sequence)
cps19aE	900 bp ClaI fragment from J87-J94 PCR product (nt 5,594-6,480 of *cps19a* sequence)
cps19aK	1.4 kb J70-J72 PCR product (nt 12,114-13,540 of *cps19a* sequence)

All primers used are described in **Table 2.4**. nt, nucleotide position.

The results for the hybridisation of both 19A and 19F genes are summarised in **Table 6.3**. The most remarkable feature of this table is that all the serotypes tested contained high stringency homologues of either *cps19fC-E* or *cps19aC-E*, except types 3 and 4 which do not have a high stringency homologue of either *cps19fE* or *cps19aE* (the gene which encodes the glucosyl transferase which adds Glc-1-phosphate to the lipid carrier). The absence of a *cpsE* homologue in types 3 and 4 is not surprising because the type 4 CPS does not contain Glc and the type 3 CPS is synthesised via a processive transferase as described in sections 1.8.4 and 8.4.2. Type 4 also contains a hybrid *cpsC* gene as previously described in chapter 4. The Southern hybridisation data presented here

Table 6.3. Hybridization of *cps19fC-E* and *cps19aC-E* genes with other pneumococcal serotypes.

	DIG labelled DNA probes					
Serotype	*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	+	+	+	-	-	-

Blue shaded '+' denotes hybridization with the gene-specific probes; - denotes no hybridization. The yellow shading represents no hybridization to either *cps19f* or *cps19a* probes.

suggests that the distinction between class I and class II genes identified for the *cpsC* gene in chapter 4, extends beyond *cpsC* to include *cpsD* and *cpsE*.

The presence of *cps19aK* homologues in serotypes 4, 9N, 9V and 12 was also examined, using a DIG-labelled *cps19aK* probe (described in **Table 6.2**) and low stringency conditions (as described in section 2.7.2.3). These serotypes are expected to contain a functional homologue to *cps19K* in their *cps* loci because Man*N*Ac is a constituent sugar in their CPS. Only type 19A chromosomal DNA had hybridised with the *cps19fK* gene probe with low stringency conditions (section 4.2.1). When *cps19aK* was used as a probe, type 12 and the other members of serogroup 19 (19F, 19B and 19C) hybridised, suggesting that the gene(s) present in the serotypes 4, 9N and 9V are not closely

related to either *cps19aK* or *cps19fK*. Indeed, examination of the *cps4* sequence (section 9.3.4) reveals that *cps4L*, which probably encodes the UDP-Glc*N*Ac-2-epimerase required for type 4 CPS biosynthesis, has 67.8% identity to the *cps19aK* gene and 65.9% identity to *cps19fK*, compared with 85.2% identity between *cps19aK* and *cps19fK*. This suggests that the UDP-Glc*N*Ac-2-epimerase gene may have been derived from distinct sources in different serotypes.

The serotype specificities of *cps19aG, H, I* and *J* were not reinvestigated as the homologues in *cps19f* were all serotype 19F and/or serogroup 19 specific.

6.2.5 Capsule transformation from type 19F to 19A

To confirm that the *cps19a* locus was sufficient for type 19A CPS biosynthesis, a 16.5 kb PCR product from the 5' end of *cps19aA* to the 5' end of *aliA*, was amplified using the primers CPS5' and J36 (**Table 2.4**). This was used to transform Rx1-19F-I, an unencapsulated, Ery-resistant derivative of Rx1-19F in which the *cps19fI* gene had been disrupted by insertion-duplication mutagenesis using pVA891, as described in section 3.2.4. Several smooth transformants were checked for Ery sensitivity, indicating loss of the pVA891 sequence. Southern hybridisation analysis was used to confirm the absence of both pVA891 and the *cps19fI* gene, and to determine if the *cps19aC, I* and *K* genes were present in three individual transformants (**Fig. 6.11**).

The production of a type 19A capsule by these three smooth transformants, designated Rx1-19A.1-3, was then confirmed by quellung reaction (section 2.3). The cross- over points between the *cps19f* locus and the type 19A PCR product were then identified by sequencing the regions where recombination was predicted to have occurred, based on the Southern hybridisation data in **Fig. 6.11**. The actual sequence data are shown in **Appendix V** and a diagram indicating the recombination points is shown in **Fig. 6.12**.

Two transformants (Rx1-19A.1 and 3) were similar, resulting in the exchange of a large region of the *cps19f* locus, from *cps19fG* to *cps19fN* (including *cps19fO* in Rx1-19A.3). On the other hand, Rx1-19A.2 is derived from exchange of a much smaller region of the *cps19f* locus, involving only *cps19H* and *cps19I* (**Fig. 6.12**). The *cps19aH* gene has 90.8% nucleotide identity to *cps19fH*, and the encoded highly conserved putative rhamnosyl transferases (95.2% amino acid identity) are predicted to be functionally identical in both type 19F and 19A CPS biosynthesis. The *cps19aI* and *cps19fI* genes are less conserved, with only 78.5% nucleotide identity, and the encoded putative polysaccharide polymerases (80.7% amino acid identity) are predicted to form different glycosidic linkages. Thus, this data shows that it is possible to alter capsule production from type 19F to type 19A by replacing only two genes in the *cps19f* locus, and that the presence of the *cps19aI* gene probably determines the 19A serotype.

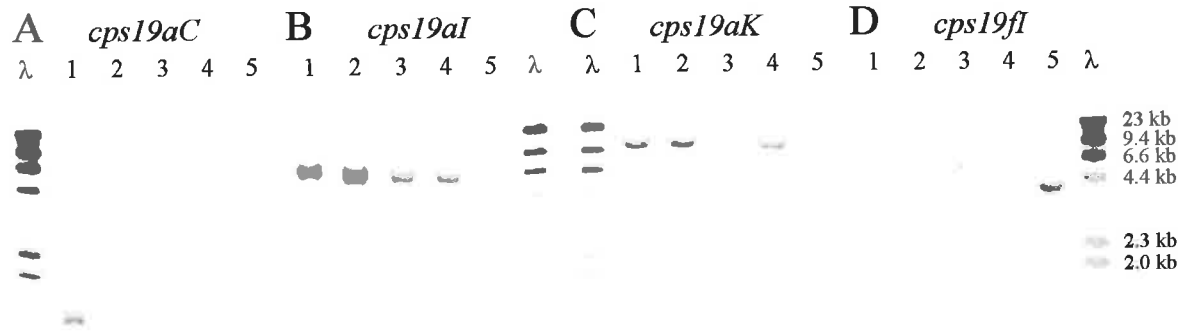


Fig. 6.11. Southern hybridisation analysis of the three Rx1-19A transformants. 19A1 (1), Rx1-19A.1 (2), Rx1-19A.2 (3), Rx1-19A.3 (4) and Rx1-19F (5) were probed with DIG-labelled probes specific for *cps19aC* (A), *cps19aI* (B), *cps19aK* (C) and *cps19fI* (D).

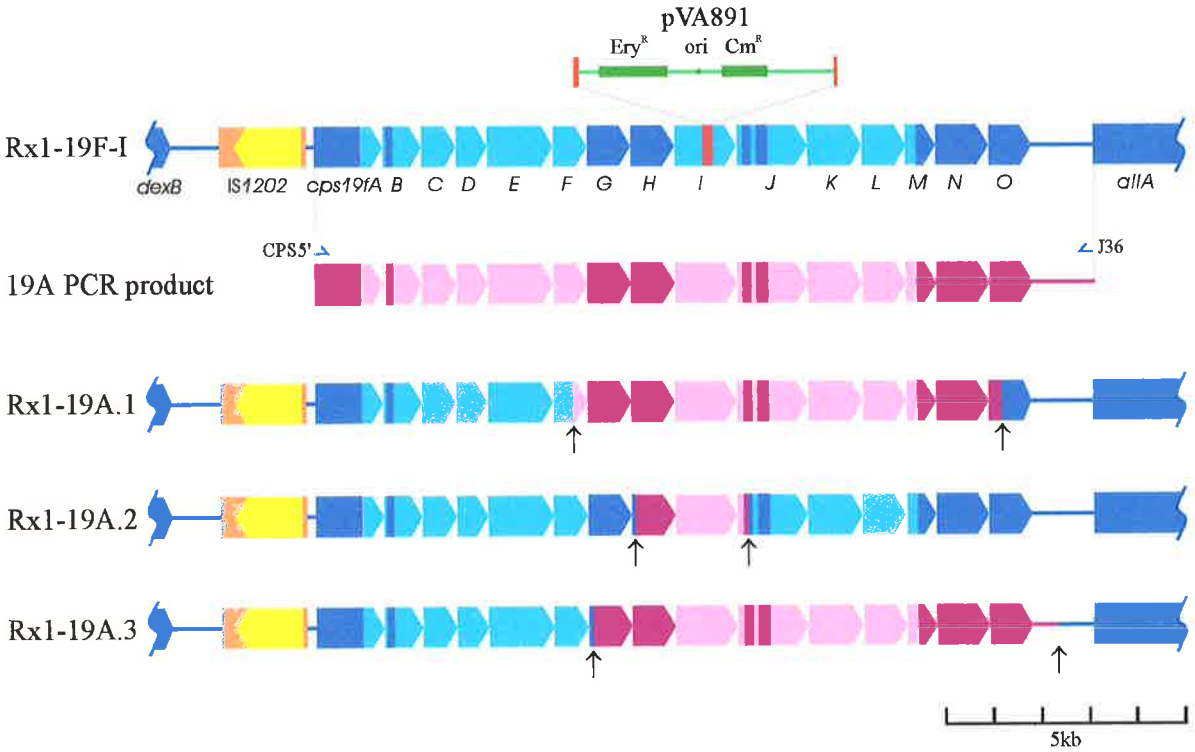


Fig. 6.12. Diagrammatic representation of the cross-over points in the Rx1-19A transformants. The *cps19f* locus is shown in blue and the *cps19a* locus is shown in pink. The dark colours represent regions with >90% DNA sequence identity and the light colours indicate 70-80% identity. The arrows indicate the points from which the sequence becomes *cps19a*-specific.

6.2.6 Sequence variation in the 5' intergenic region of serogroup 19

The 5' intergenic region of both type 19A isolates (19A1 and 19A2), as well as that for type 19F, 19B and 19C isolates were PCR amplified using the DEXB and CPSA2 primers (**Table 2.4**). The sizes of the PCR products varied between strains (**Fig. 6.13**), but those from 19A1 and 19B appeared to be the same size. The PCR products from 19A2 strain and the type 19B and 19C isolates were sequenced using specific primers and the 5' intergenic regions were compared as shown in **Fig. 6.14**.

Interestingly, the 5' intergenic regions of 19A1, 19B and 19C are all almost identical with three conserved deletions compared to 19F (**Fig. 6.14**). These three deletions remove all but 150 nucleotides of the 1-kb intergenic region between *dexB* and IS*1202*, as well as the 3' end of IS*1202* (up to the stop codon of the putative transposase).

Another mutation at nucleotide position 2,151 (as depicted in **Fig. 6.14**) introduces a stop codon which interrupts the putative transposase in 19A1, 19B and 19C. The larger size of the PCR product obtained from 19C is due to the presence of an additional IS element,

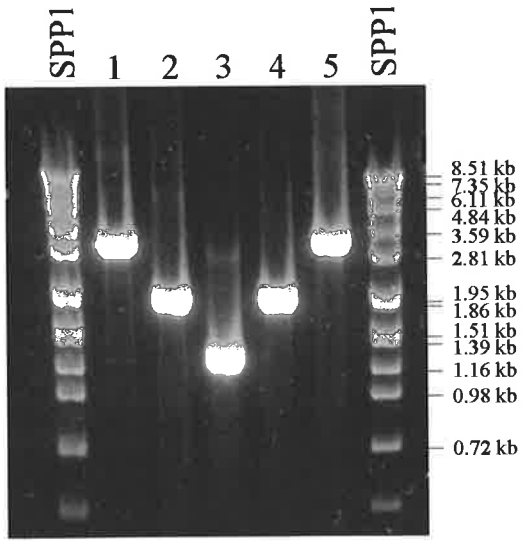


Fig. 6.13. PCR products of the 5' intergenic region. The PCR products obtained, using DEXB/CPSA2 primers, from Rx1-19F (1), 19A1 (2), 19A2 (3), 19B (4) and 19C (5) were electrophoresed in a 0.8% agarose gel and stained with ethidium bromide. *Eco*RI-digested SPP1 bacteriophage DNA was used as a size marker (section 2.6.2) and the approximate sizes of the DNA fragments are indicated on the right side of the figure.

which will be described in chapter 7. This IS element inserted into the inverted repeat of IS*1202*, adjacent to the *cps19c* locus (as shown in **Fig. 6.14**).

Analysis of the derived sequences indicated that the 5' intergenic region of 19A2 is almost identical to that of 19F, except that it does not contain a copy of IS*1202* in the 5' intergenic region, although Southern hybridisation data has previously shown that this type 19A strain does contain a copy of IS*1202* in its chromosome (Morona *et al.*, 1994a). When PCR products from the 5' intergenic region from the six Australian type 19A isolates were examined by electrophoresis, they were all the same size as that from 19A2 (**Fig. 6.15**).

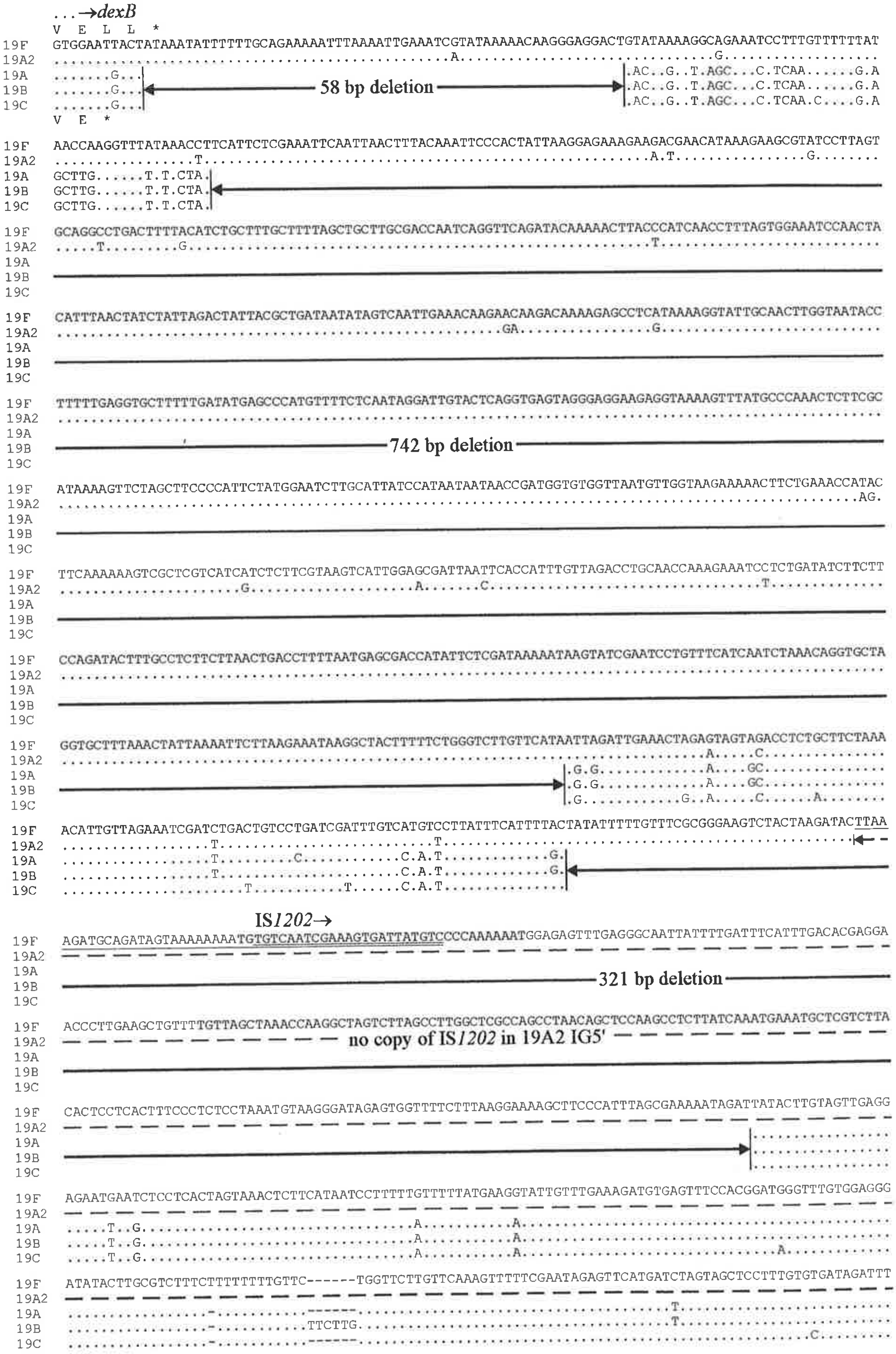


Fig. 6.14. **Continued overleaf.**

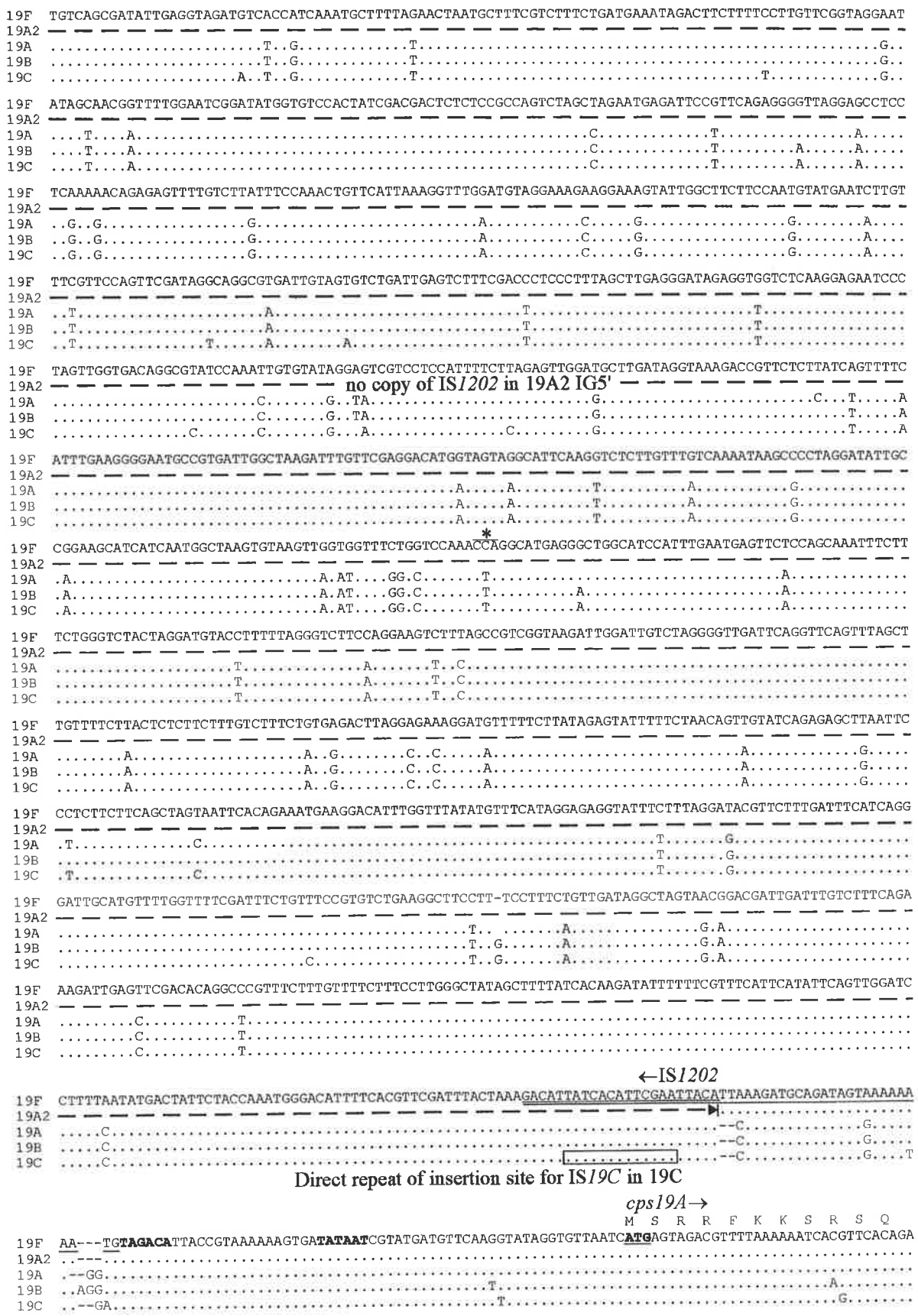


Fig. 6.14. Comparison of the 5' intergenic sequences of *S. pneumoniae* strains Rx1-19F, 19A1, 19A2, 19B and 19C. The complete 5' intergenic sequences are shown (100 nucleotides per line) and are available under accession numbers U09239, AF094575, AF105112, AF105114, and AF105115, respectively. The stop codon in the truncated IS*1202* transposase is indicated with an underlined asterisk. The -35 and -10 promoter regions are shown in bold and the start codon of *cpsA* is shown in bold and underlined. The amino acid translation of the 3' end of *dexB* and the 5' end of *cps19A* are as indicated. The direct repeat and inverted repeats associated with IS1202 are underlined and double underlined, respectively.

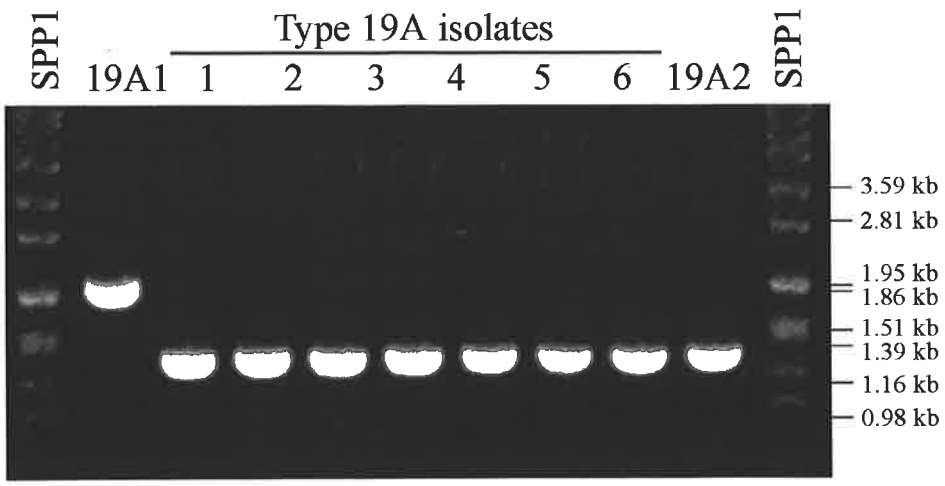


Fig. 6.15. Comparison of PCR products from the 5' intergenic regions from several Australian type 19A isolates. The PCR products were electrophoresed in a 0.8% agarose gel and stained with ethidium bromide. *Eco*RI-digested SPP1 bacteriophage DNA was used as a size marker (section 2.6.2) and the approximate sizes of the DNA fragments are indicated on the right side of the figure.

6.3 Conclusions

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, large regions of the two loci share only 70-80% nucleotide sequence identity, suggesting that they either diverged a long time ago or that portions of the loci have separate origins. Some regions within the *cps19a* locus do have greater than 90% identity to *cps19f*, which may be a consequence of either recombination between the two loci, or perhaps due to a requirement for a higher degree of conservation in regions encoding functionally important domains. Transformation studies in section 6.2.4 have demonstrated that with the exception of *cps19aI*, all *cps19a* genes are functionally homologous to their *cps19f* counterparts. Genetic exchange of a DNA fragment containing just two genes, the *cps19aI* gene (which encodes the polysaccharide polymerase) and the highly conserved *cps19aH* gene (which encodes the rhamnosyl transferase), was sufficient to convert a strain containing an interrupted *cps19f* locus to type

19A. This is consistent with the fact that according to the structure proposed by Lee and Fraser (1980) type 19A CPS differs from type 19F only by the type of glycosidic linkage between identical trisaccharide repeat units. The studies in this chapter have shown that the genes which are present in the *cps19a* locus are sufficient for type 19A CPS biosynthesis and hence, the biosynthetic pathway for type 19A CPS is essentially identical to that proposed for type 19F CPS in **Fig. 3.16**.

Of particular interest, are the two closely related polysaccharide polymerase genes (*cps19aI* and *cps19fI*). The proteins encoded by these genes have almost identical hydrophobicity profiles. A cluster of non-conservative amino acid changes between amino acids 290-320 has been identified in a hydrophilic region of the proteins and is a likely candidate for the putative catalytic domain. These two proteins provide a perfect model for further characterisation of this catalytic domain. Exchange of this small non-conserved region between the two proteins, in which the position of the potentially important aspartate residues differ by seven amino acids, could potentially alter the glycosidic linkage formed between the repeat units and thus the CPS serotype expressed. Also site-directed mutagenesis, altering the aspartate residues, might be predicted to abrogate the function of the polymerase. It would then be possible to investigate restoring the function of the polymerase, with altered specificity, by altering the position of the aspartate residue.

No additional genes, which might be involved in type 19A CPS biosynthesis, were identified either in or adjacent to the *cps19a* locus. Thus, the extra genes required to synthesise the side-chains proposed in the alternative type 19A CPS structure proposed by Lee *et al.* (1987), as shown in **Fig. 6.1**, must be located elsewhere on the *S. pneumoniae* chromosome. It is not known if these extra putative genes are present in all pneumococci or are specific to type 19A strains. The three Rx1-19A transformants described in section 6.2.4 synthesise type 19A CPS as judged by quellung reaction. However, it is not known if

they are capable of synthesising the proposed side-chains. If they are, these extra genes must be present in Rx1 (a rough derivative of a type 2 strain) and are therefore likely to be ubiquitous to all pneumococci. Chemical analysis of the CPS of these transformants grown under conditions reported to result in the biosynthesis of the alternative type 19A CPS (Lee *et al.*, 1987) is required to resolve this issue.

In chapter 4, it was noted that the *cpsC* gene could be divided into two distinct classes (I and II), depending on the degree of similarity to *cps19fC*. Whereas the *cps19fC* gene belonged to class I, the *cps19aC* gene belonged to class II. Southern hybridisation data were obtained using *cps19aC-E* probes and chromosomal DNA from the same pneumococcal serotypes analysed with *cps19fC-E* probes in **Table 4.1**. The results indicated that all three genes were present together, as either class I (*cps19fC-E* homologues) or class II (*cps19aC-E* homologues) in most serotypes tested. The *S. pneumoniae cps* loci, can be divided into class I and class II, on the basis of their *cpsC-E* genes. Using these criteria, the *cps* loci from pneumococcal types 1, 3, 4, 14, 19F, 19B, and 19C are class I and types 2, 19A, 23F, 33F are class II. The presence of *cpsB* homologues in all loci was demonstrated previously (**Table 4.1**), and the point at which the class I and class II sequences diverged was shown to occur within this gene in the serotypes tested (section 4.2.4). However, the precise point of divergence can vary, and type 19A and 19F *cps* loci appear to diverge within the *cpsA* gene (**Fig. 6.5**).

The intergenic regions flanking the *cps* loci from all six Australian *S. pneumoniae* type 19A isolates tested appear to be almost identical to type 19F except that IS*1202* was absent. A type 19F strain which lacks IS*1202* has been previously reported (Morona *et al.*, 1994a). However, two of the type 19A isolates did appear to contain extra DNA at the 3' end of the locus, which has not been investigated further, and may indicate the presence of yet another IS element in the 3' intergenic region. The common occurrence of IS elements

in the intergenic regions flanking the *cps* loci in different *S. pneumoniae* strains is discussed in section 8.4.4. The similarity in the 5' intergenic region between Rx1-19F, 19A2 and the six Australian type 19A isolates is suggestive of a common ancestry for type 19F and 19A. The unexplained rearrangement at the 3' end of the *cps* locus of 19A2, where the *cps19a*$_2$*O* gene has been inverted, is probably a subsequent event. The S. pneumoniae type 19B and 19C strains analysed had an identical series of deletions in the 5' intergenic region, suggesting that they share the same clonal origin. The additional IS element found in this region in 19C could have integrated either before or after the acquisition of the type 19C-specific gene(s). Strain 19A1, which has an identical 5' intergenic region to 19B, may have originated from a recombination event between an ancestral type 19B strain and the *cps19a* locus.

Chapter 7

CHARACTERISATION OF THE *S. PNEUMONIAE* TYPE 19C SPECIFIC *CPS* REGION

7.1 Introduction

The structure of the type 19C CPS is similar to that of type 19B. The only difference is that the type 19C repeat unit has an additional Glc side-chain, as shown in **Fig. 7.1**. Thus the type 19C capsule locus (*cps19c*) would be predicted to contain an extra gene required for the addition of this side-chain. Southern hybridisation analysis with *cps19f* gene specific probes has indicated that the *cps19b* and *cps19c* loci are almost identical with all but two of the genes hybridising to type 19B and 19C DNA (section 4.2.1). These two genes, *cps19fI* and *cps19fJ*, are located together near the middle of the *cps19f* locus and are replaced by the five genes *cps19bP, I, Q, R* and *J* in the *cps19b* locus (section 5.2.1). Homologues of these 19B-specific genes were also shown to be present in type 19C by hybridisation analysis (section 5.2.3). The extra gene required for type 19C biosynthesis may also be located in this region of the *cps19c* locus. Accordingly, the work described in this chapter was aimed at locating and characterising any additional gene(s) required for synthesis of type 19C CPS.

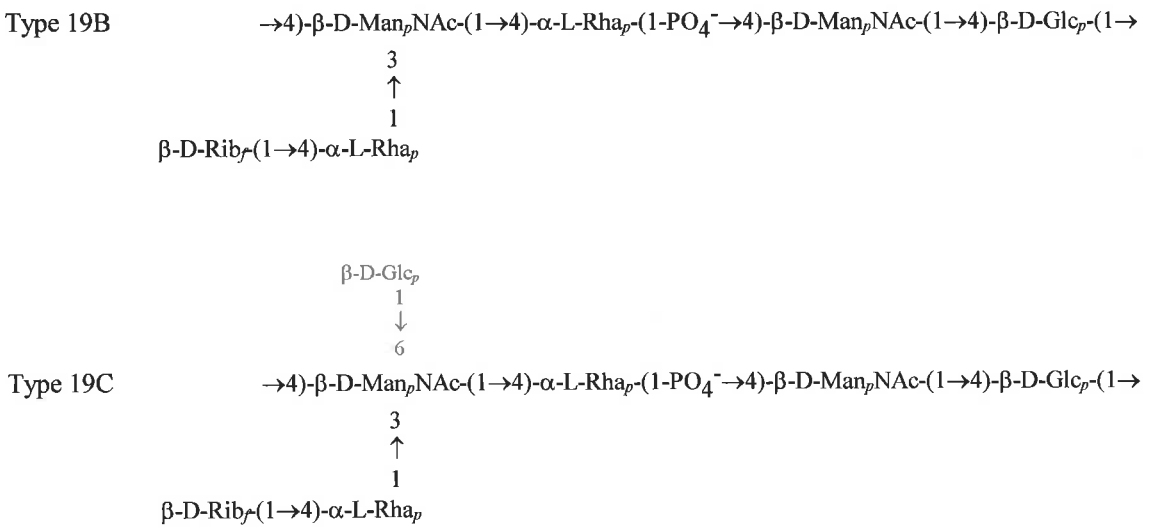


Fig. 7.1. Biological repeat units of pneumococcal type 19B and type 19C capsular polysaccharide. The order of the sugars in the repeat units have been altered compared to the published chemical structures for 19B and 19C (Beynon *et al*, 1991) as it has been determined that glucose is the first sugar in the repeat unit. D-Glu$_p$, glucose; D-Man$_p$*N*Ac, *N*-acetyl mannosamine; L-Rha$_p$, rhamnose; D-Rib$_f$, ribose; PO_4^-, phosphate.

7.2 Results

7.2.1 Isolation of the type 19C specific *cps* region

The exact location of the extra gene expected to be present in type 19C was investigated using LR-PCR as described in section 2.9.3. The PCR products amplified with J5 and J27 (the two primers used to amplify the type 19B-specific region of *cps19b* in section 5.2.1) from both type 19B and 19C were digested with restriction endonucleases and analysed by agarose gel electrophoresis. The type 19C PCR product was found to be approximately 2 kb larger than the equivalent region of *cps19b* (**Fig. 7.2A**). Thus, there is additional DNA in the *cps19c* locus between *cps19cF* and *cps19cL*. Various smaller PCR products were obtained as described in **Table 7.1** and analysed by agarose gel electrophoresis, as shown in **Fig. 7.2B**. The only PCR product which was larger using type 19C rather than type 19B template, was obtained by amplifying the region between *cps19R* and *cps19L* using the primers J49 and J27. As the PCR product amplified from the region

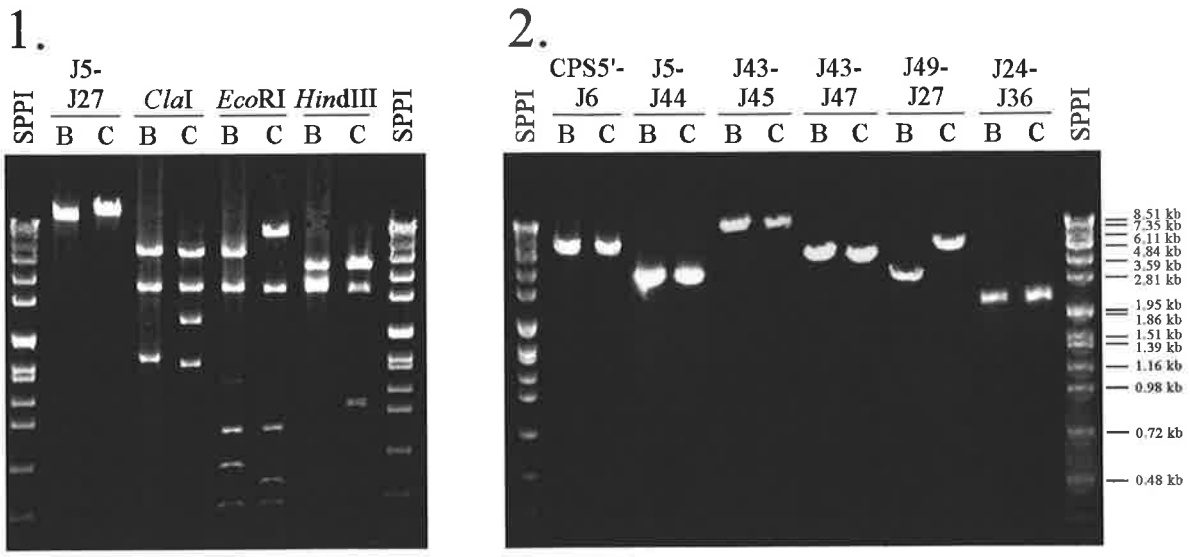


Fig. 7.2. Comparison of PCR products obtained from *S. pneumoniae* types 19B and 19C. PCR products obtained from 19B (B) and 19C (C) using various primers were electrophoresed in a 0.8% agarose gel and stained with ethidium bromide. Panel 1 shows the PCR products obtained using primers J5 and J27 either undigested or restricted with *Cla*I, *Eco*RI or *Hin*dIII prior to electrophoresis. Panel 2 shows undigested PCR products obtained using other primers as indicated. *Eco*RI-digested SPP1 bacteriophage DNA was used as a size marker (section 2.6.2); the approximate sizes of the PCR products are indicated on the left side of the figure.

Table 7.1. PCR products obtained from *cps19b* and *cps19c*.

PCR product[a]	Section of locus amplified	Size of PCR product[b]	
		Type 19B	Type 19C
CPS5'-J6	*cps19A-cps19F*	5.5 kb	5.5 kb
J5-J44	*cps19F-cps19I*	3.5 kb	3.5 kb
J43-J45	*cps19P-cps19K*	6.5 kb	6.5 kb
J43-J47	*cps19P-cps19J*	5 kb	5 kb
J49-J27	*cps19R-cps19L*	4 kb	6 kb
J24-J36	*cps19O-aliA*	2.5 kb	2.5 kb

a. PCR primers are described in **Table 2.4**. b. sizes are approximate.

between *cps19H* and *cps19K* was the same size in both types 19B and 19C, the additional gene must be located between *cps19cK* and *cps19cL* in the *cps19c* locus. The J49-J27 PCR product was purified (section 2.9.4) for further analysis. A map of the 5.3 kb PCR product was obtained, using the restriction enzymes *Bam*HI, *Cla*I, *Hin*dIII, *Nsi*I, *Nde*I and

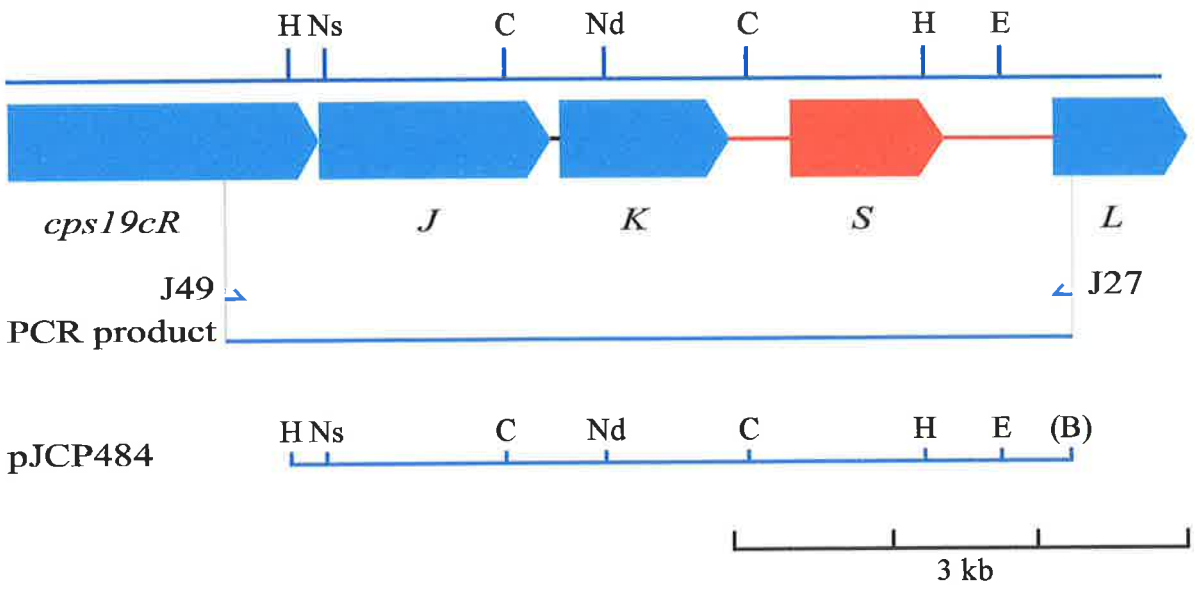


Fig. 7.3. Physical map of part of the *cps19c* locus. Boxed arrows represent potential ORFs. Gene designations are indicated below the map; *cps19cB-S* are abbreviated to *B-S*, respectively. Restriction sites are as follows; B, *Bam*HI; C, *Cla*I; E, *Eco*RI; H, *Hin*dIII; Nd, *Nde*I; Ns, *Nsi*I. The region of DNA subcloned into pBluescript KS+ is shown below the map. The *Bam*HI restriction site is bracketed because it is generated by the J27 primer and not present in the chromosome of *S. pneumoniae* type 19C.

*Eco*RI, and it was sub-cloned into pBluescript KS+, generating pJCP484, as shown in **Fig. 7.3**.

Both strands of the pneumococcal DNA insert, and nested derivatives thereof, were subjected to sequence analysis in order to compile the sequence of the portion of the *cps19c* locus, as shown in **Appendix VI**. Examination of the compiled sequence revealed, as expected, that the first 2.9 kb of sequence at the 5' end has a high degree of similarity to the *cps19b* sequence. This region contains the homologues of *cps19bR*, *cps19bJ* and *cps19bK* (*cps19cR*, *cps19cJ* and *cps19cK*) which exhibited 98.5%, 99.7% and 94.9% identity, respectively. The sequence then diverges (at nucleotide 2,906 of the *cps19c* sequence) just prior to the end of *cps19cK*; the sequence of nucleotides 2,954-3,155 exhibits 74.8% identity to the 5' region of *cps19bL* but does not encode an ORF, and the sequence then diverges as shown in **Fig. 7.4A**. An additional potential ORF, designated *cps19cS* (shown in **Fig. 7.3**), is located between *cps19cK* and *cps19cL* and has a TTG start codon, which is preceded by a ribosome binding site, as shown in **Appendix VI**. The

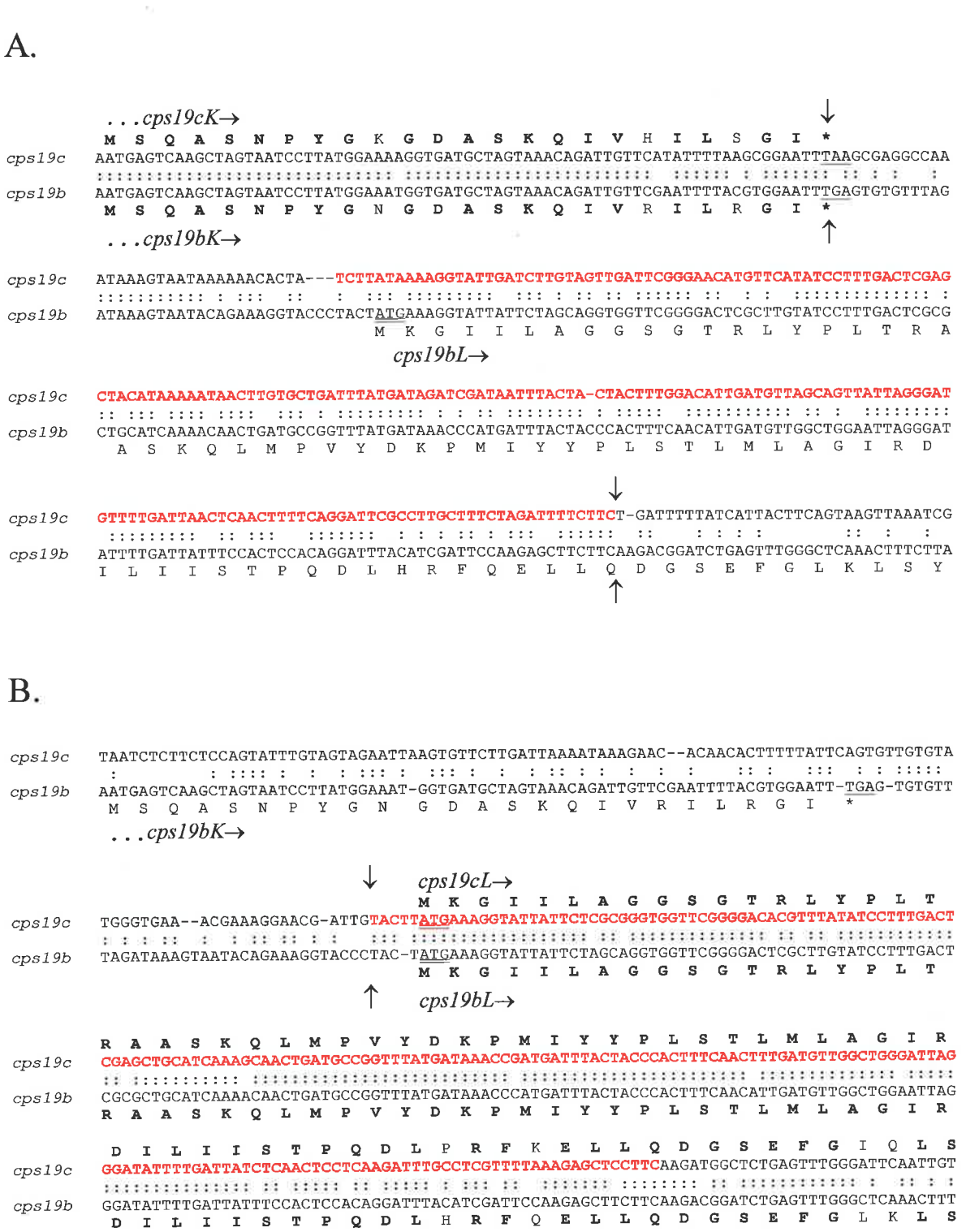


Fig. 7.4. Homology between *cps19b* and *cps19c*. Location of the approximate points of sequence divergence, at the end of *cps19K* (**A**), and immediately upstream of *cps19L* (**B**). Nucleotides 2,850-3,188 and 4,905-5,243 of the *cps19c* sequence are represented in panels A and B, respectively. The *cps19b* sequence (nucleotides 10,255-10,599 of Appendix II) is the same in both panels. The amino acid translation for the depicted ORFs are shown above the sequence for *cps19c* and below the sequence for *cps19b*. The stop codons are underlined and the ATG start codons are double underlined. The repeated sequence in *cps19c* is shown in red. Arrows indicate the approximate points of divergence. Identical amino acids are shown in bold.

closest potential ATG start codon is located 138 nucleotides downstream but it is not preceded by a ribosome binding site. As predicted, the 3' end of the *cps19c* sequence again shows similarity to the *cps19b* sequence, starting from nucleotide 5,017 (**Fig. 7.4B**); this is immediately before the start of the *cps19cL* gene, which has 90.6% identity to *cps19bL*. Thus, the first 200 nucleotides of *cps19cL* are duplicated in the *cps19c* locus, with 77.8% identity between the copies present immediately downstream of *cps19cK* and at the 5' end of *cps19cL*, as shown in **Fig. 7.4**. There are potentially significant intergenic gaps immediately before and after the *cps19cS* gene of 370 and 633 nucleotides, respectively. However, no potential stemmed-loop structures or obvious promoter sequences were found in these intergenic regions.

7.2.2 Characterisation of Cps19cS

The type 19C-specific ORF *cps19cS* is located between *cps19cK* and *cps19cL* in the *cps19c* locus (nucleotides 3,276-4,385) and encodes a putative 43.2 kDa protein containing 343 amino acids. This hydrophilic protein has a hydrophobicity index (according to Kyte and Doolittle [1982]) of -0.23 and a predicted pI of 5.18. The region from the 3' end of *cps19cK* to the 5' end of *cps19cL* has a G+C content of 30.4% increasing slightly to 31.4% for the *cps19cS* coding region. This is lower than the %G+C of the two flanking genes *cps19cK* (35.3%) and *cps19cL* (42.6%).

Database searches with Cps19cS found significant similarity to the C-termini of various glycosyl transferases as shown in **Table 7.2**. The alignment of the conserved C-terminal regions of these proteins is shown in **Fig. 7.5**. Interestingly, one of these glycosyl transferases, CpoA, is possibly involved in teichoic acid biosynthesis in *S. pneumoniae* (Grebe *et al.*, 1997). Cps19cS exhibits 21% identity along its entire length to WaaG (**Table 7.2**), an $\alpha(1\rightarrow3)$ glucosyl transferase involved in LPS core biosynthesis in *E. coli*

and *S. enterica* serovar typhimurium (Heinrichs *et al.*, 1998). Thus, Cps19cS could function as the glucosyl-transferase required for the addition of the β(1→6) linked Glc side-chain to the backbone in type 19C CPS biosynthesis.

Table 7.2. Similarity of Cps19cS to other proteins.

	% Identity[a]								
	Mj RfbU	Aa MtfC	Af Gal	Sa CapM	Ye TrsD	Vc RfbV	Ea AmsD	St WaaG	Sp CpoA
Cps19cS[b]	26	28.5	22.9	32.3	23	26.8	22.2	21	28.7
	[127]	[158]	[223]	[93]	[222]	[142]	[221]	[300]	[94]
MjRfbU[c]	100	30.3	34.2	25.4	26.3	25.4	26.9	30.4	24.3
		[195]	[190]	[244]	[179]	[213]	[119]	[161]	[173]
AaMtfC[d]		100	24.8	23.7	21.6	34.6	23	21.9	19.5
			[311]	[359]	[361]	[373]	[183]	[233]	[246]
AfGal[e]			100	25.1	20.9	24.5	25.6	27.9	21.5
				[175]	[344]	[372]	[203]	[183]	[251]
SaCapM[f]				100	25.6	20.9	17.7	13.9	15.2
					[164]	[349]	[220]	[230]	[230]
YeTrsD[g]					100	30.3	33.1	20.7	25.4
						[165]	[362]	[227]	[114]
VcRfbV[h]						100	26.3	22.8	22.4
							[152]	[136]	[250]
EaAmsD[i]							100	19.8	25.2
								[177]	[111]
StWaaG[j]								100	21.6
									[180]
SpCpoA[k]									100

a. Percentage of identical amino acids determined with FASTA as implemented in PROSIS. Numbers in parentheses indicate the number of amino acids over which the % identity occurs.
b. *S. pneumoniae* Cps19cS
c. *Methanococcus jannaschii* RfbU (GenBank accession no. F64500)
d. *Aquifex aeolicus* MtfC (GenBank accession no. AE000693)
e. *Archaeoglobus fulgidus* galactosyl transferase (GenBank accession no. AE000983)
f. *S. aureus* CapM (Lin *et al.*, 1994)
g. *Y. enterocolitica* TrsD (Skurnik *et al.*, 1995)
h. *Vibrio cholerae* RfbV (Fallarino *et al.*, 1997)
i. *E amylovora* AmsD (Bugert and Geider, 1995)
j. *S. enterica* serovar typhimurium WaaG (Heinrichs *et al.*, 1998)
k. *S. pneumoniae* CpoA (Grebe *et al.*, 1997)

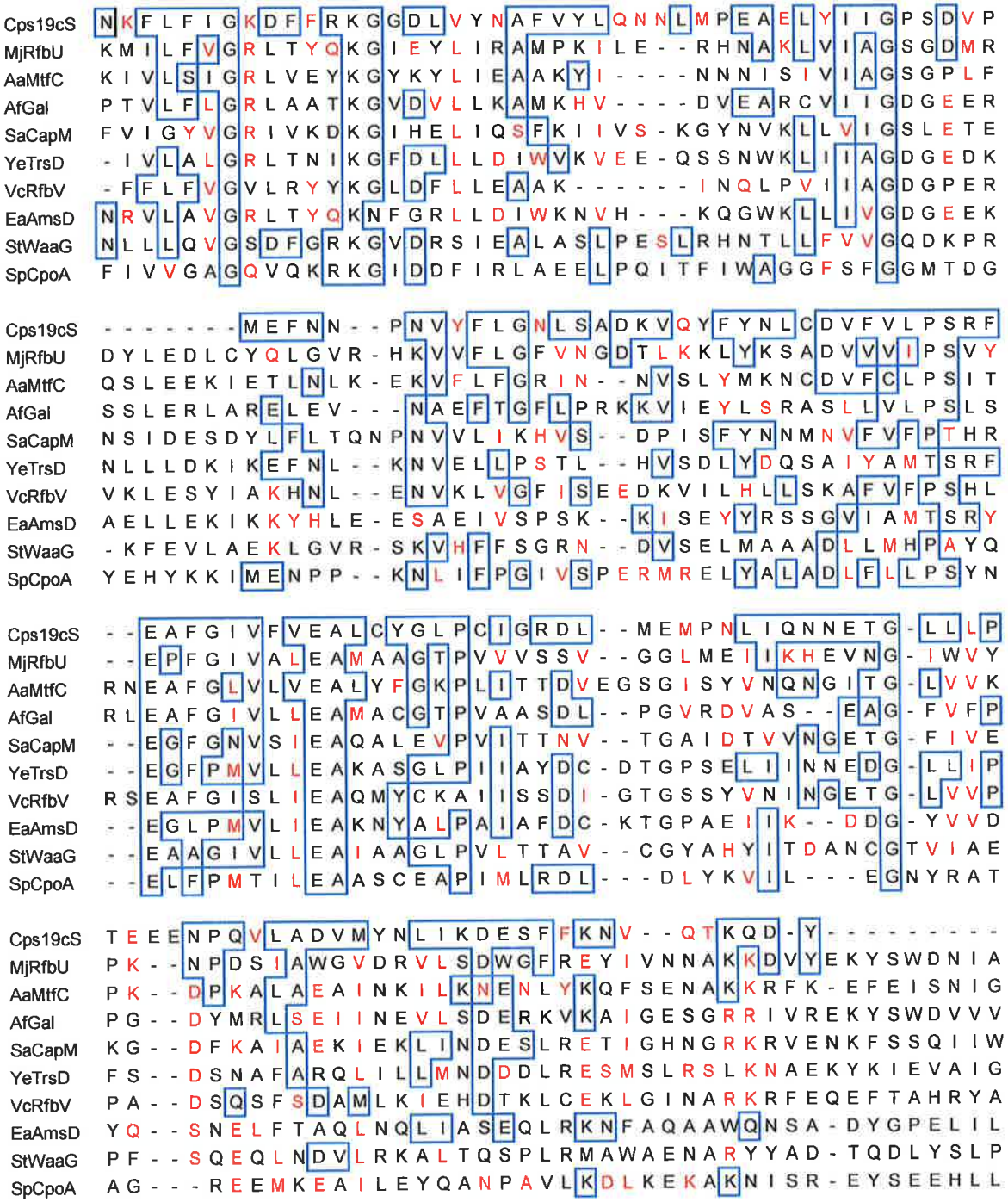


Fig. 7.5. Alignment of Cps19cS. Alignment of Cps19cS with *M. jannaschii* RfbU (MjRfbU) (GenBank accession no. F64500), *A. aeolicus* MtfC (AaMtfC) (GenBank accession no. AE000693), *A. fulgidus* Galtf (AfGal) (GenBank accession no. AE000983), *S. aureus* CapM (SaCapM) (Lin *et al.*, 1994), *Y. enterocolitica* TrsD (YeTrsD) (Skurnik *et al.*, 1995), *V. cholerae* RfbV (VcRfbV) (Fallarino *et al.*, 1997), *E. amylovora* AmsD (EaAmsD) (Bugert and Geider, 1995),and *S. enterica* serovar typhimurium WaaG (StWaaG) (Heinrichs *et al.*, 1998), as determined using the default settings of the program CLUSTAL. Residues identical to Cps19cS are boxed; similar residues are shown in red; - indicates absence of a residue.

7.2.3 Serotype specificity of *cps19cS*

To examine the relationship between *cps19cS* and encapsulation loci of other *S. pneumoniae* serotypes, a *Sac*I-*Hin*dIII DNA fragment from a nested deletion derivative of pJCP484 corresponding to nucleotides 3,160-4,269 of the *cps19c* sequence was labelled with DIG and used to probe (at high stringency) Southern blots of restricted chromosomal DNA from representative pneumococci belonging to the following types/groups: 2, 3, 4, 6, 7F, 7B, 8, 9N, 9V, 12, 14, 16, 17, 18, 19F, 19A, 19B, 20, 22, 23F and 24. None of these serotypes had a high stringency homologue to *cps19cS* (result not shown). However, this is not surprising when the structures for their CPS are examined, because none contain a Glc side-chain with a β(1→6) linkage (van Dam *et al*, 1990).

7.2.4 Transformation of *S. pneumoniae* type 19B to type 19C

In section 5.2.4, capsule production was altered from type 19F to type 19B by replacing *cps19fIJ* with the central region of *cps19b*, which contains the *cps19bPIQRJ* genes, and determines the 19B serotype. A similar approach was taken to determine whether *cps19cS* is indeed the gene responsible for the additional Glc side-chain which distinguishes type 19C CPS. A large PCR product of the *cps19c* region between *cps19cF* and *aliA* was amplified using primers J5 and J36 (**Table 2.4**) and transformed into Rx1-19F-I (an unencapsulated, Ery-resistant derivative of Rx1-19F, in which the *cps19fI* gene had been disrupted by insertion-duplication mutagenesis using pVA891, as described in section 3.2.4). The resultant transformant, expressing type 19C CPS, would be predicted to contain the *cps19cPIQRJ* genes required for both type 19B and 19C CPS biosynthesis as well as *cps19cS*. The *cps19cK* gene, which is located between *cps19cJ* and *cps19cL*, would also replace the almost identical *cps19fK* gene (94.9% identity). However, the

encoded UDP-Glc*N*Ac-2-epimerase, while essential for CPS biosynthesis in all group 19 members, is not serotype determining.

A smooth transformant was checked for Ery sensitivity, indicating loss of the pVA891 sequence. Southern hybridisation was used to confirm the absence of both pVA891 and the *cps19fI* gene, and the presence of each of the *cps19cP*, *J*, and *S* genes (**Fig. 7.6**). The presence of *cps19cP* and *J* genes in the Rx1-19C transformant were confirmed using the *cps19bP* and *cps19bJ* gene probes because these *cps19b* genes are >95% identical to those in *cps19c*. The production of a type 19C capsule by the transformant, designated Rx1-19C, was then confirmed by quellung reaction. This shows that it is possible to alter capsule production from type 19F to type 19C by replacing *cps19fIJ* with the *cps19cPIQRJ* genes (required for both type 19B and 19C CPS biosynthesis) and the *cps19cS* gene, which determines the 19C serotype.

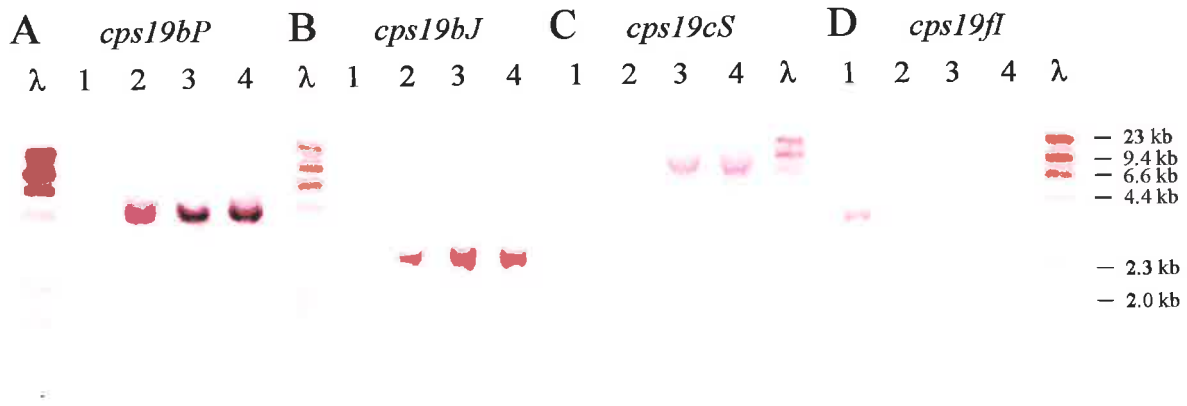


Fig. 7.6. Southern hybridisation of Rx1-19F, 19B, Rx1-19C and 19C. Rx1-19F (1), 19B (2), Rx1-19C (3), and 19C (4) were probed with DIG-labelled probes specific for *cps19bP* (**A**), *cps19bJ* (**B**), *cps19cS* (**C**) and *cps19fI* (**D**).

7.2.5 Characterisation of IS*19C*, located in the 5' intergenic region

Sequence analysis of the 5' intergenic region between *dexB* and the *cps19c* locus (section 6.2.5), indicated the presence of an additional insertion sequence designated

IS*19C*. The insertion point was within the inverted repeat of IS*1202*, just upstream of the *cps19c* locus, as shown in **Fig. 6.13**. This 1.2-kb IS element is flanked at both ends by a 13-bp direct repeat, followed by 14 bp of unique DNA and then a 14-bp inverted repeat as shown in **Fig. 7.7**. The ORF which encodes the putative transposase in IS*19C* lies in the same orientation as that for IS*1202* and opposite to the *cps19c* genes. This putative transposase has 67.5% identity to the transposase encoded by IS*1239* from *S. pyogenes*, but at the DNA sequence level these IS elements exhibit negligible similarity. The transposases from these two IS elements also share amino acid similarity to other transposases found in several different bacterial species including IS*30* from *E. coli*, as shown in **Table 7.3**.

```
1     TTCGAATGTGATATGAAAGTTCTATAATGAAGTTAGCCACCTTACCTTAGTCAAGAATTAGATGTTTCACTATGTTTGAGTAAGTTGATGATTTCATTTG

                                                                                                    -35
101   ATAACAGGTTTGAAACTGTTAGGCTAGGTGGCCAAGGCTAATCATAGCCTTGGTTTAGCTGAAAAAACAGGTTCAAGGGTTCCTGTTGTCAAATGAAATG

                -10                                                                  M  Q  E  H  Y  T  P  K  G  K
201   TGATTTAAGGTATAAGAAAACACCTCTGTGCTATACTTGTTGTTCACCACAAACACAAGGAAAGGCACAGAGATGCAAGAACATTATACCCCAAAAGGGA

        H  L  T  I  D  N  R  R  L  I  E  R  W  K  N  E  N  K  S  N  R  E  I  A  G  L  L  G  K  A  P  Q  T
301   AACATTTGACAATAGATAACCGTCGCTTGATTGAGCGGTGGAAGAATGAAAATAAGTCCAATCGTGAAATTGCAGGCTTGTTAGGAAAGGCGCCTCAAAC

        I  H  N  E  V  K  R  G  T  T  L  Q  Q  V  R  K  G  L  Y  K  K  V  Y  S  A  D  Y  A  Q  T  V  Y  Q
401   GATTCATAATGAAGTCAAAAGAGGTACAACTTTACAACAAGTGAGAAAAGGGCTATACAAAAAGGTCTATTCTGCCGATTACGCACAAACTGTTTACCAA

        F  N  R  K  R  S  V  K  K  L  I  L  T  K  E  I  R  E  K  I  L  H  Y  H  K  Q  K  F  S  P  E  M  M  V
501   TTCAATCGAAAACGGTCGGTGAAAAAGTTAATTTTAACAAAGGAAATCAGAGAGAAGATCTTACACTATCATAAGCAAAAATTTTCGCCTGAAATGATGG

        N  K  K  Q  V  K  V  G  I  S  T  I  Y  Y  W  F  H  N  G  H  L  R  L  T  K  A  D  M  L  Y  P  R  K
601   TTAACAAGAAGCAAGTGAAAGTTGGTATTTCAACCATCTACTACTGGTTTCATAATGGTCATTTAAGATTAACGAAGGCCGACATGCTTTATCCCAGAAA

        R  K  G  V  K  K  Q  A  S  P  N  F  K  P  A  G  K  S  I  E  E  R  P  D  V  I  N  L  R  L  E  N  G
701   AAGGAAAGGTGTCAAGAAGCAAGCTAGTCCGAACTTTAAGCCGGCAGGTAAATCTATCGAAGAACGTCCTGACGTTATTAATCTTCGCTTGGAAAATGGT

        H  Y  E  I  D  T  V  L  L  T  K  I  K  N  Y  C  L  L  V  L  T  D  R  R  S  R  H  Q  I  I  R  L  I  P
801   CATTATGAAATTGATACCGTCCTACTGACTAAGATAAAAAATTATTGCCTGTTAGTCTTAACCGACCGGCGGAGCAGACACCAAATTATAAGGTTAATTC

        N  K  T  A  E  S  V  N  Q  A  L  T  L  L  L  G  E  H  R  I  L  S  I  T  A  D  N  G  S  E  F  K  R
901   CAAATAAAACTGCTGAATCTGTCAATCAGGCGCTTACGTTACTATTAGGGGAGCATCGTATTCTGTCCATTACTGCAGATAATGGTTCGGAGTTCAAACG

        L  S  E  V  F  P  E  E  H  I  Y  Y  A  H  A  Y  S  S  W  E  R  G  S  N  E  N  H  N  R  L  I  R  R
1001  ATTGTCTGAGGTATTTCCTGAGGAACATATCTACTACGCACATGCTTACTCTTCATGGGAGAGAGGTTCAAATGAAAATCATAATCGATTAATTCGGAGA

        W  L  P  E  G  T  K  K  T  T  P  K  E  V  A  F  I  E  N  W  I  N  N  Y  P  K  K  C  L  D  Y  K  S  P
1101  TGGTTACCTGAAGGAACCAAGAAAACGACTCCGAAAGAAGTAGCTTTTATCGAAAATTGGATTAACAACTACCCTAAAAAATGCTTGGACTACAAGTCGC

        N  E  F  L  L  G  G  *
1201  CAAATGAATTTCTTTTGGGTGGCTAACTTCAACTTGAAATTTGGGTTCGAATGTGATA
```

Fig. 7.7. IS*19C*. The sequence of IS*19C* as shown above corresponds to nucleotides 1,644-2,875 of the sequence available under GenBank accession no. AF105115. The translation of the putative transposase is shown in single letter code above the first nucleotide of each codon. The direct repeats are underlined and the inverted repeats are doubly underlined and are both shown in bold. A potential -35 and -10 promoter sequence for the putative IS*19C* transposase is underlined with a broken line. The ribosome binding site is underlined, and the amino acid translation is represented by single letter code above the first nucleotide of each codon; the stop codon is indicated by an asterisk.

Table 7.3. Similarity of the putative transposase encoded by IS*19C* to other transposases.

	% Identity[a]								
	IS*19c*	IS*1239*	IS*1161*	IS*4351*	IS*1470*	IS*1070*	IS*1086*	IS*30*	IS*1394*
IS*19c*[b]	100	67.5 [314]	36.2 [232]	33 [318]	31.5 [305]	31.1 [254]	28.8 [326]	28 [321]	26.9 [308]
IS*1239*[c]		100	29 [314]	31.5 [321]	31.8 [333]	30.7 [241]	28.8 [326]	26.2 [321]	29 [328]
IS*1161*[d]			100	27.7 [328]	36.3 [344]	34.5 [203]	26.5 [336]	28.7 [328]	28.1 [320]
IS*4351*[e]				100	41.2 [170]	30.7 [313]	30.5 [325]	31.6 [326]	33 [324]
IS*1470*[f]					100	31.5 [302]	26.8 [340]	30.8 [334]	29.2 [332]
IS*1070*[g]						100	30.8 [208]	28.6 [280]	33.3 [204]
IS*1086*[h]							100	41.1 [326]	58.3 [326]
IS*30*[i]								100	42.2 [329]
IS*1394*[j]									100

a. Percentage of identical amino acids determined with FASTA as implemented in PROSIS. Numbers in parentheses indicate the number of amino acids over which the % identity occurs.
b. *S. pneumoniae* IS*19C* putative transposase
c. *S. pyogenes* IS*1239* putative transposase (Kapur *et al.*, 1994)
d. *Streptococcus salivarius* IS*1161* putative transposase (Giffard *et al.*, 1993)
e. *Bacteroides fragilis* IS*4351* putative transposase (Smith, 1987)
f. *Clostridium perfringus* IS*1470* putative transposase (Brynestad *et al.*, 1997)
g. *Leuconostoc lactis* IS*1070* putative transposase (Vaughan and de Vos, 1995)
h. *Ralstonia eutropha* IS*1086* putative transposase (Dong *et al.*, 1992)
i. *E. coli* IS*30* transposase (Dalrymple *et al.*, 1984)
j. *Pseudomonas alcaligenes* IS*1394* putative transposase (GenBank accession no. U37284)

7.3 Conclusions

The *cps19c* locus is almost identical to the *cps19b* locus except that an extra gene (*cps19cS*) has inserted between *cps19cK* and *cps19cL*. This gene is most likely to encode the glucosyl transferase required for the addition of the Glc side-chain in the type 19C repeat unit. Interestingly, all three putative transferases involved in the addition of side-chains to type 19B and/or 19C CPS, Cps19cS, Cps19P and Cps19Q, appear to be cytoplasmic enzymes, as they lack both a leader sequence for export to the cell surface and

a hydrophobic transmembrane sequence which could anchor them to the cell membrane. Thus, the Rha-Rib disaccharide side-chain present in both type 19B and 19C CPS, and the Glc side-chain specific to type 19C CPS are most probably added to the repeat units in the cytoplasm, before translocation to the outer surface by Cps19J and subsequent polymerisation by Cps19I. It is interesting that the Glc side-chain does not appear to interfere with the function of either the repeat unit transporter (Cps19bJ and Cps19cJ) or the polysaccharide polymerase (Cps19bI and Cps19cI), as the proteins encoded by *cps19b* and *cps19c* are almost identical (99.7% and >95%, respectively) and are able to function in the biosynthesis of both the type 19B and type 19C CPS.

Chapter 8

DISCUSSION

8.1 Introduction

S. pneumoniae group 19 is the first group for which the *cps* loci from all the members (19F, 19A, 19B and 19C) have been completely characterised. Functions have been assigned to the majority of the *cps19* gene products, based on either gene complementation or similarity to other proteins with known functions. The ability of PCR products containing either complete or partial *cps* loci to transform pneumococci from one serotype to another demonstrated that the *cps19* loci contain sufficient genetic information for expression of type-specific CPSs.

Comparison of the *cps* loci from all four members of group 19 and to the *cps* loci from other pneumococcal serotypes will be discussed below.

8.2 Analysis of the *S. pneumoniae* group 19 *cps* loci

8.2.1 Comparison of the group 19 *cps* loci

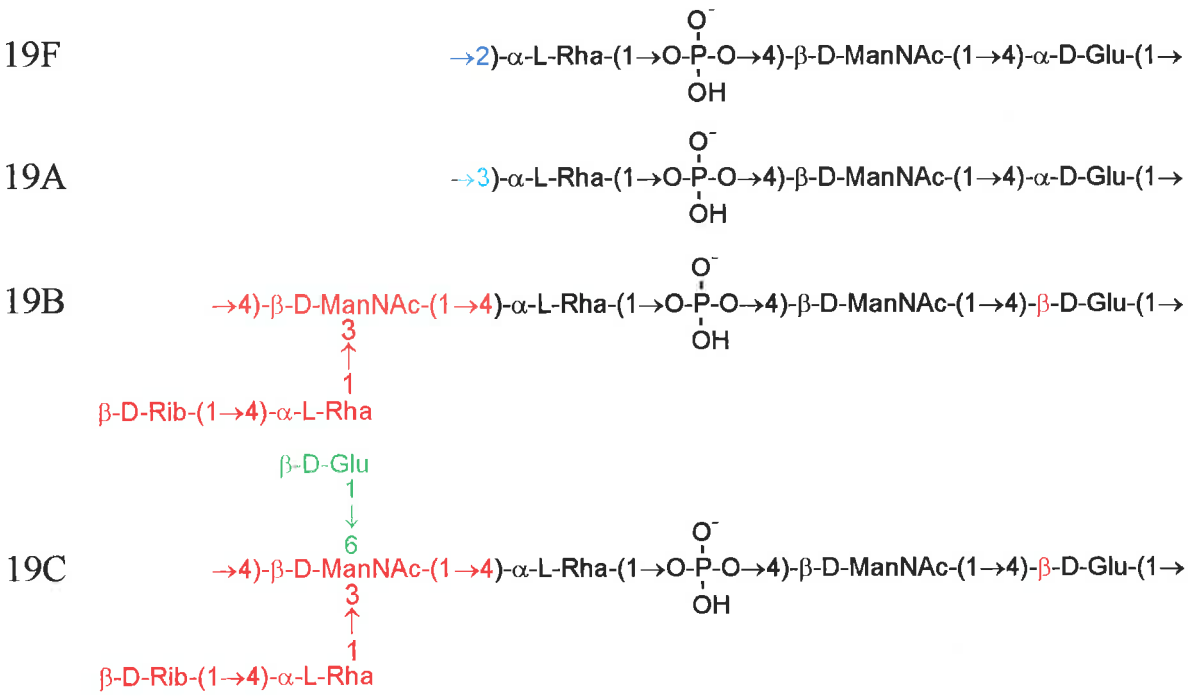
The structural similarities between the CPS repeat units from all four members of

serogroup 19 (19F, 19A, 19B and 19C) are reflected in the highly conserved arrangement of their *cps* loci, with 13 genes (*cps19A-H*, and *K-O*) common to all four serogroup members, as shown in **Fig. 8.1**. Nearly all of the common genes from types 19F, 19B and 19C are >95% identical to each other, whereas those from the type 19A *cps* locus are more divergent.

Although the *cps19a* and *cps19f* loci are identical in the number and arrangement of the genes present, the similarity between individual genes varies from 70% to 99% identity (for both the nucleotide and the deduced amino acid sequences). This sequence divergence is surprising given that the only difference between their CPS repeat units is the glycosidic linkage which joins the repeat units together ($\alpha(1\rightarrow2)$ for 19F and $\alpha(1\rightarrow3)$ for 19A) (**Fig. 8.1A**). Hypothetically, only a difference in the *cps19aI* gene, which presumably encodes the polysaccharide polymerase responsible for this linkage, is required to change a type 19F pneumococcus into type 19A. The results of the present study are consistent with this hypothesis. A transformation event in which the region of the *cps19a* locus encoding most of Cps19aH, all of Cps19aI and the first 76 amino acids of Cps19aJ replaced the homologous portion of the *cps19f* locus was sufficient to convert CPS type from 19F to 19A (section 6.2.4). Given that Cps19fH and Cps19aH are >95% identical, it seems probable that Cps19aI is solely responsible for the observed alteration in CPS type.

The degree of divergence between the type 19F and 19A loci suggests that they may have diverged early in their evolutionary past, or that their component genes originated from different sources. The latter alternative is supported by the fact that the G+C content of the *cps19a* locus is higher than that for *cps19f* (section 8.2.2). In chapters 4 and 6, sequence variations between the *cpsCDE* genes among different pneumococcal serotypes identified two distinct classes of *cps* loci. Type 19A was recognised as a class II locus whereas the other members of group 19 belonged to class I. Thus, the marked difference

A.



B.

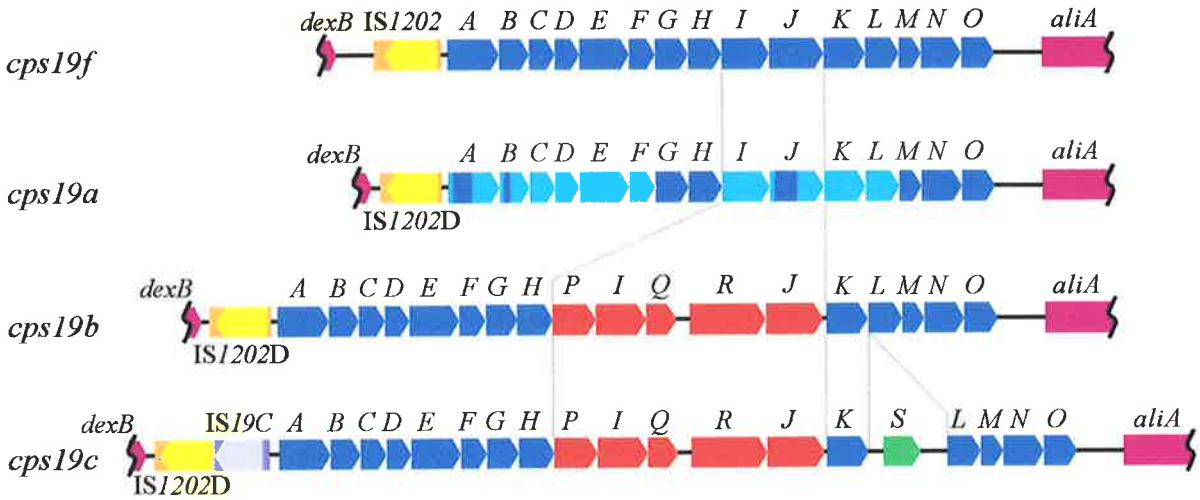


Fig. 8.1 Comparison of the CPS structures and the *cps* loci of *S. pneumoniae* group 19. A. CPS repeat unit structures. The part of the CPS repeat unit structure shown in black is common to all group 19 members. The differences in their CPS structures are highlighted in different colours: the different repeat unit linkages are shown in dark blue for 19F and light blue for 19A, the additional sugars common to 19B and 19C are shown in red, and the additional glucose side-chain in 19C is shown in green. **B. The *cps* loci.** The genes of the *cps19f* locus are shown in dark blue, and the *cps19f* homologues from other group 19 members with >90% identity are also shown in dark blue. The *cps19a* genes which have 70-90% identity to the *cps19f* homologue are shown in light blue. The 19B and 19C specific genes are shown in red and the additional 19C specific gene is shown in green.

between the genes at the 5' end of the type 19A *cps* locus and homologues in other members of group 19 is almost certainly a consequence of recombination between portions of class I and class II loci, rather than divergence over a long period of time.

Closely related homologues of *cps19fI* and *J*, which encode the type 19F polysaccharide polymerase and repeat unit transporter, respectively, are not found in the type 19B *cps* locus. In type 19B this region of the *cps* locus (between *cps19bH* and *cps19bK*) contains five genes (**Fig. 8.1B**), encoding an unrelated polymerase and repeat unit transporter, as well as two additional putative glycosyl transferases and a protein which may be involved in synthesis of an activated Rib precursor (section 5.2.2.4). Transformation studies indicated that these five genes encode all of the functions required to convert a type 19F pneumococcus to type 19B (section 5.2.4).

The type 19C locus is virtually identical to the type 19B locus, with the *cps19c* genes exhibiting >95% identity to all 18 of the *cps19b* genes, differing only in the insertion of a glucosyl transferase gene (*cps19cS*) between *cps19cK* and *cps19cL*. Transformation studies have shown that the presence of this gene accounts for the additional Glc side chain in the otherwise identical repeat unit structures (**Fig. 8.1A**). The type 19C *cps* locus contains 19 genes, and at 21 kb, it is the largest pneumococcal capsule gene cluster characterised to date.

The 13 genes common to all members of group 19 encode functions required for the synthesis of the shared trisaccharide component of the group 19 CPS structures. Furthermore, the genetic differences between the group 19 *cps* loci identified are also consistent with the differences in the CPS structures of individual serotypes. This information has been used to propose biosynthetic pathways for each of the serotypes as shown in **Fig. 8.2**. Only the functions of Cps19A and Cps19B are still unknown, and hence, are not included in **Fig. 8.2**.

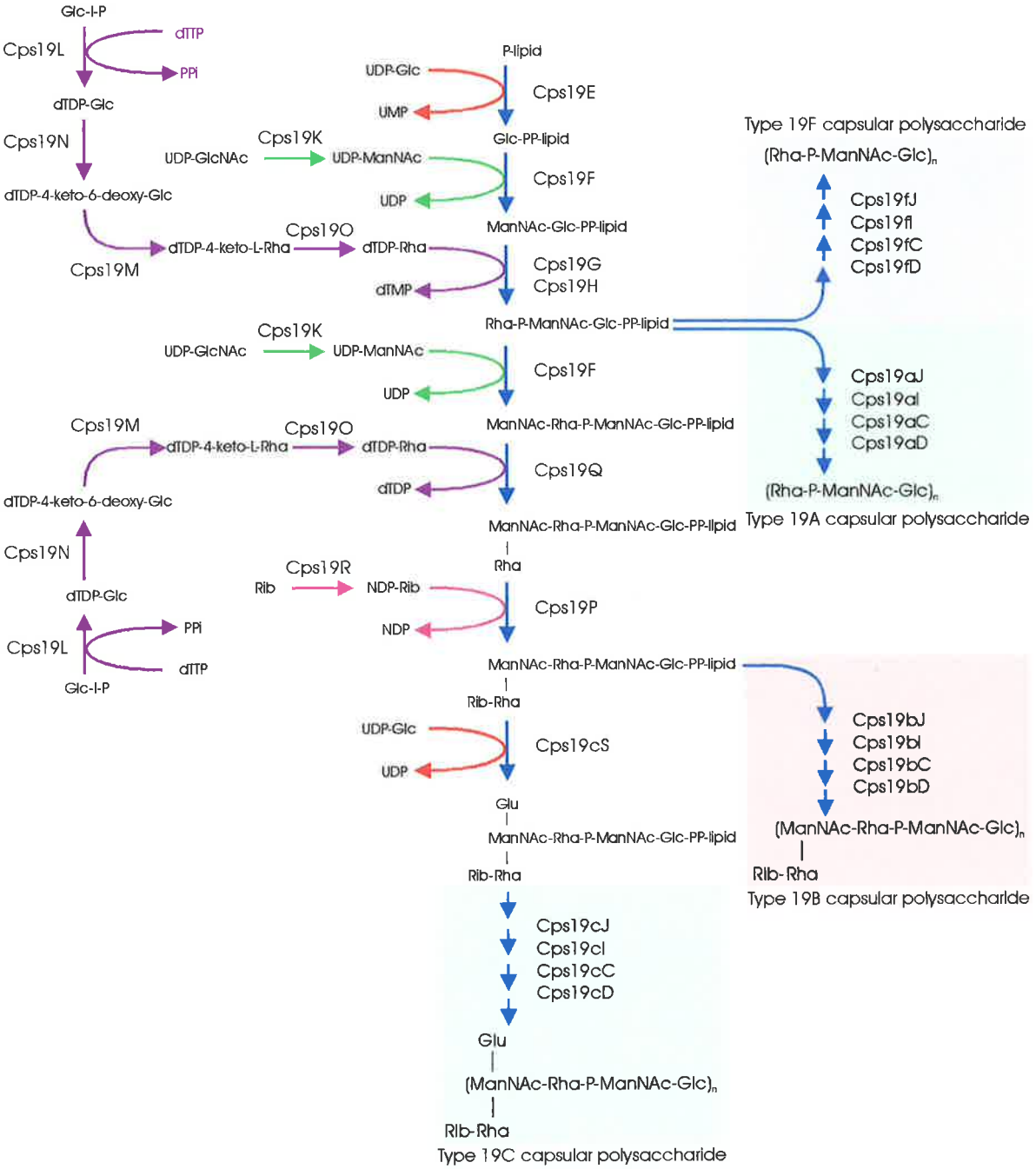


Fig. 8.2. The putative biosynthetic pathways for 19F, 19A, 19B and 19C.

8.2.2 Distribution of the *cps19* genes in other pneumococcal serotypes

When the distribution of the *cps19* genes in other serotypes was examined, homologues of the first four genes, *cpsA-D,* were identified in the *cps* loci of all serotypes tested to date, and furthermore, the *cpsC* and *cpsD* genes existed as two distinct classes, designated class I and class II. The *cpsE* gene, encoding the glucosyl transferase which adds Glc-1-phosphate to the lipid carrier, is also conserved as two distinct alleles in the class I and class II *cps* loci of all serotypes tested which contain Glc in their CPS (except type 3). Thus, pneumococcal *cps* loci can be segregated into class I and class II loci which may have evolved from two distinct clonal ancestors, based on the degree of nucleotide and amino acid similarity between them. Whereas, *cpsC-E* genes in the same class share >95% identity at both the DNA and amino acid level, they share only 70-80% identity to genes belonging to the other class. The precise point at which these sequences diverge into the two distinct classes appears to vary between serotypes but probably occurs within *cpsB* in most serotypes.

Homologues of the four genes involved in dTDP-Rha biosynthesis (*cps19L-O*) were also present in all serotypes tested that contained Rha in their CPS. A homologue to *cps19fF*, a putative Man*N*Ac transferase responsible for linking Man*N*Ac via a β(1→4) linkage to Glc in the type 19 CPS, was also present in types 9N and 9V. This was not surprising as their CPS repeat units also contain this same glycosidic linkage. The remainder of the *cps19* genes were all specific to one or more members of group 19.

8.2.3 G+C content

The G+C content of the individual genes in the *cps19* locus is quite variable, ranging from 27-43%, with genes of similar G+C content clustered together in a mosaic structure. The G+C content of some of the *cps* genes are atypical for pneumococci and this

suggests that *S. pneumoniae* may have acquired these genes from a different bacterial source. The G+C content of the individual genes is shown in **Table 8.1** and is represented diagrammatically in **Fig 8.3**. Interestingly, the G+C content of most of the genes in the *cps19a* locus are consistently higher (up to 6.5%) than that for the respective *cps19f* gene, suggesting a distinct ancestral origin for these two loci.

Also of interest is the low G+C content of the type 19B-specific gene cluster (*cps19bPIQRJ*), which ranges from 27.2-29.7% for the individual genes (**Table 8.1**). This is comparable to the G+C content for the *cps19fIJ* gene cluster of 29.7%, but differs considerably from the remainder of either of the loci, which ranges from 30.3-42.3% G+C in the *cps19f* locus (**Table 8.1**). Conversely, the four type 19F genes encoding dTDP-Rha biosynthesis (*cps19fLMNO*) have G+C contents ranging from 41.5-42.3%, which is significantly higher than that of flanking sequences and for pneumococcal genes in general, the average G+C content of which have been estimated at 35.1% (Bridge and Sneath, 1983). This is consistent with acquisition of these different gene clusters from distinct sources.

8.2.4 Transcription and translation of the *cps19* loci

All the *cps19* loci examined contained an identical consensus σ^{70} promoter (TAGACA-17 bp-TATAAT) just upstream of the first ORF, as described for *cps19f* (Guidolin et al., 1994). All *cps19* ORFs identified were preceded by a ribosome binding site and are closely linked with most intergenic gaps being less than 100 nucleotides. Even the four intergenic gaps larger than 100 nucleotides did not appear to contain any sequences which resemble either transcription terminators or promoter sequences. The only stemmed-loop structures likely to function as transcription terminators are found downstream of the *cps19* loci in the 3' intergenic region (section 3.2.1). Thus the entire

Table 8.1. The G+C content of the *cps19* genes.

	%G+C			
Gene	19F	19A	19B	19C
cps19A	38.1	39.5	NS	NS
cps19B	38	41.3	NS	NS
cps19C	38.2	42.1	NS	NS
cps19D	34.5	41	NS	NS
cps19E	33.2	37.7	NS	NS
cps19F	33.6	34.1	34.2*	NS
cps19G	36.3	37.2	36.6	NS
cps19H	30.3	32.2	30.2	NS
cps19I	29.7	32.7	27.2	NS
cps19J	29.7	33	29.5	29.9
cps19K	35.2	36.9	35.2	35.3
cps19L	42.3	43.3	41.9*	42.6*
cps19M	41.5	41.2	NS	NS
cps19N	42.1	42.4	NS	NS
cps19O	41.5	41.3	NS	NS
cps19P	-	-	29.5	NS
cps19Q	-	-	29.7	NS
cps19R	-	-	27.2	26.4*
cps19cS	-	-	-	31.4

NS, denotes not sequenced; "-", denotes absence of gene in that serotype; "*" denotes only part of the gene was sequenced.

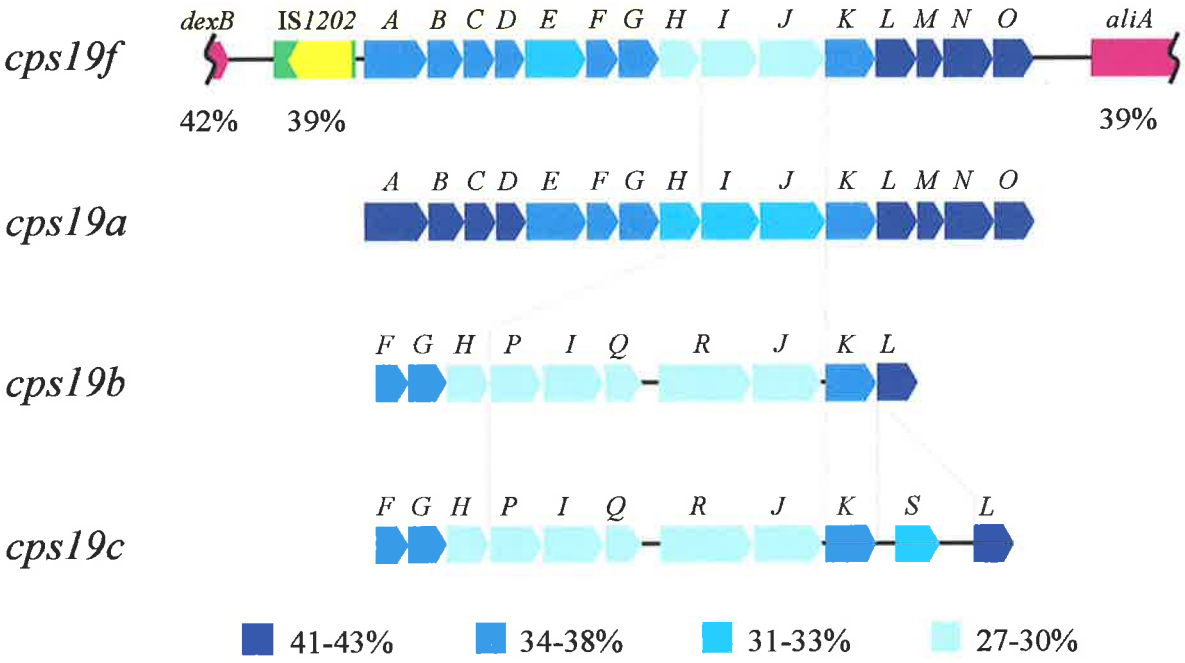


Fig. 8.3. Diagrammatic representation of the G+C content of the group 19 *cps* loci. Individual *cps* genes are shaded in accordance with their approximate G+C content. For flanking genes, the % G+C is shown on the map. Portions of the *cps19b* and *cps19c* loci are not shown, as they are >95% identical to *cps19f*.

loci are probably co-transcribed in each of the four serotypes.

Although most of the *cps19* genes used an ATG initiation codon, a few of the genes were predicted to use a TTG initiation codon. For one of these (*cps19fH*) use of a TTG initiation codon was confirmed by T7 expression studies in *E. coli* (section 3.2.5). The *cps19H* homologues in the *cps* loci of the other members of group 19 also have TTG initiation codons, as does *cps19aG*.

8.3 The *cps* loci from other *S. pneumoniae* serotypes

Since the publication of the first six genes of the *cps19f* locus (Guidolin *et al*, 1994) sequence data for *cps* loci from *S. pneumoniae* serotypes 1, 3, 14 and 23F have been published (Arrecubieta *et al.*, 1995; Dillard *et al.*, 1995; Kolkman *et al.*, 1997b; Muñoz *et al.*, 1997; Ramirez and Tomasz, 1998). The sequence of the type 4 *cps* locus has also become available by analysis of the partial genome sequence for *S. pneumoniae* type 4 released by the TIGR Microbial Database (http://www.tigr.org/pub/data/s_pneumoniae/). The *cps* loci from a different *S. pneumoniae* type 23F strain and from types 2 and 33F also have recently been determined and are either submitted for publication or are currently in press (Iannelli *et al.*, 1998; Llull *et al.*, 1998; Morona *et al.*, 1998). An overview of the genes present in these loci is provided below to enable a comparison with those from the group 19 *cps* loci.

8.3.1 The *S. pneumoniae* type 1 *cap* locus

The *S. pneumoniae* type 1 *cps* locus (Muñoz *et al*, 1997) has 11 ORFs, designated *cap1A-K*, and the predicted functions of their protein products are listed in **Table 8.2**. The

cps locus is flanked on both sides by IS*1167* and redundant copies of the genes required for dTDP-Rha biosynthesis (*orf1-4*) were found downstream of the 3' copy of IS*1167*. The functions of the protein products encoded by both *cap1H* and *cap1I* have been deduced by examination of their hydropathy profiles (data not shown).

Table 8.2. Predicted functions of *S. pneumoniae cap1*[a] ORFs

ORF	Predicted or known function	Similar proteins (% identity)
cap1A	regulation?	*S. pneumoniae* Cps19fA (>98%)
cap1B	unknown	*S. pneumoniae* Cps19fB (>98%)
cap1C	chain length regulation/export	*S. pneumoniae* Cps19fC (>98%)
cap1D	chain length regulation/export	*S. pneumoniae* Cps19fD (>98%)
cap1E	galacturonic acid transferase	*Proteus mirabilis* GpsF[b] (33.5%) *S. dysenteriae* Rfp[c] (29.1%)
cap1F	acetyl transferase	*Rhizobium leguminosarium* NodL[d] (30.3%) *E. coli* K12 LacA[e] (29.7%)
cap1G	glycosyl transferase	*S. aureus* CapJ[f] (30.1%) *E. coli* K12 WcaL[g] (19.5%)
cap1H	polysaccharide polymerase	-
cap1I	repeat unit transporter	*M. jannaschii* MJ1068[h] (22.7%) *Shigella sonnei* form 1 ORF3[i] (22.3%)
cap1J	galacturonic acid epimerase	*V. cholerae* Orf9[j] (53.2%) *E. coli* K12 GalE[k] (27.5%)
cap1K	UDP-glucose dehydrogenase	*S. pneumoniae* Cap3A[l] (61.7%)
orf1	glucose-1-phosphate thymidylyl transferase	*S. pneumoniae* Cps19fL (96.1%)
orf2	dTDP-4-keto-6-deoxyglucose-3,5-epimerase	*S. pneumoniae* Cps19fM (98%)
orf3	dTDP-glucose-4,6-dehydratase	*S. pneumoniae* Cps19fN (98.8%)
orf4	dTDP-L-rhamnose synthase	*S. pneumoniae* Cps19fO (97.5%)

The sequences can be found under GenBank accession nos: a, Z83335; b, L36873; c, S27671; d, X17557; e, X51872; f, U10927; g, X56793; h, U67549; i, U34305; j, U47057; k, X06226; l, Z47210.

8.3.2 The *S. pneumoniae* type 2 *cps* locus

The *S. pneumoniae* type 2 *cps* locus has recently been sequenced by Iannelli *et al.* (1998), and consists of 17 ORFs, designated *cps2A-O*, *P* and *T*, and the predicted functions

of their protein products are listed in **Table 8.3**. Interestingly, the *cps2F* gene product has significant similarity to the experimentally determined galactosyl transferases, Cps14J and Cps14I, from *S. pneumoniae* type 14, as well as the putative rhamnosyl transferase RgpBc from *S. mutans*. As type 2 CPS does not contain Gal, Cps2F is probably one of the rhamnosyl transferases required for type 2 CPS biosynthesis.

Table 8.3 Predicted functions of *S. pneumoniae cps2*[a] ORFs

ORF	Predicted function	Similar proteins (% identity)
cps2A	regulation?	*S. pneumoniae* Cps19fA (96%)
cps2B	unknown	*S. pneumoniae* Cps19fB (84%)
cps2C	chain length regulation/export	*S. pneumoniae* Cps19fC (67%)
cps2D	chain length regulation/export	*S. pneumoniae* Cps19fD (78%)
cps2E	glucosyl-1-phosphate transferase	*S. pneumoniae* Cps23fE[b] (95%) *S. pneumoniae* Cps14E[c] (60%) *S. pneumoniae* Cps19fE (60%)
cps2T	rhamnosyl transferase	*S. pneumoniae* Cps23fT[b] (81%)
cps2F	rhamnosyl transferase	*S. mutans* RgpBc[d] (33%) *S. pneumoniae* Cps23fU[b] (23%) *S. pneumoniae* Cps14J[c] (19%) *S. pneumoniae* Cps14I[c] (18%)
cps2G	glycosyl transferase	*S. pneumoniae* RPN00103[e] (35%) *S. pneumoniae* RPN00642[e] (20%) *N. meningitidis* IcsA[f] (19%)
cps2H	polysaccharide polymerase	*S. pneumoniae* Cps14H[c] (21%) *S. pneumoniae* Cps19fI (20%) *S. flexneri* Rfc[g] (22%)
cps2I	glycosyl transferase	*N. meningitidis* RfaK[h] (25%) *N. meningitidis* IcsA[f] (24%)
cps2J	repeat unit transporter	*S. pneumoniae* Cps23fJ[b] (36%)
cps2K	UDP-glucose dehydrogenase	*S. pneumoniae* Cap1K[i] (89%) *S. pneumoniae* Cps3D[j] (Cap3A[k])(74%)
cps2P	UDP-galactopyranose mutase	*E. coli* Glf[l] (67%) *Mycobacterium tuberculosis* Glf[m] (58%)
cps2L	glucose-1-phosphate thymidylyl transferase	*S. pneumoniae* Cps19fL (91%)
cps2M	dTDP-4-keto-6-deoxyglucose-3,5-epimerase	*S. pneumoniae* Cps19fM (96%)
cps2N	dTDP-glucose-4,6-dehydratase	*S. pneumoniae* Cps19fN (99%)
cps2O	dTDP-L-rhamnose synthase	*S. pneumoniae* Cps19fO (99%)

The sequences can be found under GenBank accession nos: a, AF026471; b, AF030373; c, X85787; d, AB010970; e, GenBank unfinished genomes (TIGR Genomic Database); f, U39810; g, X71970; h, U58765; i, Z83335; j, U15171; k, Z47210; l, U09876 ; m, U96128.

8.3.3 The *S. pneumoniae* type 3 *cps* locus

The sequence of the *S. pneumoniae* type 3 *cps3/cap3* locus contains six ORFs, only four of which are transcribed. These four genes, have been designated *cps3D, S, U* and *M* by Dillard *et al.* (1995) and *cap3B, A, C* and *D* by Arrecubieta *et al.* (1995). The predicted functions of their protein products are listed in **Table 8.4**.

Table 8.4 Predicted functions of *S. pneumoniae cps3*[a]*/cap3*[b] ORFs

ORF	Predicted or known function	Similar proteins (% identity)
cps3D/cap3A	UDP-glucose dehydrogenase	*S. pyogenes* HasB[c] (57%) *E. coli* Ugd[d] (55.3%)
cps3S/cap3B	type 3 synthase	*S. pyogenes* HasA[e] (25.8%) *R. meliloti* NodC[f] (21%) *S. cerviciae* Csd2[g] (24.7%)
cps3U/cap3C	glucose-1-phosphate uridylyltransferase	*S. pyogenes* HasC[h] (78%) *B. subtilis* GtaB[i] (55%)
cps3M/orf3	phosphoglucomutase	*B. subtilis* YhxB[j] (38.6%) *E. coli* PgmU[k] (22.6%)

The sequences can be found under GenBank accession nos: a, U66846, U15171, and U66845; b, Z47210; c, L08444; d, U90519; e, L21187; f, X01649; g, M73697; h, Q54713; i, Z99122; j, Y14079; k, AE000172.

8.3.4 The *S. pneumoniae* type 4 *cps* locus

The sequence of the *S. pneumoniae* type 4 *cps* locus (located on contig no. 4108 of the partial *S. pneumoniae* genome sequence available from the TIGR Microbial Database) was examined and the *cps4* locus, located between *dexB* and *aliA* in the *S. pneumoniae* chromosome, was found to have 15 ORFs, designated *cps4A-O*. The predicted functions of the protein products encoded by *cps4A-O*, based on homologies to other proteins in the GenBank database, are listed in **Table 8.5**, and are sufficient for biosynthesis of type 4 CPS repeat unit which is shown in **Table 1.2**.

Table 8.5 Predicted functions of *S. pneumoniae cps4* ORFs

ORF	Predicted function	Similar proteins (% identity)
cps4A	regulation?	*S. pneumoniae* Cps19fA (94.2%)
cps4B	unknown	*S. pneumoniae* Cps19fB (86.8%)
cps4C	chain length regulation/export	*S. pneumoniae* Cps19fC (86.5%)
cps4D	chain length regulation/export	*S. pneumoniae* Cps19fD (94.3%)
cps4E	galactosyl-1-phosphate transferase	*S. aureus* Cap8M[a] (48.6%) *S. aureus* Cap5M[b] (49.2%) *S. enterica* serovar typhimurium RfbP[c] (36.5%)
cps4F	N-acetyl fucosamine transferase	*S. aureus* Cap8L[a] (23.9%) *S. aureus* Cap5L[b] (23.9%) *E. coli* K12 WcaI[d] (22.6%)
cps4G	N-acetyl mannosamine transferase	*S. aureus* Cap8H[a] (32%)
cps4H	N-acetyl galactosamine transferase	*Streptococcus thermophilus* EpsF[e] (25.8%) *H. influenzae* lsg Orf3[f] (19.8%)
cps4I	polysaccharide polymerase	-
cps4J	pyruvyl transferase	*E. coli* K12 WcaK[d] (17.8%)
cps4K	repeat unit transporter	*E. coli* K12 RfbX[g] (20.9%) *Y. enterocolitica* TrsA[h] (19.1%) *S. enterica* serovar typhimurium RfbX[i] (20.1%)
cps4L	UDP-N-acetyl glucosamine-2-epimerase	*B. subtilis* YvyH[j] (63.2%) *S. aureus* Cap5P[b] (57.2%) *S. aureus* Cap8P[a] (57.2%) *E. coli* K12 RffE[k] (53%)
cps4M	UDP-N-acetyl galactosamine dehydratase (UDP-FucNAc synthesis)	*S. aureus* Cap8E[a] (61.3%) *S. aureus* Cap5E[b] (61%) *Y. enterocolitica* TrsG[l] (33.4%)
cps4N	UDP-N-acetyl fucosamine synthase (UDP-FucNAc synthesis)	*S. aureus* Cap8F[a] (45.6%) *S. aureus* Cap5F[b] (46.2%) *S enterica* serovar typhimurium RfbE[c] (25%)
cps4O	UDP-N-acetyl glucosamine-4-epimerase (UDP-FucNAc synthesis)	*S. aureus* Cap8G[a] (58.9%) *S. aureus* Cap5G[b] (58.9%) *B. subtilis* YvyH[j] (25.3%)

The sequences can be found under GenBank accession nos: a, U73374; b, U81973; c, X56793; d, U38473; e, U40830; f, M94855; g, AF013583; h, Z47767; i, X60665; j, Z99122, k, L18799; l, Z47767.

8.3.5 The *S. pneumoniae* type 14 *cps* locus

The *S. pneumoniae* type 14 cps locus (Kolkman *et al.*, 1997b) consists of 12 ORFs, designated *cps14A-L* and the predicted functions of their protein products are listed in **Table 8.6**. The functions of the glycosyl transferases have been experimentally determined.

Table 8.6. Predicted functions of *S. pneumoniae cps14*[a] ORFs

ORF	Predicted or known function	Similar proteins (% identity)
cps14A	regulation?	*S. pneumoniae* Cps19fA (96.9%)
cps14B	unknown	*S. pneumoniae* Cps19fB (96.3%)
cps14C	chain length regulation/export	*S. pneumoniae* Cps19fC (92.6%)
cps14D	chain length regulation/export	*S. pneumoniae* Cps19fD (95.2%)
cps14E	glucosyl-1-phosphate transferase	*S. pneumoniae* Cps19fE (95.8%) *S. enterica* serovar typhimurium RfbP[b] (31.3%)
cps14F	transferase co-factor	*Sphyngomonas* S88 SpsK[c] (33.8% to N-terminal)
cps14G	β-1,4-galactosyl transferase	*Sphyngomonas* S88 SpsK[c] (27.7% to C-terminal)
cps14H	polysaccharide polymerase	*S. flexneri* Rfc[d] (25.6%) *S. enterica* serovar typhimurium Rfc[b] (21%)
cps14I	β-1,3-N-acetylglucosaminyl transferase	*R. meliloti* ExoO[e] (26.6%) *S. thermophilus* EpsI[f] (26.1%)
cps14J	β-1,4-galactosyl transferase	*S. thermophilus* EpsI[f] (43.9%) *Y. enterocolitica* TrsB[g] (26.9%)
cps14K	unknown	*H. influenzae* type b Orf4[h] (28.3%)
cps14L	repeat unit transporter	*S. dysenteriae* RfbX[i] (26.1%) *E. coli* K12 RfbX[j] (33%) *Y. enterocolitica* O:3 Wzx[g] (23.5%) *S. pneumoniae* Cap1I[k] (20.7%)
orfX	glycerol-phosphate transferase	*B. subtilis* TagF[l] (26.6%)

The sequences can be found under GenBank accession nos: a, X85785; b, X56793; c, U51197; d, X71970; e, Z22636; f, U40830; g, Z47767; h, X78559; i, L07293; j, AF013583; k, Z83335; l, Z99122.

8.3.6 The *S. pneumoniae* type 23F *cps* locus

The *S. pneumoniae* type 23F *cps* locus has recently been determined from two different isolates (Morona *et al.*, 1998; Ramirez and Tomasz, 1998). The two loci are almost identical but the gene designations differ. Whereas Ramirez and Tomasz (1998) have designated the 18 ORFs *cps23fA-R*, Morona *et al.* (1998) have designated the genes *cps23fA-E, I, J, L-O* and *T-Z*. The latter gene designations are used below because they take into account the previously published sequence of the dTDP-Rha biosynthesis genes (*cps23fL-O*) (Coffey *et al.*, 1998a). The predicted functions of the protein products encoded by the *cps23f* genes are listed in **Table 8.7**.

Table 8.7. Predicted functions of *S. pneumoniae cps23f*[a] ORFs

ORF	Predicted function	Similar proteins (% identity)
cps23fA	regulation?	*S. pneumoniae* Cps19fA (96%)
cps23fB	unknown	*S. pneumoniae* Cps19fB (88.5%)
cps23fC	chain length regulation/export	*S. pneumoniae* Cps19fC (69.6%)
cps23fD	chain length regulation/export	*S. pneumoniae* Cps19fD (79.3%)
cps23fE	glucosyl-1-phosphate transferase	*S. pneumoniae* Cps19fE (71.2%)
cps23fT	rhamnosyl transferase	*A. aeolicus* MtfB[b] (18.9%) *Synechocystis* sp. RfbW[c] (22.3%) *A. fulgidus* WbaZ-1[d] (25.5%)
cps23fI	polysaccharide polymerase	-
cps23fU	galactosyl transferase	*S. pneumoniae* Cps14J[e] (33.5%) *S. thermophilus* EpsI[f] (36%) *Y. enterocolitica* TrsB[g] (29.7%)
cps23fV	rhamnosyl transferase	-
cps23fJ	repeat unit transporter	*S. thermophilus* EpsM[f] (32.1%) *S. pneumoniae* Cps14L[e] (16.7%) *Y. enterocolitica* TrsA[g] (17.9%)
cps23fW	glycerol-2-phosphate transferase	*B. subtilis* TagF[h] (27.8%) *S. pneumoniae* type 14 OrfX[e] (37%)
cps23fX	glyceraldehyde-2-phosphate dehydrogenase (CDP-2-glycerol biosynthesis)	*A. fulgidus* GldA[i] (32.3%) *M. jannaschii* GldA[j] (34.4%) *B. subtilis* AraM[k] (28.6%)
cps23fY	glycerol-2-phosphate cytidyltransferase (CDP-2-glycerol biosynthesis)	*A. fulgidus* RfbF[l] (27.5%)
cps23fZ	glyceraldehyde-2-phosphotransferase (CDP-2-glycerol biosynthesis)	*E. coli* NagD[m] (32%) *B. subtilis* AraL[k] (26.3%)
cps23fL	glucose-1-phosphate thymidylyl transferase	*S. pneumoniae* Cps19fL (95.5%)
cps23fM	dTDP-4-keto-6-deoxyglucose-3,5-epimerase	*S. pneumoniae* Cps19fM (99%)
cps23fN	dTDP-glucose-4,6-dehydratase	*S. pneumoniae* Cps19fN (99.7%)
cps23fO	dTDP-L-rhamnose synthase	*S. pneumoniae* Cps19fO (99.3%)

The sequences can be found under GenBank accession nos: a, AF030373 and AF057294; b, .AE000693; c, D64002; d, AE001104; e, X85785; f, U40830; g, Z47767; h, Z99122; i, AE000988; j, V67518; k, Z99118; l, AE001025; m, AF052007.

8.3.7 The *S. pneumoniae* type 33F *cps* locus

The *S. pneumoniae* type 33f *cps* locus has been sequenced by Llull *et al.* (1998) and consists of 15 ORFs, designated *cap33fA-O*. The *cap33fA-O* genes and the predicted functions of their protein products are listed in **Table 8.8**. The authors did not assign functions to Cap33fF, G, I, K, M and O, however on closer examination of the similarities

to other proteins, putative functions could be assigned to all of these proteins, except Cap33fI.

Table 8.8. Predicted functions of *S. pneumoniae cps33f*[a] ORFs

ORF	Predicted function	Similar proteins (% identity)
cap33fA	regulation?	*S. pneumoniae* Cps19fA (94.2%)
cap33fB	unknown	*S. pneumoniae* Cps19fB (85.6%)
cap33fC	chain length regulation/export	*S. pneumoniae* Cps19fC (71.3%)
cap33fD	chain length regulation/export	*S. pneumoniae* Cps19fD (78.4%)
cap33fE	glucosyl-1-phosphate transferase	*S. pneumoniae* Cps14E[b] (70.5%) *S. pneumoniae* Cps19fE (71%)
cap33fF	glycosyl transferase	*K. pneumoniae* RfbC[c] (34%)
cap33fG	glycosyl transferase	*K. pneumoniae* Orf7[d] (28.7%)
cap33fH	galactosyl transferase	*S. pneumoniae* Cps14I[b] (26.5%) *S. pneumoniae* Cps14J[b] (24.9%)
cap33fI	unknown	*S. pneumoniae* Cps14K[b] (22.4%)
cap33fJ	galactosyl transferase	*S. pneumoniae* Cps14J[b] (31.2%)
cap33fK	polysaccharide polymerase	-
cap33fL	repeat unit transporter	*S. pneumoniae* Cps14L[b] (34.5%)
cap33fM	acetyl transferase	*E. coli* YinH[e] (23.7%)
cap33fN	UDP-galactopyranose mutase	*E. coli* Glf[f] (61.6%)
(*cap33fO*)	acetyl transferase	*B. subtilis* YkrP[g] (28.4%)

The sequences can be found under GenBank accession nos: a, AJ006986; b, X85787; c, L41518; d, D21242; e, P37669; f, U09876, g, Z99111.

8.4 Analysis of the *cps* loci from *S. pneumoniae*

8.4.1 Organisation of pneumococcal *cps* loci

All of the *S. pneumoniae cps* loci which have been sequenced to date are located on the *S. pneumoniae* chromosome between *dexB* and *aliA*, and with the exception of type 3, show a remarkably similar arrangement of the *cps* genes (**Fig. 8.4**). For most of these, functions of the gene products have been proposed on the basis of amino acid sequence

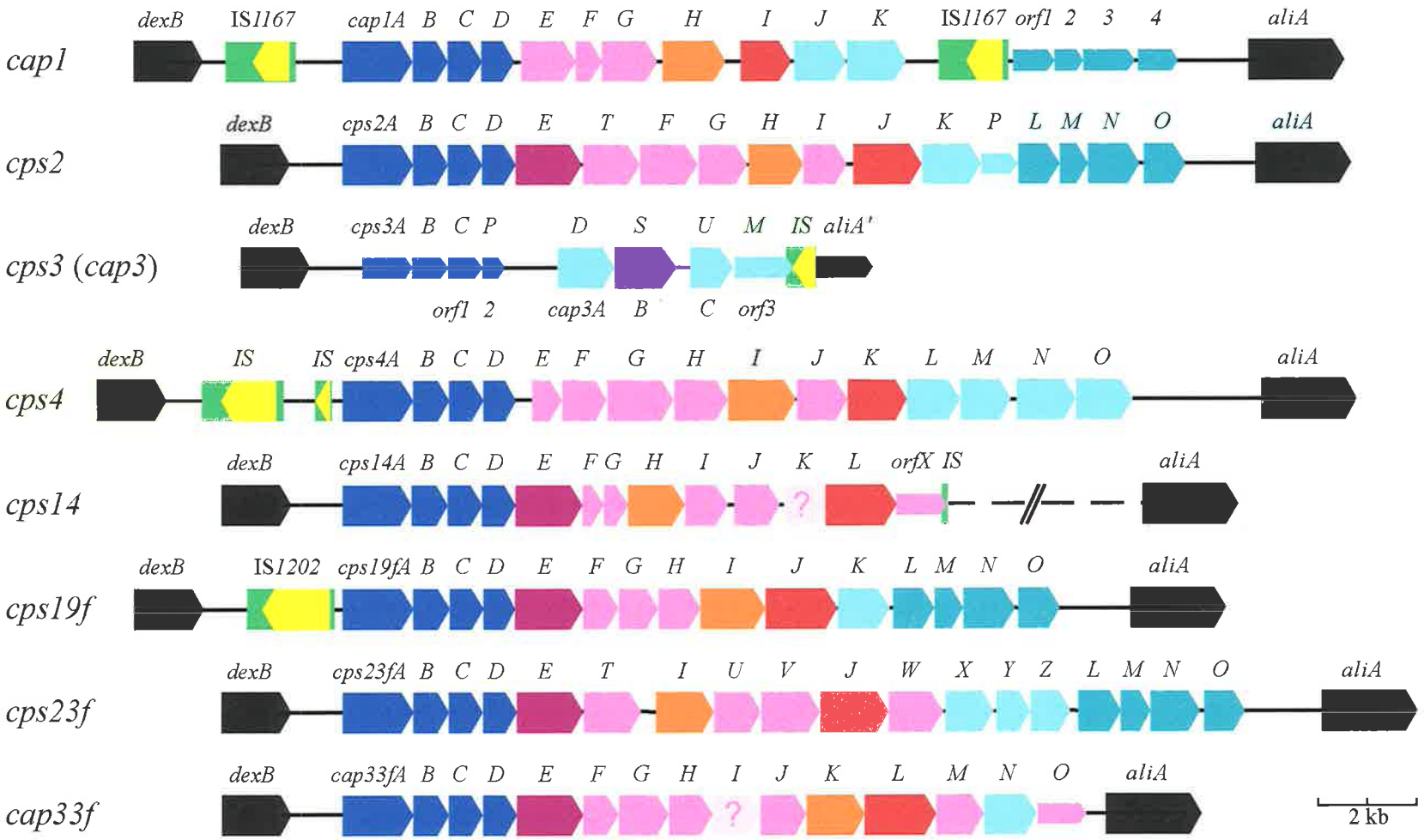


Fig 8.4. The *cps* loci from *S. pneumoniae* serotypes 1, 2, 3, 4, 14, 19F, 23F and 33F. The large boxed arrows represent ORFs involved in CPS biosynthesis and the narrow boxed arrows represent cryptic genes not involved with CPS biosynthesis in that serotype. The ORFs common to all serotypes are shown in dark blue. The specific transferases are shown in pink and the Glc-1-phosphate transferase which adds Glc to the lipid carrier is shown in dark pink. The type 3 synthase is shown in purple, the repeat unit transporters are shown in red and the polysaccharide polymerases are shown in orange. The ORFs involved in nucleotide-sugar biosynthesis are shown in pale blue and those involved in the biosynthesis of dTDP-Rha are shown a shade darker. The green boxes represent IS elements and the yellow boxed arrows represent their ORFs. The dashed line represents the unsequenced region between *dexB* and *aliA* (black boxed arrows). The two ORFs with question marks are of unknown function.

similarities with known proteins. However, Kolkman *et al.* (1996; 1997) have biochemically characterised the four glycosyl transferases encoded by genes in the type 14 *cps* locus. Also, a gene in the type 1 locus (*cap1K*) has been proven to encode a UDP-Glc dehydrogenase by complementation analysis (Muñoz et al., 1997).

The first four genes (*cpsA-D*) are conserved at the 5' end of all loci except type 3, although type 3 retains extensive DNA similarity to these genes. However, there are no intact type 3 ORFs corresponding to *cpsA* and *B* due to several frame-shift mutations and deletions in this region. Two ORFs corresponding to *cpsC* and *D* are present, but *cps3P/orf2* (the *cpsD* homologue) is truncated compared to *cpsD* in other serotypes, and no promoter sequence for these ORFs has been identified in type 3 (García and López, 1997). This is consistent with the distinct mechanism of biosynthesis of type 3 CPS which is described below.

The central portion contains genes which encode the specific glycosyl transferases, the polysaccharide polymerase and the repeat unit transporter required for synthesis of the individual CPSs, and the 3' end contains the genes which encode the biosynthetic enzymes required to synthesise the individual nucleotide sugars. However, this arrangement is not absolute, as the type 14 *cps* locus does not appear to contain any genes which encode biosynthetic enzymes (the sequence downstream of *orfX* has not been determined). However, this is not surprising as type 14 CPS contains only Glc, Glc*N*Ac and Gal, the nucleotide precursors of which are ubiquitous within the cell. Exceptions have also been identified within the group 19 *cps* loci. Types 19B and 19C contain a gene which encodes an enzyme putatively involved in the biosynthesis of an activated Rib precursor (*cps19R*) in their central region, and 19C also contains an additional putative glucosyl transferase gene (*cps19cS*) which has inserted between two genes which encode biosynthetic enzymes (**Fig. 8.1**).

Interestingly, the position of the genes which encode enzymes involved in the synthesis of dTDP-Rha is conserved, at the 3' end of the locus, in all the serotypes examined which contain Rha in their CPS. Type 1 CPS does not contain Rha, but cryptic copies of these genes (designated *orf1-4*) are present in the same position relative to *aliA* in the *cap1* locus of all type 1 isolates examined to date (Muñoz et al., 1997). These cryptic genes are separated from the *cap1* locus by a copy of IS*1167* and having no obvious promoter, are presumably not transcribed. Also *orf4* has a frame-shift mutation which would truncate the protein product. This suggests that the type 1 serotype may have arisen from a recombination event resulting in the insertion of type 1 specific DNA into the *cps* locus of a Rha-containing serotype.

8.4.2 Comparison of the *cps* genes and their protein products

As already described above, the type 3 locus appears to be atypical. The type 3 CPS consists of a simple disaccharide repeat unit (**Table. 1.2**) and the *cps3/cap3* locus contains only three intact genes which are transcribed as a single unit (Arrecubieta *et al.*, 1995; Dillard *et al.*, 1995). The first gene (*cps3D* or *cap3A*) encodes a UDP-Glc dehydrogenase required for the synthesis of UDP-GlcA (Arrecubieta *et al.*, 1994). The second gene (*cps3S* or *cap3B*) encodes the type 3 synthase, a processive β-glycosyl transferase which links the alternating Glc and GlcA moieties via distinct glycosidic bonds (Arrecubieta *et al.*, 1995; Dillard *et al.*, 1995). There is a significant degree of amino acid sequence similarity between Cps3S/Cap3B and other bacterial polysaccharide synthases, including HasA which synthesises the hyaluronic acid capsule of group A streptococci. These synthases have a common predicted architecture and are capable of forming the two different glycosidic linkages, and extruding the growing polysaccharide chain, is as it is synthesised, through a pore or channel formed in the membrane (section 1.8.4; Keenleyside and

Whitfield, 1996). Interestingly, transformation with plasmids carrying *cps3S/cap3B* alone is sufficient to direct synthesis of type 3 CPS in *E. coli* or in *S. pneumoniae* serotypes 1, 2, 5 or 8, all of which contain UDP-GlcA. In smooth heterologous *S. pneumoniae* hosts, expression of *cps3S/cap3B* resulted in binary encapsulated strains producing type 3 as well as the original CPS type (Arrecubieta *et al.* 1996). In the *E. coli* transformants, a significant proportion of the type 3 CPS appeared in the periplasm, indicating that additional type-specific genes are not required for transport of CPS across the cell membrane (Arrecubieta *et al.*, 1996). The final complete gene in the *cps3/cap3* locus (*cps3U* or *cap3C*) encodes a Glc-1-phosphate uridylyltransferase (or UDP-Glc pyrophosphorylase), which is capable of complementing an *E. coli galU* mutant, thereby confirming its function (Arrecubieta *et al.*, 1995). The product of this enzyme (UDP-Glc) is present in all pneumococci, regardless of serotype, and the pneumococcal *galU* gene (a functional homologue of *cps3U/cap3C*) has recently been identified and shown to be located elsewhere in the chromosome (Mollerach *et al.*, 1998). This is consistent with the finding that insertion-duplication mutagenesis of *cps3U/cap3C* does not abrogate type 3 CPS biosynthesis (Arrecubieta *et al.*, 1995; Dillard *et al.*, 1995). On the other hand, insertion-duplication mutagenesis of the *galU* gene, which is 77% identical to *cap3C*, resulted in loss of CPS expression in *S. pneumoniae* type 1 and type 3 (Mollerach *et al.*, 1998). Thus it appears that *cps3U/cap3C* is not capable of replacing the *galU* gene in *S. pneumoniae* either due to poor expression of the type 3 enzyme or poor affinity for the substrate (Mollerach *et al.*, 1998). The *galU* gene is present in all pneumococci tested, regardless of serotype and predicted to be absolutely essential for CPS production as most if not all serotypes contain Glc, and/or other sugars, such as Gal, GlcA, and GalA, which require UDP-Glc for biosynthesis of their activated precursors (Mills and Smith, 1965). No other *S. pneumoniae cps* locus examined to date contains a *galU* homologue. An

additional ORF, designated *cps3M*, is located 3' to *cps3U/cap3C* and has similarity to genes encoding phosphoglucomutases (Dillard *et al.*, 1995). However, it is truncated at its C-terminus and may not encode a functional enzyme; insertion-duplication mutagenesis of *cps3M* also has no impact on encapsulation (Caimano *et al.*, 1998; Dillard *et al.*, 1995).

The other *S. pneumoniae* serotypes have considerably more complex CPSs, consisting of 3-7 sugars, and as a consequence their *cps* loci are also more complex, containing 11 (*cap1*) to 19 (*cps19c*) genes. These loci all contain specific glycosyl transferases, a putative repeat unit transporter and a putative polysaccharide polymerase suggesting a common mode of CPS biosynthesis involving the sequential assembly of the CPS repeat units immobilised on a lipid carrier, followed by transport across the membrane and polymerisation prior to attachment to the cell surface. Thus, CPS biosynthesis in most *S. pneumoniae* serotypes proceeds as described in section 1.8.2 and resembles that of O-antigen biosynthesis in Gram-negative bacteria (Whitfield, 1995).

The serotype specificity of pneumococcal CPSs is determined by both the individual sugars present in the CPS and the nature of the glycosidic bonds joining them. This, in turn, is determined by both the substrate specificity and the nature of the glycosidic linkage formed by the glycosyl transferases encoded by the individual *cps* genes. Thus, each pneumococcal *cps* locus would be expected to include genes which encode mostly unique, potentially serotype determining, glycosyl transferases. For example, serotypes 4, 14, 23F and 33F all contain functionally different galactosyl transferases, which form different glycosyl linkages to different acceptor sugars, in accordance with their distinct CPS structures. Although the amino acid sequences share some similarities, there is no significant DNA similarity between these galactosyl transferases, suggesting that, as expected, they were probably acquired from diverse origins and are unlikely to have arisen from gene duplication.

However, the *cps* loci from all the serotypes examined, except types 1, 3 and 4, contain a homologue of *cps19fE* (70-99% identity). The closely related *cps14E* gene has been shown experimentally to encode a glucosyl transferase which adds Glc-1-phosphate to the lipid carrier in the cell membrane (Kolkman *et al.*, 1996), a common first step in CPS biosynthesis. The absence of a *cps19fE* homologue in types 1 and 4 is consistent with the absence of Glc in their CPS. The only other transferase genes which have significant DNA sequence identity (71.6%) are *cps2T* and *cps23fT* which encode putative β(1→4) rhamnosyl transferases. Both these genes are located in the same position within their respective *cps* loci, just downstream of *cpsE*, raising the possibility that type 2 and type 23F strains may have had a common ancestor. Southern hybridisation data also indicates that a high stringency homologue of *cps19fF*, which encodes a putative Man*N*Ac transferase, is present in type 9N and 9V strains. This transferase is predicted to link Man*N*Ac to Glc via a β(1→4) linkage in group 19 CPS production. Group 9 CPS also contains these two sugars joined by the same linkage, thus the presence of a *cps19fF* homologue in group 9 strains is not surprising.

The DNA similarity between the genes involved in nucleotide sugar biosynthesis in different serotypes is much greater than that observed for transferases. The *cap1J*, *cap3A* and *cps2K* genes all encode UDP-Glc dehydrogenase, which catalyses synthesis of UDP-GlcA; *cap1J* and *cps2K* share 75% identity, and the type 3 gene *cap3A* is slightly more divergent with only 60% identity to both *cap1J* and *cps2K*. UDP-Glc*N*Ac-2-epimerase, which catalyses synthesis of UDP-Man*N*Ac, is encoded by both *cps4L* and *cps19fK* which share 70% identity. Both *cps2P* and *cap33fN* (83.6% identity) are predicted to encode proteins similar to UDP-galactopyranose mutase which converts UDP-galactopyranose to UDP-galactofuranose. However, the *cps2P* gene is truncated at the 5' end compared to *cap33fN* and as UDP-galactofuranose is not required for type 2 CPS biosynthesis, *cps2P*

may be a cryptic gene within the *cps2* locus. The four dTDP-Rha biosynthesis genes are highly conserved in all serotypes containing Rha in their CPS, and have 90-99% nucleotide sequence identity. The high degree of sequence identity and the conserved position of these genes at the 3' end of the respective loci suggests the possibility of a common ancestor for these serotypes, and for type 1 which contains cryptic copies of these genes.

Both type 14 and type 33F *cps* loci contain copies of genes which encode redundant transferases at the 3' end of the loci. The redundant acetyl transferase encoded by *cap33fO* is truncated due to a frame-shift mutation in type 33F but is intact in the type 33A *cps* locus (Llull *et al.*, 1998). The *orfX* gene at the 3' end of the *cps14* locus is also truncated (Kolkman *et al.*, 1997b), but a glycerol-phosphate transferase, the predicted product of this gene, would be required for the biosynthesis of the immuno-cross reactive type 15A CPS (van Dam *et al.*, 1990). These redundant genes may be the remnants of a recombinational event within an ancestral *cps* locus, which resulted in the generation of a new serotype for which these distal genes were no longer required.

8.4.3 Capsular transformation *in vivo*

There is growing evidence that the phenomenon of capsular transformation first observed by Griffith (section 1.1.6) is a common phenomenon *in vivo*. Application of modern molecular typing techniques has resulted in the detection of otherwise genetically indistinguishable pneumococci expressing different capsular types. This has been particularly evident in clonal groups which are resistant to multiple antibiotics. Indeed, derivatives of a highly successful multiply-resistant type 23F clone (which originated in Spain) expressing types 3, 9N, 14, 19A and 19F capsules have been isolated (Barnes *et al.*, 1995; Coffey *et al.*, 1991; Coffey *et al.*, 1998a; Nesin *et al.*, 1998).

Completion of the *cps19f* sequence enabled further characterisation of a parental multiresistant type 23F strain and eight otherwise genetically indistinguishable type 19F clinical isolates. This involved PCR amplification, and subsequent sequencing, of the regions from *dexB-cpsB* and from *cpsL-aliA* for all of these strains (Coffey *et al.*, 1998b). Examination of polymorphisms in the conserved regions of the two *cps* loci indicated that in each case, the 5' recombination occurred upstream of *dexB*. In six of the eight type 19F strains, the 3' cross-over point was downstream of *aliA*. However, in the other two, a recombination cross-over point between the introduced type 19F sequences and the type 23F chromosome was identified; this was in *cpsM* in one strain and *cpsN* in the other (Coffey *et al.*, 1998b). Thus, capsule switching involves exchange of very large DNA fragments, ranging from at least 15 kb to over 22.5 kb. The existence of multiple cross-over points as well as additional minor polymorphisms within the type 19F-derived *cps* genes also indicated that the eight multiply resistant type 19F strains that were studied arose as a consequence of a minimum of four independent transformation events involving different type 19F donors. It therefore appears that these capsule switching events may be relatively common among pneumococci in nature (Coffey *et al.*, 1998b).

Multiple serotypes of *S. pneumoniae* are frequently carried concurrently in the human nasopharynx (Austrian, 1981b) providing ample opportunity for exchange of DNA between types. In addition to enhancing the spread of drug resistance amongst diverse capsular types, these exchanges may also provide a mechanism for evasion of serotype-specific host immune defences, such as those resulting from immunisation with pneumococcal CPS-protein conjugate vaccines which provide cover against a limited range of serotypes.

8.4.4 The presence of IS elements in the 5' and 3' intergenic regions

Complete or partial IS elements have been located adjacent to 9 of the 11 *cps* loci sequenced to date, and in type 1 the locus is flanked at both the 5' and 3' ends by IS*1167*. The first IS element identified just upstream of the *S. pneumoniae cps* loci was IS*1202* which is associated with all four members of group 19 (Morona *et al.*, 1994a). Since then several more IS elements have been identified in both the 5' and 3' intergenic regions of *S. pneumoniae cps* loci. In fact, 5' and 3' intergenic regions themselves contain sequences with some similarity to IS elements, and are repeated many times in the chromosome, although the complete IS elements are no longer present.

The *cps*-flanking regions appear to be common targets for IS elements, and this has led to the suggestion that they may play a role in horizontal transfer of *cps* genes (Kolkman *et al.*, 1998; Muñoz *et al.*, 1997). There are precedents for this in other bacteria; the *H. influenzae* type b capsule genes, for example, are located on a 17-kb compound transposon (Kroll *et al.*, 1991). However, the identification of cross-over points within the *cps* loci of the type 19F derivatives of the multiresistant type 23F *S. pneumoniae* clone referred to above confirm that at least in these cases, capsular exchange occurred as a consequence of homologous recombination rather than by transposition. Nevertheless, Muñoz *et al.* (1997) have demonstrated that IS*1167* sequences flanking part of the *cap1* locus cloned in a plasmid could direct ectopic integration of these genes into copies of IS*1167* located elsewhere in the pneumococcal chromosome, resulting in genetically binary strains.

8.4.5 Transcription of the pneumococcal *cps* locus

As mentioned previously, the only promoter sequence identified in the *S. pneumoniae cps19f* locus is immediately upstream of *cps19fA* (Guidolin et al., 1994), and the only stemmed-loop structures likely to function as transcription terminators are found

downstream of the *cps19f* locus in the 3' intergenic region (section 3.2.2). The consensus σ^{70} promoter sequence (TAGACA-17 bp-TATAAT) is also present in the same position in all of the other *cps* loci which have been sequenced, except type 3, which has a deletion of 280 nucleotides at the 5' end of the *cpsA*-related region resulting in the loss of the promoter sequence. The three *cps3* genes are transcribed as a single operon from a promoter immediately upstream of *cps3D* (*cap3A*) (Arrecubieta *et al.*, 1995). Llull *et al.* (1998) have shown that the promoter sequences identified for the *cps* loci in types 1, 3, 14 and 33F are all functional in *S. pneumoniae*. The transcriptional start site has also been determined by primer extension analysis of the type 1 locus (Muñoz *et al.*, 1997).

It is possible, however, that the level of transcription of *cps* loci varies from strain to strain. Although the -10 and -35 sequences themselves are highly conserved, Llull *et al.* (1998) have noted minor variations in flanking sequences. In type 37, for example, the -10 and -35 sequences are separated by 16 nucleotides rather than 17, and this correlated with a markedly lower promoter strength. A 4-nucleotide deletion immediately 5' to the -35 sequence in type 33F, relative to type 1, also correlated with a slight reduction in promoter strength (Llull *et al.*, 1998). The extent to which such differences impact on the level of encapsulation is uncertain, but it is curious in the light of these findings that type 37, which had the weakest *cps* promoter, produces one of the thickest capsules of all *S. pneumoniae* serotypes (Austrian, 1981b).

Many polysaccharide loci from Gram-negative bacteria (including: *E. coli kps*, *E. coli* K12 *rfa*, *S. flexneri rfb*, *S. enterica rfb*, *Vibrio cholerae rfb*, and *Y. enterocolitica rfb*) have a large leader sequence which is involved in transcriptional regulation of the locus (Marolda and Valvano, 1998). This region includes the conserved JUMPstart sequence (Hobbs and Reeves, 1994) which is part of two *cis*-acting sequences known as *ops* (for operon polarity suppressor) and are located upstream of the coding regions of all RfaH-

regulated operons (Marolda and Valvano, 1998). These leader sequences form a series of stem-loop structures; one of these contains the JUMPstart and *ops* sequences and interacts with RfaH, a transcription anti-terminator, and the RNA polymerase. In the absence of RfaH or with mutations affecting this stemmed-loop, premature termination of the transcript occurs when the RNA polymerase reaches the other stemmed-loop structures in this region. In the presence of RfaH, the JUMPstart stemmed-loop would serve to bring RfaH and possibly other factors together with the RNA polymerase complex, preventing the formation of the other stemmed-loops and hence, premature termination (Marolda and Valvano, 1998). However, in *S. pneumoniae cps* loci this leader sequence is entirely absent, with only 20 bp separating the transcriptional start site and the ATG start codon of *cpsA*. This is too short for formation of any stemmed-loop structures. Thus, transcription of the *cps* genes and hence CPS production in *S. pneumoniae* is not regulated in the same manner as polysaccharide production in Gram-negative bacteria.

8.4.6 Regulation of CPS production in *S. pneumoniae*

Colonisation of the nasopharyngeal mucosa is an essential first step in the pathogenesis of pneumococcal disease, and is presumed to involve interaction between pneumococcal adhesins and specific receptors on host epithelial cells. In a proportion of cases, asymptomatic carriage progresses to invasive disease, and although the events involved are poorly understood, it is clearly a watershed in the bacteria-host relationship. In recent years evidence has emerged that this transition involves a major switch in expression of important virulence determinants, as the pneumococcus adapts to the altered microenvironment (Tuomanen and Masure, 1997). Maximal expression of capsule is clearly essential for systemic virulence, but the degree of exposure of other important pneumococcal surface structures, such as the adhesins, may also be influenced by capsular

thickness. Non-encapsulated pneumococci exhibit higher adherence to human respiratory epithelial (A549) cells *in vitro* than otherwise isogenic derivatives expressing either type 3 or type 19F capsules (Talbot *et al.*, 1996). Thus, the very feature (encapsulation) which is absolutely essential for systemic virulence of *S. pneumoniae* could be disadvantageous during the colonisation phase. Pneumococci have recently been shown to undergo a bidirectional phase variation between two distinct colonial morphologies, described as "opaque" and "transparent". The transparent phenotype exhibits increased *in vitro* adherence to buccal epithelial cells and cytokine-activated A549 cells relative to opaque variants of the same strain, as well as an enhanced capacity to colonise the nasopharynx of infant rats (Cundell *et al.*, 1995; Weiser *et al.*, 1994). On the other hand, the opaque form is associated with massively increased virulence in animal models of systemic disease, and this correlates with increased production of CPS relative to cell wall teichoic acid compared with the transparent phenotype (Kim and Weiser, 1998). Phase variation also correlated with alteration in levels of several surface proteins, but the molecular mechanism involved is yet to be elucidated.

Clearly, the capacity to regulate CPS production is important for the survival of the pneumococcus in different host environments. The presence of a 115-bp repeated element has been noted upstream of the *cps* promoter in *S. pneumoniae* types 1, 3, 14, and 19F. This element appears to be specific to pneumococci, and copies have been found in the vicinity of other genes believed to be associated with virulence (Kolkman *et al.*, 1997b; Muñoz *et al.*, 1997). These sequences have no obvious function, although Kolkman *et al.* (1997) have suggested that they might have a regulatory function for coordinately controlled expression of virulence-related genes. However, examination of sequence data for *S. pneumoniae* type 4 released by TIGR indicates that there are at least 40 copies of sequences with > 80% identity to the element in the genome. In type 19F, the copy near

the *cps* locus is actually separated from the *cps* promoter by IS*1202*, and so it is difficult to imagine its involvement in a global regulatory mechanism. Moreover, insertion of the mutagenesis vector pVA891 into IS*1202*, which places the 115-bp element more than 8 kb upstream of the promoter, is known not to affect type 19F CPS production (Morona *et al.*, 1994a). As previously mentioned, in type 19F, IS*1202* is inserted immediately 5' to the -35 sequence of the *cps* promoter and so the above result, in all probability, precludes the involvement of any upstream region in transcriptional regulation of *cps19f*.

At present the mechanism of regulation of CPS production in *S. pneumoniae* is not understood. The only product of the *cps* loci likely to be involved is CpsA, which resembles a *B. subtilis* transcriptional attenuator. Recently, a derivative of *S. pneumoniae* type 19F with an in-frame deletion mutation in *cps19fA* was constructed. This mutant still produces type 19F CPS, as judged by quellung reaction. Also, colonies growing on BA appeared smooth, but they were much smaller than the parental type 19F strain, in spite of the lack of any obvious difference in growth rate (J. Morona, unpublished data). Additional phentoypic characterisation of this mutant, including transcriptional studies of *cps19f* and quantitation of total CPS production are in progress.

It is also possible that genes located elsewhere on the chromosome may influence CPS production in *S. pneumoniae*. Watson *et al.* (1995) suggested that a region 3' to the pneumococcal *lytA* (autolysin) gene might be involved in regulation of capsular expression, as this appeared to be the site of Tn*916* insertion in a type 3 mutant which had lost the capacity to produce CPS. However, García *et al.* (1996) subsequently demonstrated that deletion of this region of the chromosome had no impact an encapsulation. An additional possibility is that regulation of CPS production is indirect, and mediated by availability of precursors or co-factors. For example, the mechanism of synthesis of the pneumococcal cell wall teichoic acid (see **Fig. 1.2**) is likely to be similar to CPS. Accordingly, if teichoic

acid synthesis is subject to direct regulation, this may influence the availability of common precursors (e.g. UDP-Glc or the lipid carrier) for CPS biosynthesis. Such a phenomenon would be consistent with the observation that phase variation in *S. pneumoniae* has opposite effects on the total amounts of CPS and teichoic acid (Kim and Weiser, 1998). Saturation of a common lipid carrier pool with the disaccharide precursor for peptidoglycan synthesis has previously been proposed to explain blockade of EPS production in *Streptococcus thermophilus* with a mutation in *pbp2b*, one of its penicillin-binding protein genes (Stingele and Mollet, 1996). Similarly, competition for the supply of GDP-mannose between the alginate and LPS biosynthetic pathways is thought to explain why mucoid *P. aeruginosa* strains lack the LPS O-antigen (May and Chakrabarty, 1994).

8.4.7 The function of CpsC and CpsD

The *cpsC* and *D* gene products are predicted to function in CPS chain length regulation and export in an analogous fashion to ExoP from *R. meliloti* (Paulsen *et al.*, 1997). CpsC has two putative membrane-spanning hydrophobic domains which are found in all chain length regulators (Rol- and Cld-like proteins) and its topology is predicted to be similar these proteins. Thus, the N- and C-termini are predicted to be located in the cytoplasm, while the central portion is predicted to be exposed on the external side of the cell membrane. CpsD is presumed to be a cytoplasmic protein, and like the C-terminal portion of ExoP, contains a putative ATP binding domain (Guidolin *et al.*, 1994). The C-terminal "CpsD" domain of ExoP is required for synthesis of high molecular weight EPS (Becker *et al.*, 1995). This observation suggests that CpsD may interact with CpsC in the pneumococcus. Recently, derivatives of *S. pneumoniae* type 19F with in-frame deletion mutations in *cps19fC* and *cps19fD* were constructed. These mutants no longer produced

type 19F CPS, as judged by quellung reaction (J. Morona, unpublished data), indicating that both CpsC and CpsD are important for CPS production in *S. pneumoniae*.

Two distinct classes of *cpsC* and *cpsD* genes (class I and class II) have been identified in pneumococcal *cps* loci. However, the exact mechanism by which CpsC and CpsD (and their homologues) regulate polysaccharide chain length is still unknown. The effect of the two different classes on CPS chain length, and hence, capsule thickness await analysis of defined strains in which the class I and class II genes have been interchanged.

Genes which encode functional homologues of both CpsC and D are found in many polysaccharide biosynthesis loci from a large variety of Gram-negative and Gram-positive bacteria. The CpsC homologues (Rol- and Cld-like proteins) are present in all polysaccharide loci where biosynthesis occurs via lipid-linked repeat unit intermediates. However, CpsD homologues are not always present; they are not found in loci involved in O-antigen biosynthesis, but they are usually present in loci involved in CPS and EPS biosynthesis. This may have implications for the function of CpsC homologues in O-antigen and CPS biosynthesis. The average number of repeat units is much larger for CPS than for O-antigen, and CpsD may be required to enable CpsC to function such that a long polysaccharide chain can be synthesised (R. Morona, unpublished results). Gram-negative CPS and EPS biosynthesis loci generally contain a gene encoding one large protein (such as ExoP in *R. meliloti*, AmsA in *E. amylovora* and Orf6 in *K. pneumoniae* O1:K2) with separate domains homologous to either CpsC or CpsD (Whitfield *et al.*, 1997). However, in Gram-positive CPS and EPS biosynthesis loci, two separate genes encode CpsC and CpsD homologues, as discussed below.

8.4.8 Relationship of the *S. pneumoniae cps* locus to other Gram-positive *cps* loci

To date, CpsA and CpsB homologues have only been found in ORFs cloned from

other Gram-positive bacteria, indicating that their function is probably Gram-positive specific. The only protein of known function similar to either of these proteins is LytR (a transcriptional attenuator of the *B. subtilis* autolytic enzymes) which has 28% identity to Cps19fA. However, the function of CpsA in pneumococcal CPS production has yet to be determined, as does that for CpsB. Homologues to CpsC and CpsD are more wide-spread, found in both Gram-positive and Gram-negative polysaccharide biosynthesis loci, as described above.

The presence of homologues of the first four *S. pneumoniae cps* genes, *cpsA-D*, was examined in other Gram-positive polysaccharide loci. Interestingly, homologues to CpsA were only found in other *Streptococcus* species, with the first four genes arranged in the same order in all of the loci examined (**Table 8.9**). CPS and EPS biosynthesis loci from other Gram-positive bacteria contained homologues to CpsB, C and D, also in the 5' region of the loci, but the gene order differed to that in the *Streptococcus* species. In *S. aureus* types 1, 5 and 8 capsules, the *L. lactis* EPS and a polysaccharide from *B. subtilis*, the homologues are ordered *cpsC*, *D* and then *B*. However, the *cpsC* homologue is not necessarily the first gene in the locus; in both the *eps* locus from *L. lactis* and the *ywq* locus from *B. subtilis*, there are two genes preceding the *cpsC* homologue (unrelated to the pneumococcal *cpsA* or *B* genes) one of which encodes a protein (EpsR) with similarity to DNA-binding regulatory proteins (van Kranenburg *et al.*, 1997). Interestingly, an ORF designated *orfY*, located after the transcription terminator at the 3' end of the *eps* locus from *Lactococcus lactis*, and transcribed in the opposite direction to the *eps* genes, has significant similarity to LytR (23.6% identity), but not to CpsA. Any role this gene may have in EPS production in *L. lactis* has yet to be determined (van Kranenburg *et al.*, 1997). In group B streptococci, an additional gene located upstream of the *cpsA* homologue, but transcribed in the opposite direction, encodes a protein (CpsY) with similarity to the LysR

family of DNA-binding transcriptional regulators (Koskiniemi *et al.*, 1998). However, no such gene is present adjacent to any of the pneumococcal *cps* loci examined to date. Thus, it appears that distinct transcriptional regulatory mechanisms may exist for polysaccharide biosynthesis loci in different Gram-positive species.

Table 8.9. CpsA-D homologues present in polysaccharide biosynthesis loci of Gram-positive bacteria

Homologues to:	CpsA	CpsB	CpsC	CpsD
Streptococcus pneumoniae type 19F	Cps19fA	Cps19fB	Cps19fC	Cps19fD
Streptococcus pneumoniae type 19A	Cps19aA	Cps19aB	Cps19aC	Cps19aD
Streptococcus pneumoniae type 19B	Cps19bA	Cps19bB	Cps19bC	Cps19bD
Streptococcus pneumoniae type 19C	Cps19cA	Cps19cB	Cps19cC	Cps19cD
Streptococcus pneumoniae type 1[a]	Cap1A	Cap1B	Cap1C	Cap1D
Streptococcus pneumoniae type 2[b]	Cps2A	Cps2B	Cps2C	Cps2D
Streptococcus pneumoniae type 4[c]	Cps4A	Cps4B	Cps4C	Cps4D
Streptococcus pneumoniae type 14[d]	Cps14A	Cps14B	Cps14C	Cps14D
Streptococcus pneumoniae type 23F[e]	Cps23fA	Cps23fB	Cps23fC	Cps23fD
Streptococcus pneumoniae type 33F[f]	Cap33fA	Cap33fB	Cap33fC	Cap33fD
Streptococcus thermophilus[g]	EpsA	EpsB	EpsC	EpsD
Streptococcus salivarius[h]	CpsA	CpsB	CpsC	CpsD
Streptococcus agalactiae[i]	CpsX	CpsA	CpsB	CpsC
Staphylococcus aureus type 1[j]		Cap1C	Cap1A	Cap1B
Staphylococcus aureus type 5[k]		Cap5C	Cap5A	Cap5B
Staphylococcus aureus type 8[l]		Cap8C	Cap8A	Cap8B
Lactococcus lactis[m]		EpsC	EpsA	EpsB
Bacillus subtilis[n]		YwqE	YwqC	YwqD

a, Muñoz *et al.* (1997); b, Iannelli *et al.* (1998); c, TIGR Microbial Database; d, Kolkman *et al.* (1997); e, Morona *et al.* (1998); f, Llull *et al.* (1998); g, Stingele *et al.* (1996); h, Griffin *et al.* (1996); i, Koskiniemi *et al.* (1998) and Rubens *et al.* (1993); j, Wen *et al.* (1994); k, Sau *et al.* (1997); l, Sau and Lee (1996); m, van Kranenburg *et al.* (1997); n, Presecan *et al.* (1997).

The polysaccharide loci examined above are all presumed to share a similar mode of biosynthesis, involving the assembly of repeat units on lipid carriers which are exported prior to polymerisation. The *S. pneumoniae* type 3 and the Group A Streptococcus

hyaluronic acid CPS loci were not included in **Table 8.9** because the mode of CPS biosynthesis differs, involving a synthase which exports the growing polysaccharide chain, as described in sections 1.8.4 and 8.4.2.

The similarities between the CPS and EPS loci listed in **Table 8.9** extend over the entire locus. They all appear to be arranged as tightly linked, single transcriptional units with the common genes located at the 5' end. The serotype-specific genes encoding specific transferases, a repeat unit transporter, and a polysaccharide polymerase are located in the central portion. The genes encoding enzymes involved in biosynthesis of nucleotide sugars are generally located at the 3' end of the loci. This arrangement has some similarity to CPS loci in Gram-negative bacteria which are divided into three separate regions with the serotype-specific region in the centre (Roberts, 1996; section 1.8.6).

8.5 Future studies

Much remains to be learnt about the precise molecular events involved in the mechanisms of CPS biosynthesis, and about how CPS production in pneumococci is regulated. Detailed biochemical and mutational analyses are also required to confirm proposed functions of specific genes and in particular to elucidate the precise functions of the four genes at 5' end of the *cps* loci. These presumably encode important common steps in polysaccharide biosynthesis in pneumococci, as well as in certain other Gram-positive genera. As a first step, in-frame deletion mutants of *cps19fA*, *cps19fC* and *cps19fD* have already been constructed in *S. pneumoniae* strain Rx1-19F. Given the importance of capsules to the virulence of *S. pneumoniae* and several other Gram-positive pathogens, such conserved components of the CPS biosynthesis machinery may prove to be useful targets for novel antimicrobial strategies.

REFERENCES

Aarson, VA., Kristinsson, KG., Sigurdsson, JA., Stefansdottir, G., Molstad, S. and Gudmundsson, S. 1996. Do antimicrobials increase the carriage rate of penicillin resistant pneumococci in children? Cross sectional prevalence study. Br. Med. J. 313:387-391.

Adams, WG., Deaver, KA., Cochi, SL., Plikaytis, BD., Zell, ER., Broome, CV. and Wenger, JD.; for the *Haemophilus influenzae* Study Group. 1993. Decline of childhood *Haemophilus influenzae* Type b (Hib) disease in the Hib vaccine era. JAMA. 269:221-226.

Alloing, G., de Philip, P. and Claverys, J-P. 1994, Three highly homologous membrane-bound lipoproteins participate in oligopeptide transport by the Ami system of the Gram-positive *Streptococcus pneumoniae*. J. Mol. Biol. 241:44-58.

Alloway, H. 1932. The transformation in vitro of R pneumococci into S forms of different specific types by use of filtered pneumococcus extracts. J. Exp. Med. 55:91-99.

Alloway, H. 1933. Further observations on the use of pneumococcus extracts in effecting transformation of type in vitro. J. Exp. Med. 57:265-278.

AlonsoDeVelasco, E., Verheul, AFM., Verhoef, J. and Snippe, H. 1995. *Streptococcus pneumoniae*: Virulence factors, pathogenesis, and vaccines. Microbiol. Rev. 59:591-603.

Altschul, SF., Gish, W., Miller, W., Myers, EW. and Lipman, DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Altschul, SF., Madden, TL., Schaffer, AA., Zhang, J., Zhang, Z., Miller, W. and Lipman, DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Appelbaum, PC. 1992. Antimicrobial resistance in *Streptococcus pneumoniae*: An overview. Clin. Infect. Dis. 15:77-83.

Arakawa, Y., Wacharotayankun, R., Nagatsuka, T., Ito, H., Kato, N. and Ohta, M. 1995. Genomic organisation of the *Klebsiella pneumoniae cps* region responsible for serotype K2 capsular polysaccharide synthesis in the virulent strain Chedid. J. Bacteriol. 177:1788-1796.

Armstrong, RR. 1931. A swift and simple method for detecting pneumococcal "type". Br. Med. J. 1:214-215.

Arrecubieta, C., García, E. and López, R. 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.

Arrecubieta, C., López, R. and García, E. 1994. Molecular characterisation of cap3A, a gene from the operon required for the synthesis of the capsule of *Streptococcus pneumoniae* type 3: sequencing of mutations responsible for the unencapsulated phenotype and localisation of the capsular cluster on the pneumococcal chromosome. J. Bacteriol. 176:6375-6383.

Arrecubieta, C., López, R. and García, E. 1996. Type 3-specific synthase of *Streptococcus pneumoniae* (Cap3B) directs type 3 polysaccharide biosynthesis in *Escherichia coli* and in pneumococcal strains of different serotypes. J. Exp. Med. 184:449-455.

Austrian, R. 1981a. Pneumococcus: The first one hundred years. Rev. Infect. Dis. 3:183-189.

Austrian, R. 1981b. Some observations on the pneumococcus and on the current status of pneumococcal disease and its prevention. Rev. Infect. Dis. 3(Suppl.):S1-S17.

Austrian, R., Bernheimer, HP., Smith, EEB. and Mills, GT. 1959. Simultaneous production of two capsular polysaccharides by pneumococcus. II. The genetic and biochemical bases of binary capsulation. J. Exp. Med. 110:585-602.

Avery, OT. and Dubos, R. 1931. The protective action of a specific enzyme against type 3 pneumococcus infection in mice. J. Exp. Med. 54:73-89.

Avery, OT., MacLeod, CM. and McCarty, M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J., Exp. Med. 79:137-157.

Barnes, DM., Whittier, S., Gilligan, PH., Soares, S., Tomasz, A. and Henderson, FW. 1995. Transmission of multidrug-resistant serotype 23F *Streptococcus pneumoniae* in group day care: evidence suggesting capsular transformation of the resistant strain *in vivo*. J. Infect. Dis. 171:890-896.

Barry, MA., Craven, DE. and Finland, M. 1984. Serotypes of *Streptococcus pneumoniae* isolated from blood cultures at Boston City Hospital between 1979 and 1982. J. Infect. Dis. 149:449-452.

Bechthold, A., Sohng, JK., Smith, TM., Chu, X. and Floss, HG. 1995. Identification of *Streptomyces violaceoruber* Tu22 genes involved in the biosynthesis of granaticin. Mol. Gen. Genet. 248:610-620.

Becker, A., Niehaus, K. and Pühler, A. 1995. Low-molecular-weight succinoglycan is predominantly produced by Rhizobium meliloti strains carrying a mutated ExoP protein characterised by a periplasmic N-terminal domain and a missing C-terminal domain. Mol. Microbiol. 16:191-203.

Bernheimer, HP., Wermundsen, IE. and Austrian, R. 1967. Qualitative differences in the behaviour of pneumococcal deoxyribonucleic acids transforming to the same capsular type. J. Bacteriol. 93:320-333.

Berry, AM., Lock, RA., Thomas, SM., Rajan, DP., Hansman, D. and Paton, JC. 1994. Cloning and nucleotide sequence of the *Streptococcus pneumoniae* hyaluronidase gene, and purification of the enzyme from recombinant *Escherichia coli*. Infect. Immun. 62:1101-1108.

Berry, AM., Lock, RA. and Paton, JC. 1996. Cloning and characterisation of *nanB*, a second *Streptococcus pneumoniae* neuraminidase gene, and purification of the NanB enzyme from recombinant *Escherichia coli*. J. Bacteriol. 16:4854-4860.

Berry, AM. and Paton, JC. 1996. Sequence heterogeneity of PsaA, a 37-kilodalton putative adhesin essential for virulence of *Streptococcus pneumoniae*. Infect. Immun. 64:5255-5262.

Berry, AM., Paton, JC., Glare, EM. Hansman, D. and Catcheside, DEA. 1988. Cloning and expression of the pneumococcal neuraminidase gene in *Escherichia coli*. Gene. 71:299-305.

Berry, AM., Yother, J., Briles, DE., Hansman, D. and Paton, JC. 1989. Reduced virulence of a defined pneumolysin-negative mutant of *Streptococcus pneumoniae*. Infect. Immun. 57:2324-2330.

Beynon, LM., Richards, JC., Perry, MB. and Kniskern. 1991. Antigenic and structural relationships within group 19 *Streptococcus pneumoniae*: chemical characterisation of the specific capsular polysaccharides of type 19B and 19C. Can. J. Chem. 70:131-137.

Bluestone, CD., Stephenson, JS. and Martin, LM. 1992. Ten-year review of otitis media pathogens. Pediatr. Infect. Dis. J. 11:S7-S11.

Bogaerts, J., Lepage, P., Taelman, H., Rouvroy, D., Batungwanayo, J., Kestelyn, P., Hitimana, DG., Van de Perre, P., Vandepitte, J., Verbist, L. and Verhaegen, J. 1993. Antimicrobial susceptibility and serotype distribution of *Streptococcus pneumoniae* from Rwanda, 1984-1990. J. Infect. 27:157-168.

Briles, DE., Yother, J. and McDaniel, LS. 1988. Role of pneumococcal surface protein A in the virulence of *Streptococcus pneumoniae*. Rev. Infect. Dis. 3:S82-S88.

Bridge, PD. and Sneath, PHA. 1983. Numerical taxonomy of Streptococcus. J. Gen. Microbiol. 129:565-597.

Bronner, D., Clarke, BR. and Whitfield, C. 1994. Identification of an ATP-binding cassette transport system required for translocation of lipopolysaccharide O-antigen side-chains across the cytoplasmic membrane of *Klebsiella pneumoniae* serotype O1. Mol. Microbiol. 14:505-519.

Brown, MCM., Weston, A., Saunders, JR. and Humphreys, GO. 1979. Transformation of *E. coli* C600 by plasmid DNA at different phases of growth. FEMS Microbiol. Lett. 5:219-222.

Brown, PK., Romana, LK. and Reeves, PR. 1992. Molecular analysis of the *rfb* gene cluster of *Salmonella* serovar muenchen (strain M67): the genetic basis of the polymorphism between groups C2 and B. Mol. Microbiol. 6:1385-1394.

Brynestad, S., Synstad, B. and Granum, PE. 1997. The *Clostridium perfringens* enterotoxin gene is on a transposable element in type A human food poisoning strains. Microbiol. 143:2109-2115.

Bugert, P. and Geider, K. 1995. Molecular analysis of the *ams* operon required for exopolysaccharide synthesis of *Erwinia amylovora*. Mol. Microbiol. 15:917-933.

Burke, JP., Klein, JO., Gezon, HM. and Finland, M. 1971. Pneumococcal bacteraemia. Amer. J. Dis. Child. 121:353-359.

Butler, JC., Breiman, RF., Lipman, HB., Hofmann, J. and Facklam, RR. 1995. Serotype distribution of *Streptococcus pneumoniae* infections among preschool children in the United States, 1978-1994: Implications for development of a conjugate vaccine. J. Infect. Dis. 171:885-889.

Cabib, E., Bowers, B. and Roberts, RL. 1983. Vectorial synthesis of a polysaccharide by isolated plasma membranes. Proc. Natl. Acad. Sci. USA. 80:3318-3321.

Caimano, MJ., Hardy, GG. and Yother, J. 1998. Capsule genetics in *Streptococcus pneumoniae* and a possible role for transposition in the generation of the type 3 locus. Microb. Drug. Resist. 4:11-23.

Camara, M., Mitchell, TJ, Andrew, PW. and Boulnois, GJ. 1991. *Streptococcus pneumoniae* produces at least two distinct enzymes with neuraminidase activity: cloning and expression of a second neuraminidase gene in *Escherichia coli*. Infect. Immun. 59:2856-2858.

Centers for Disease Control and Prevention. 1997. Prevention of pneumococcal disease: Recommendations of the Advisory Committee on Immunisation Practices (ACIP). MMWR 46(No. RR-8):1-24.

Cherian, T., Steinhoff, MC., Harrison, LH., Rohn, D., McDougal, L. and Dick, J. 1994. A cluster of invasive pneumococcal diseases in young children in child care. JAMA. 271:695-698.

Coffey, TJ., Dowson, CG., Daniels, M., Zhou, J., Martin, C., Spratt, BG. and Musser, JM. 1991. Horizontal gene transfer of multiple penicillin-binding protein genes and capsular biosynthetic genes in natural populations of *Streptococcus pneumoniae*. Mol. Microbiol. 5:2255-2260.

Coffey, TJ., Enright, MC., Daniels, M., Morona, JK., Morona, R., Hryniewicz, W., Paton, JC. and Spratt. BG. 1998a. Recombinational exchanges at the capsular polysaccharide biosynthetic locus lead to frequent serotype changes among natural isolates of *Streptococcus pneumoniae*. Mol. Microbiol. 27:73-83.

Coffey, TJ., Enright, MC., Daniels, M., Wilkinson, P., Berron, S., Fenoll, A. and Spratt, BG. 1998b. Serotype 19A variants of the Spanish serotype 23F multiresistant clone of *Streptococcus pneumoniae*. Microb. Drug Resist. 4:51-55.

Collignon, PJ. and Bell, JM. 1996. Drug resistant *Streptococcus pneumoniae*: the beginning of the end for many antibiotics? Med. J. Aust. 164:64-67.

Collins, LV. and Hackett, J. 1991. Molecular cloning, characterisation, and nucleotide sequence of the *rfc* gene, which encodes an O-antigen polymerase of *Salmonella typhimurium*. J. Bacteriol. 173:2521-2529.

Crain, MJ., Waltman, WD. II, Turner, JS., Yother, J., Talkington, DF., McDaniel, LS., Gray, BM. and Briles, DE. 1990. Pneumococcal surface protein A (PspA) is serologically highly variable and is expressed by all clinically important capsular serotypes of *Streptococcus pneumoniae*. Infect. Immun. 58:3293-3299.

Cundell, DR., Weiser, JN., Shen, J., Young, A. and Tuomanen, EI. 1995. Relationship between colonial morphology and adherence of *Streptococcus pneumoniae*. Infect. Immun. 63:757-761.

Dagan, R., Engelhard, D., Piccard, E., and the Israeli Paediatric Bacteraemia and Meningitis Group. 1992. Epidemiology of invasive childhood pneumococcal infections in Israel. JAMA. 268:3328-3332.

Dagan, R., Melamed, R., Muallem, M., Piglansky, L., Greenberg, D., Abramson, O., Mendelman, PM., Bohidar, N. and Yagupsky, P. 1996. Reduction of nasopharyngeal carriage of pneumococci during the second year of life by a heptavalent conjugate pneumococcal vaccine. J. Infect. Dis. 174:1271-1278.

Dalrymple, B., Caspers, P. and Arber, W. 1984. Nucleotide sequence of the prokaryotic mobile genetic element IS*30*. EMBO J. 3:2145-2149.

Daniels, C., Vindurampulle, C. and Morona, R. 1998. Overexpression and topology of the *Shigella flexneri* O-antigen polymerase (Rfc/Wzy). Mol. Microbiol. 28:1211-1222.

Daniels, DL., Plunkett III, G. Burland, VD. and Blattner, FR. 1992. Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. Science 257:771-778.

Dawson, MH. and Sia, RHP. 1931. In vitro transformation of pneumococcal types: 1. A technique for inducing transformation of pneumococcal types in vitro. J. Exp. Med. 54:681-699.

DeAngelis, PL., Papaconstantinou, J. and Weigel, PH. 1993. Molecular cloning, identification, and sequence of the hyaluronan synthase gene from Group A *Streptococcus pyogenes*. J. Biol. Chem. 268:19181-19184.

DeAngelis, PL. and Weigel, PH. 1994. Immunochemical confirmation of the primary structure of streptococcal hyaluronan synthase and synthesis of high molecular weight product by the recombinant enzyme. Biochem. 33:9033-9039.

Dillard, JP., Vandersea, MW. and Yother, J. 1995. Characterisation of the cassette containing genes for type 3 capsular polysaccharide biosynthesis in *Streptococcus pneumoniae*. J. Exp. Med. 181:973-983.

Dintilhac, A., Alloing, G., Granadel, C. and Claverys, JP. 1997. Competence and virulence of *Streptococcus pneumoniae*: Adc and PsaA mutants exhibit a requirement for Zn and Mn resulting from inactivation of putative ABC metal permeases. Mol. Microbiol. 25:727-739.

Distler, J., Mansouri, K., Mayer, G., Stockman, M. and Piepersberg, W. 1992. Streptomycin biosynthesis and its regulation in Streptomyces. Gene. 115:105-111.

Dochez, AR. and Avery, OT. 1917. Soluble substance of pneumococcus origin in the blood and urine during lobar pneumonia. Proc. Soc. Exp. Biol. Med. 14:126-127.

Dolph, PJ., Majerczak, DR. and Coplin, DL. 1988. Characterisation of a gene cluster for exopolysaccharide biosynthesis and virulence in *Erwinia stewartii*. J. Bacteriol. 170:865-871.

Dong, Q., Sadouk, A., van der Lelie, D., Taghavi, S., Ferhat, A., Nuyten, JM., Borremans, B., Mergeay, M. and Toussaint, A. 1992. Cloning and sequencing of IS*1086*, an *Alcaligenes eutrophus* insertion element related to IS*30* and IS*4351*. J. Bacteriol. 174:8133-8138.

Douglas, RM., Paton, JC., Duncan. SJ. and Hansman, DJ. 1983. Antibody response to pneumococcal vaccination in children younger than five years of age. J. Infect. Dis. 148:131-137.

Drake, CR., Roberts, IS., Jann, B., Jann, K. and Boulnois, G. 1990. Molecular cloning and expression of the genes encoding the *Escherichia coli* K4 capsular polysaccharide, a fructose-substituted chondroitin. FEMS Microbiol. Lett. 66:227-30.

Eriksen, KR. 1945. Studies on induced resistance to penicillin in a pneumococcus type 1. Acta Pathol. Microbiol. Scand. 22:398-405.

Fallarino, A., Mavrangelos, C., Stroeher, UH. and Manning, PA. 1997. Identification of additional genes required for O-antigen biosynthesis in *Vibrio cholerae* O1. J. Bacteriol. 179:2147-2153.

Fang, GD., Fine, M., Orloff, J., Arisumi, D., Yu, VL., Kapoor, W., Grayston, JT., Wang, SP., Kohler, R., Muder, RR., Yee, YC., Rihs, JD. and Vickers RM. 1990. New and emerging aetiologies for community-acquired pneumonia with implications for therapy: A prospective multicenter study of 359 cases. Medicine. 69:307-316.

Ferrante, A., Rowan-Kelly, B. and Paton, JC. 1984. Inhibition of *in vitro* human lymphocyte response by the pneumococcal toxin pneumolysin. Infect. Immun. 46:585-589.

Finland, M. and Barnes, MW. 1977. Changes in occurrence of capsular serotypes of *Streptococcus pneumoniae* at Boston City Hospital during selected years between 1935 and 1974. J. Clin. Microbiol. 5:154-166.

Fischer, H. and Tomasz, A. 1985. Peptidoglycan cross-linking and teichoic acid attachment in *Streptococcus pneumoniae*. J. Bacteriol. 163:46-54.

Frisch, AW., Price, AE. and Meyers, GB. 1943. Type VIII pneumococcus: Development of sulphadiazine resistance transmission by cross infection and persistence in carriers. Ann. Intern. Med. 18:271-278.

Frosch, M. and Müller, A. 1993. Phospholipid substitution of capsular polysaccharides and mechanisms of capsule formation in *Neisseria meningitidis*. Mol. Microbiol. 8:483-493.

Frosch, M., Weisgerber, C. and Meyer, T. 1989. Molecular characterisation and expression in *Escherichia coli* of the gene complex encoding the polysaccharide capsule of *Neisseria meningitidis* group B. Proc. Natl. Acad. Sci. USA. 86:1669-1673.

Fuchs-Cleveland, E. and Gilvarg, CC. 1976. Oligomeric intermediates in peptidoglycan biosynthesis in *Bacillus megaterium*. Proc. Natl. Acad. Sci. USA. 73:4200-4204.

Gallo, MA., Ward, J. and Hutchinson, CR. 1996. The *dnrM* gene in *Streptomyces peucetius* contains a naturally occurring frameshift mutation that is suppressed by another locus outside of the daunorubicin-production gene cluster. Microbiol. 142:269-275.

García, E., Arrecubieta, C. and López, R. 1996. The *lytA* gene and the region located downstream of this gene are not involved in the formation of the type 3 capsular polysaccharide of *Streptococcus pneumoniae*. Curr. Microbiol. 33:133-135.

García, P., García, JL., García, E. and López, R. 1986. Nucleotide sequence and expression of the pneumococcal autolysin gene from its own promoter in *Escherichia coli*. Gene. 43:265-272.

García, E. and López, R. 1997. Molecular biology of the capsular genes of *Streptococcus pneumoniae*. FEMS Microbiol. Lett. 149:1-10.

Garcia-Bustos, JF., Chait, BT. and Tomasz, A. 1987. Structure of the peptide network of pneumococcal peptidoglycan. J. Biol. Chem. 262:15400-15405.

Garcia-Bustos, JF. Tomasz, A. 1990. A biological price of antibiotic resistance: Major changes in the peptidoglycan structure of penicillin-resistant pneumococci. Proc. Natl. Acad. Sci. USA. 87:5414-5419.

Giffard, PM., Rathsam, C., Kwan, E., Kwan, DW., Bunny, KL., Koo, SP. and Jaques, NA. 1993. The *ftf* gene encoding the cell-bound fructosyl transferase of *Streptococcus salivarius* ATCC 25975 is preceded by an insertion sequence and followed by FUR1 and *clpP* homologues. J. Gen. Microbiol. 139:913-920.

Giudicelli, S. and Tomasz, A. 1984. Attachment of pneumococcal autolysin to wall teichoic acids, an essential step in enzymatic wall degradation. J. Bacteriol. 158:1188-1190.

Glucksmann, MA., Reuber, TL. and Walker, GC. 1993. Genes needed for the modification, polymerisation, export, and processing of succinoglycan by *Rhizobium meliloti*: a model for succinoglycan biosynthesis. J. Bacteriol. 175:7045-7055.

Gram, C. 1884. Über die isolierte färbung der schizomyceten in schnitt- und trockenpräparaten. Fortschr. Med. 2:185-189.

Gratten, M., Barker, J., Shann, F., Gerega, G., Montgomery, J., Kajoi, M. and Lupiwa, T. 1985. The aetiology of purulent meningitis in highland children: a bacteriological study. PNG. Med. J. 28:233-240.

Gratten, M. and Montgomery J. 1991. The bacteriology of acute pneumonia and meningitis in children in Papua New Guinea: assumptions, facts and technical strategies. PNG. Med. J. 34:185-198.

Gratten, M., Naraqi, S. and Hansman, D. 1980. High prevalence of penicillin-insensitive pneumococci in Port Morseby, Papua New Guinea. Lancet. 2:192-195.

Gratten, M., Torzillo, P., Morey, F., Dixon, J., Erlich, J., Harrer, J. and Henrichsen, J. 1996. Distribution of capsular types and antibiotic susceptibility of invasive *Streptococcus pneumoniae* isolated from Aborigines in Central Australia. J. Clin. Microbiol. 34:338-341.

Gray, BM., Converse III, GM. and Dillon Jr., HC. 1979. Serotypes of *Streptococcus pneumoniae* causing disease. J. Infect. Dis. 140:979-983.

Gray, BM., Converse III, GM. and Dillon Jr., HC. 1980. Epidemiologic studies of *Streptococcus pneumoniae* in infants: Acquisition, carriage, and infection during the first 24 months if life. J. Infect. Dis. 142:923-933.

Grebe, T., Paik, J. and Hakenbeck, R. 1997. A novel resistance mechanism against beta-lactams in *Streptococcus pneumoniae* involves CpoA, a putative glycosyl transferase. J. Bacteriol. 179:3342-3349.

Griffin, AM., Morris, VJ. and Gasson, MJ. 1996. The *cpsABCDE* genes involved in polysaccharide production in *Streptococcus salivarius* ssp. thermophilus strain NCBF 2393. Gene. 183:23-27.

Griffith, F. 1928. The significance of pneumococcal types. J. Hyg. 27:113-159.

Gross, M., Geier, G., Rudolph, K. and Geider, K. 1992. Levan and levansucrase synthesised by the fireblight pathogen *Erwinia amylovora*. Physiol. Mol. Plant Pathol. 40:371-381.

Guidolin, A., Morona, JK., Morona, R., Hansman, D. and Paton, JC. 1994. Nucleotide sequence analysis of genes essential for capsular polysaccharide biosynthesis in *Streptococcus pneumoniae* type 19F. Infect. Immun. 62:5384-5396.

Guirguis, NI., Helmy, MF., Mohamed, MR. and Ali, RH. 1990. A suggested vaccine formulation for the control of pneumococcal meningitis in Egypt. J. Egypt Public Health Assoc. 65:291-303.

Hammerschmidt, S., Birkholz, C., Zahringer, U., Robertson, BD., van Putten, J., Ebeling, O. and Frosch, M. 1994. Contribution of genes from the capsule gene complex (*cps*) to lipooligosaccharide biosynthesis and serum resistance in *Neisseria meningitidis*. Mol. Microbiol. 11:885-896.

Hammerschmidt, S., Taley, SR., Brandtzaeg, P. and Chatwal, S. 1997. SpsA, a novel pneumococcal surface protein with specific binding to secretory Immunoglobulin A and secretory component. Mol. Microbiol. 25:1113-1124.

Hammond, SM., Lambert, PA and Rycroft, AN. 1984. The bacterial cell surface. Croon Helm, London.

Hanahan, D. 1983. Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. 166:557-580.

Hansman, D. 1983. Serotypes in pneumococcal disease: a ten year study in Australia 1970 through 1979. Aust. NZ. J. Med. 13:359-64.

Hansman, D. and Bullen, MM. 1967. A resistant pneumococcus [letter]. Lancet. 2:264-265.

Hansman, D., Devitt, L., Miles, H. and Riley, I. 1974. Pneumococci relatively insensitive to penicillin in Australia and New Guinea. Med. J. Aust. 2:353-356.

Hashimoto, Y., Li, N., Yokoyama, H. and Ezaki, T. 1993. Complete nucleotide sequencing and characterisation of ViaB region encoding Vi antigen in *Salmonella typhi*. J. Bacteriol. 175:4456-4465.

Heidelberger, M. 1927. Immunologically specific polysaccharides. Chem. Rev. 3:403-423.

Heidelberger, M., and Avery, OT. 1923. The soluble specific substance of pneumococcus. J. Exp. Med. 38:73-79.

Heinrichs, DE., Monteiro, MA., Perry, MB. and Whitfield, C. 1998. The assembly system for the lipopolysaccharide R2 core-type of *Escherichia coli* is a hybrid of those found in *Escherichia coli* K-12 and *Salmonella enterica*. Structure and function of the R2 WaaK and WaaL homologues. J. Biol. Chem. 23:8849-8859.

Henikoff, S. 1984. Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene. 28:351-359.

Henrichsen, J. 1995. Six newly recognised types of *Streptococcus pneumoniae.* J. Clin. Microbiol. 33:2759-2762.

Higgins, DG. and Sharp, PM. 1988. CLUSTAL: a package for performing multiple sequence alignments on a microcomputer. Gene. 73:237-244.

Hilleman, MR., Carlson Jr., AM., McLean, AA., Vella, PP., Weibel RE. and Woodhour, AF. 1981. *Streptococcus pneumoniae* polysaccharide vaccine: age, and dose responses, safety, persistence of antibody, revaccination, and simultaneous administration of pneumococcal and influenza vaccines. Rev. Infect. Dis. 3(Suppl):S31-S42.

Hobbs, M. and Reeves, PR. 1994. The JUMPstart sequence: a 39 bp element common to several polysaccharide gene clusters. Mol. Microbiol. 12:855-856.

Hofmann, K. and Stöffel, W. 1989. PROFILEGRAPH: an interactive graphical tool for protein sequence analysis. Comput. Appl. Biosci. 5:151-153.

Hook, EW., Horton, CA. and Scharberg, DR. 1983. Failure of intensive care unit support to influence mortality from pneumococcal bacteraemia. JAMA. 249:1055-1057.

Hotchkiss, RD. 1951. Transfer of penicillin resistance in pneumococci by the desoxynucleate derived from resistant cultures. Cold Spring Harb. Symp. Quant. Biol. 16:457-461.

Huang, J. and Schell, M. 1995. Molecular characterisation of the *eps* gene cluster of *Pseudomonas solanacearum* and its transcriptional regulation at a single promoter. Mol. Microbiol. 16:977-989.

Iannelli, F., Pearce, BJ. and Pozzi, G. 1998. The type 2 capsule locus of *Streptococcus pneumoniae*. J. Bacteriol. (in press).

Ielpi, L., Couso, RO. and Dankert, MA. 1993. Sequential assembly and polymerisation of the polypreol-linked pentasaccharide repeating unit of the xanthan polysaccharide in *Xanthomonas campestris*. J. Bacteriol. 175:2490-2500.

Issaeff, B. 1893. Contribution à l'étude de l'immunité aquise contre le pneumocoque. Ann. de l'Institut Pasteur. 7:260-279.

Jacobs, MR., Koornhof, HJ., Robins-Browne, RM. Stevenson, CM., Vermaak, ZA., Freiman, I., Miller, GB., Witcomb, MA., Isaacson, M., Ward, JI. and Austrian, R. 1978. Emergence of multiply resistant pneumococci. New Engl. J. Med. 299:734-740.

Jacobs, NM., Lerdkachornsuk, S. and Metzger, W. 1979. 1. Pneumococcal bacteraemia in infants and children: a ten-year experience at the Cook County Hospital with special reference to the pneumococcal serotypes isolated. Pediatrics. 64:296-300.

Jansson, PE., Lindberg, B., Andersson, M., Lindquist, U. and Henrichsen, J. 1988. Structural studies of the capsular polysaccharide from *Streptococcus pneumoniae* type 2, a reinvestigation. Carbohyd. Res. 182:111-117.

Jansson, PE., Lindberg, B. and Lindquist, U. 1985. Structural studies of the capsular polysaccharide from *Streptococcus pneumoniae* type 5. Carbohyd. Res. 140:101-110.

Jennings, HJ., Lugowski, C. and Young, NM. 1980. Structure of the complex polysaccharide C-substance from *Streptococcus pneumoniae* type 1. Biochemistry. 19:4712-4719.

Jiang, X-M., Neal, B., Santiago, F., Lee, SJ., Romana, LK. and Reeves, PR. 1991. Structure and sequence of the *rfb* (O antigen) gene cluster of *Salmonella* serovar typhimurium (strain LT2). Mol. Microbiol. 5:695-713.

Jobling, MG. and Holmes, RK. 1990. Construction of vectors with the p15a replicon, kanamycin resistance, inducible *lacZ*α and pUC18 or pUC19 multiple cloning sites. Nucleic Acids Res. 18:5315-5316.

Jones, C. and Currie, F. 1988. The pneumococcal polysaccharide S4: A structural re-assessment. Carbohyd. Res. 184:279-284.

Jones, C., Mulloy, B., Wilson, A., Dell, A. and Oates, JE. 1985. Structure of the capsular polysaccharide from *Streptococcus pneumoniae* type 9. J. Chem. Soc. Perkin I:1665-1673.

Jones, JNK. and Perry, MB. 1957. The structure of the type VIII pneumococcus specific polysaccharide. J. Amer. Chem. Soc. 79:2787-2793.

Kaneko, T., Tanaka, A., Sato, S., Kotani, H., Sazuka, T., Miyajima, N., Sugiura, M. and Tabata, S. 1995. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to 92% of the genome. DNA Res. 2:153-166.

Kapur, V., Reda, KB., Li, LL., Rich, RR. and Musser JM. 1994. Characterisation and distribution of insertion sequence IS*1239* in *Streptococcus pyogenes*. Gene 150:135-140.

Katzenellenbogen, E. and Jennings. HJ. 1983. Structural determination of the capsular polysaccharide of *Streptococcus pneumoniae* type 19A (57). Carbohyd. Res. 124:235-245.

Keenleyside, WJ. and Whitfield, C. 1995. Lateral transfer of *rfb* genes: a mobilizable ColE1-type plasmid carries the *rfb*O:54 (O:54 antigen biosynthesis) gene cluster from *Salmonella enterica* serovar Borreze. J. Bacteriol. 177:5247-5253.

Keenleyside, WJ. and Whitfield, C. 1996. A novel pathway for O-polysaccharide biosynthesis in *Salmonella enterica* serovar Borreze. J. Biol. Chem. 271:28581-28592.

Kelly, T., Dillard, JP. and Yother, J. 1994. Effect of genetic switching of capsular type on virulence of *Streptococcus pneumoniae.* Infect. Immun. 62:1813-1819.

Kenne, L., Lindberg, B. and Madden, JK. 1979. Structural studies of the capsular antigen from *Streptococcus pneumoniae* type 26. Carbohyd. Res. 73:175-182.

Kessler, AC., Haase, A. and Reeves, PR. 1993. Molecular analysis of the 3,6-dideoxyhexose pathway genes of *Yersinia pseudotuberculosis* serogroup IIA. J. Bacteriol. 175:1412-1422.

Kiino, DR., Licudine, R., Wilt, K., Yang, DHC. and Rothman-Denes, LB. 1993. A cytoplasmic protein, NfrC, is required for bacteriophage N4 adsorption. J. Bacteriol. 175:7074-7080.

Kim JO. and Weiser, JN. 1998. Association of intrastrain phase variation in quantity of capsular polysaccharide and teichoic acid with the virulence of *Streptococcus pneumoniae*. J. Infect. Dis. 177:368-377.

Klee, SR., Tzschaschel, BD., Timmis, KN. and Guzman, CA. 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.

Klemperer, G and Klemperer, F. 1891. Versuche über immunisirung und heilung bei der pneumokokkeninfection. Berliner Klinische Wochenschrift. 28:833-835, 869-875.

Klena, JD. and Schnaitman, CA. 1993. Function of the *rfb* gene cluster and the *rfe* gene in the synthesis of O-antigen by *Shigella dysenteriae* 1. Mol. Microbiol. 9:393-402.

Klugman, KP. and Koornhof, HJ. 1988. Drug resistance patterns and serogroups of serotypes of pneumococcal isolates from cerebrospinal fluid of blood, 1979-1986. J. Infect. Dis. 158:956-964.

Klugman, KP. 1996. Serotypes and clones of antibiotic-resistant pneumococci. Abstract from *Streptococcus pneumoniae*: Molecular biology and mechanisms of disease. Oeiras, Portugal, September, 1996.

Kolkman, MAB., Morrison, DA., van der Zeijst, BAM. and Nuijten, PJM. 1996. The capsule polysaccharide synthesis locus of *Streptococcus pneumoniae* serotype 14: identification of the glycosyl transferase gene *cps14E*. J. Bacteriol. 178:3736-3741.

Kolkman, MAB., van der Zeijst, BAM. and Nuijten, PJM. 1997a. Functional analysis of glycosyl transferases encoded by the capsular polysaccharide biosynthesis locus of *Streptococcus pneumoniae* serotype 14. J. Biol. Chem. 272:19502-19508.

Kolkman, MAB., van der Zeijst, BAM. and Nuijten, PJM. 1998. Diversity of capsular polysaccharide synthesis gene clusters in *Streptococcus pneumoniae*. J. Biochem. (Tokyo) 123:937-945.

Kolkman, MAB., Wakarchuk, W., Nuijten, PJM. and van der Zeijst, BAM. 1997b. Capsular polysaccharide synthesis in *Streptococcus pneumoniae* serotype 14: molecular analysis of the complete *cps* locus and identification of genes encoding glycosyl transferases required for the biosynthesis of the tetrasaccharide subunit. Mol. Microbiol. 26:197-208.

Koplin, R., Wang, G., Hotte, B., Priefer, UB. and Puhler, A. 1993. A 3.9-kb DNA region of *Xanthomonas campestris* pv. campestris that is necessary for lipopolysaccharide production encodes a set of enzymes involved in the synthesis of dTDP-rhamnose. J. Bacteriol. 175:7786-7792.

Koskiniemi, S., Sellin, M. and Norgren, M. 1998. Identification of two genes, *cpsX* and *cpsY*, with putative regulatory function on capsule expression in group B streptococci. FEMS Immun. Med. Microbiol. 21:159-168.

Kroll, JS., Loynds, BM. and Moxon, ER. 1991. The *Haemophilus influenzae* capsulation gene cluster: a compound transposon. Mol. Microbiol. 5:1549-1560.

Kroll, S., Zamze, S., Loynds, B. and Moxon, E. 1989. Common organisation of chromosomal loci for production of different capsular polysaccharides in *Haemophilus influenzae*. J. Bacteriol. 171:3343-3347.

Kumar, S., Tamura, K. and Nei, M. 1994. MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. Comput. Appl. Biosci. 10:189-191.

Kyte, J. and Doolittle, RF. 1982. A simple method for displaying the hydrophobic character of a protein. J. Mol. Biol. 157:105-132.

Laemmli, UK. 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature. 227:680-685.

Lazarevic, V. and Karamata, D. 1995. The *tagGH* operon of Bacillus subtilis 168 encodes a two-component ABC transporter involved in the metabolism of two wall teichoic acids. Mol. Microbiol. 16:345-355.

Lazarevic, V., Margot, P., Soldo, B. and Karamata, D. 1992. Sequencing and analysis of the *Bacillus subtilis lytRABC* divergon: a regulatory unit encompassing the structural genes of the *N*-acetylmuramoyl-L-alanine amidase and its modifier. J. Gen. Microbiol. 138: 1949-1961.

Leach, A., Ceesay, SJ., Banya, WAS. and Greenwood, BM. 1996. Pilot trial of a pentavalent pneumococcal polysaccharide/protein conjugate vaccine in Gambian infants. Pediatr. Infect. Dis. J. 15:333-339.

Lee, C-J. 1987. Bacterial capsular polysaccharide biochemistry, immunity and vaccine. Mol. Immun. 24:1005-1019.

Lee, C-J. and Fraser, BA. 1980. The structures of the cross-reactive types 19 (19F) and 57 (19A) pneumococcal polysaccharides. J. Biol. Chem. 255:6847-6853.

Lee, C-J., Fraser, BA., Boykins, RA. and Li JP. 1987. Effect of culture conditions on the structure of *Streptococcus pneumoniae* type 19A (57) capsular polysaccharide. Infect. Immun. 55:1819-1823.

Lee, SJ., Romana, LK. and Reeves, PR. 1992. Cloning and structure of group C1 O antigen (*rfb* gene cluster) from *Salmonella enterica* serovar montevideo. J. Gen. Microbiol. 138:305-312.

Lin, WS., Cunneen, T. and Lee, CY. 1994. Sequence analysis and molecular characterisation of genes required for the biosynthesis of type 1 capsular polysaccharide in *Staphylococcus aureus*. J. Bacteriol. 176:7005-7016.

Lindberg, B., Lönngren, J. and Powell, DA. 1977. Structural studies on the specific type 14 pneumococcal polysaccharide. Carbohyd. Res. 58:177-186.

Lindberg, B., Lindqvist, B., Lönngren, J. and Powell, DA. 1980. Structural studies of the capsular polysaccharide from *Streptococcus pneumoniae* type 1. Carbohyd. Res. 78:111-117.

Linton, KJ., Jarvis, BW. and Hutchinson, CR. 1995. Cloning of the genes encoding thymidine diphosphoglucose 4,6-dehydratase and thymidine diphospho-4-keto-6-deoxyglucose 3,5-epimerase from the erythromycin-producing *Saccharopolyspora erythraea*. Gene. 153:33-40.

Liu, D., Cole, RA. and Reeves, PR. 1996. An O-antigen processing function for Wzx (RfbX): a promising candidate for O-unit flippase. J. Bacteriol. 178:2102-2107.

Llull, D., López, R., García, E. and Muñoz, R. 1998. Molecular structure of the gene cluster responsible for the synthesis of the polysaccharide capsule of *Streptococcus pneumoniae* type 33F. Biochim. Biophys. Acta. (in press).

Lock, RA., Paton, JC. and Hansman, D. 1988a. Purification and immunological characterisation of neuraminidase produced by *Streptococcus pneumoniae*. Microb. Pathogen. 4:33-43.

Lock, RA., Paton, JC. and Hansman, D. 1988b. Comparative efficacy of pneumococcal neuraminidase and pneumolysin as immunogens protective against *Streptococcus pneumoniae* infection. Microb. Pathogen. 5:461-467.

Lu, C-D. and Abdelal, AT. 1993. The *Salmonella typhimurium* uracil-sensitive mutation *use* is in *argU* and encodes a minor arginine tRNA. J. Bacteriol. 175:3897-3899.

Lugowski, C. and Jennings, HJ. 1984. Structural determination of the capsular polysaccharide of *Streptococcus pneumoniae* type 18C (56). Carbohyd. Res. 131:119-129.

MacLeod, CM. and Krauss, MR. 1950. Relation of virulence of pneumococcal strains to the quantity of capsular polysaccharide formed *in vitro*. J. Exp. Med. 92:1-9.

McCarty, M. and Avery, OT. 1946. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: II. Effect of desoxyribonuclease on the biological activity of the transforming substance. J., Exp. Med. 83:89-96.

McDaniel, LS., Scott, G., Kearney, JF. and Briles, DE. 1984. Monoclonal antibodies against protease-sensitive pneumococcal antigens can protect mice from fatal infection with *Streptococcus pneumoniae*. J. Exp. Med. 160:386-397.

McDaniel, LS., Yother, J., Vijayakamur, M., McGarry, L., Guild, WR. and Briles, DE. 1987. Use of insertional inactivation to facilitate studies of biological properties of pneumococcal surface protein A (PspA). J. Exp. Med. 165:381-394.

McIntyre, P. 1997. Epidemiology and prevention of pneumococcal disease. Comm. Dis. Intell. 21:41-46.

Macpherson, DF., Manning, PA. and Morona, R. 1994. Characterisation of the dTDP-rhamnose biosynthetic genes encoded in the *rfb* locus of *Shigella flexneri*. Mol. Microbiol. 11:281-292.

Macpherson, DF., Manning, PA. and Morona, R. 1995. Genetic analysis of the *rfbX* gene of *Shigella flexneri*. Gene. 155:9-17.

Macrina, FL., Evans, RP., Tobian, JA., Hartley, DL., Clewell, DB. and Jones, KR. 1983. Novel shuttle plasmid vehicles for *Escherichia-Streptococcus* transgeneric cloning. Gene. 25:145-150.

Maharaj, R., May, TB., Wang, SK. and Chakrabarty, AM. 1993. Sequence of the *alg8* and *alg44* genes involved in the synthesis of alginate by *Pseudomonas aeruginosa*. Gene. 136:267-269.

Maniatis, T., Fritsch, EF. and Sambrook, J. 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Marolda, CL. and Valvano, MA. 1995. Genetic analysis of the dTDP-rhamnose biosynthesis region of the *Escherichia coli* VW187 (O7:K1) *rfb* gene cluster: identification of functional homologues of *rfbB* and *rfbA* in the *rff* cluster and correct location of the *rffE* gene. J. Bacteriol. 177:5539-5546.

Marolda, CL. and Valvano, MA. 1998. Promoter region of the *Escherichia coli* O7-specific lipopolysaccharide gene cluster: structural and functional characterisation of an upstream untranslated mRNA sequence. J. Bacteriol. 180:3070-3079.

Mastro, TD., Ghafoor, A., Nomani, NK., Ishaq, Z., Anwar, F., Granoff, DM., Spika, JS., Thornsberry, C. and Facklam RR. 1991. Antimicrobial resistance of pneumococci in children with acute lower respiratory tract infection in Pakistan. Lancet. 1:156-159

Mauel, C., Young, M. and Karamata, D. 1991. Genes concerned with synthesis of poly(glycerol phosphate), the essential teichoic acid in *Bacillus subtilis* strain 168, are organised in two divergent transcriptional units. J. Gen. Microbiol. 137:929-941.

May, TB. and Chakrabarty, AM. 1994. *Pseudomonas aeruginosa*: genes and enzymes of alginate synthesis. Trends Microbiol. 2:151-157.

Meier-Dieter, U., Starman, R., Barr, K., Mayer, H. and Rick, PD. 1990. Biosynthesis of enterobacterial common antigen in *Escherichia coli*. Biochemical characterisation of Tn10 insertion mutants defective in enterobacterial common antigen synthesis. J. Biol. Chem. 265:13490-13497.

Merson-Davies, LA. and Cundliffe, E. 1994. Analysis of five tylosine biosynthetic genes from the *tylIBA* region of the *Streptomyces fradiae* genome. Mol. Microbiol. 13:349-355.

Mills, GT. and Smith, EEB. 1965. Biosynthesis of capsular polysaccharides in the pneumococcus. Bull. Soc. Chim. Biol. 47:1751-1765.

Mitchell, TJ., Andrew, PW., Saunders, FK., Smith, AN. and Boulnois, GJ. 1991. Complement activation and antibody binding by pneumolysin via a region of the toxin homologous to a human acute phase protein. Mol. Microbiol. 5:1883-1888.

Mitchison, MD. Bulach, M., Vinh, T., Rajakumar, K., Faine, S. and Adler, B. 1997. Identification and characterisation of the dTDP-rhamnose biosynthesis and transfer genes of the lipopolysaccharide-related *rfb* locus in *Leptospira interrogans* serovar Copenhageni. J. Bacteriol. 179:1262-1267.

Mogdasy, MC., Camou, T., Fajardo, C. and Hortal, M. 1992. Colonising and invasive strains of *Streptococcus pneumoniae* in Uruguayan children: Type distribution and patterns of antibiotic resistance. Pediatr. Infect. Dis. J. 11:648-652.

Mollerach, M., López, R. and García, E. 1998. Characterisation of the *galU* gene of *Streptococcus pneumoniae* encoding a UDP-glucose pyrophosphorylase: a gene essential for capsular polysaccharide biosynthesis. J. Exp. Med. (in press).

Montgomery, JM., Lehman, D., Smith, T., Michael, A., Joseph, B., Lupiwa, T., Coakley, C., Spooner, V., Best, B., Riley, ID. and Alpers, MP. 1990. Bacterial colonisation of the upper respiratory tract and its association with acute lower respiratory tract infections in Highland children of Papua New Guinea. Rev. Infect. Dis. 12:S1006-S1016.

Moore, HF. and Chesney, AM. 1917. A study of ethylhydrocuprein (optichin) in the treatment of acute lobar pneumonia. Ach. Intern. Med. 19:611-682.

Mooser, G., Hefta, SA., Paxton, RJ., Shively, JE. and Lee, TD. 1991. Isolation and sequence of an active-site peptide containing a catalytic aspartic acid from two *Streptococcus sorbinus* alpha-glucosyl transferases. J. Biol. Chem. 266:8916-8922.

Mooser, G. and Wong, C. 1988. Isolation of a glucan binding domain of glucosyl transferase (1,6-α-glucan synthase) from *Streptococcus sorbinus.* Infect. Immun. 56:880-884.

Moreau, M., Richards, JC., Perry, MB. and Kniskern, PJ. 1988. Application of high resolution n. m. r. spectroscopy to the elucidation of the structure of the specific capsular polysaccharide of *Streptococcus pneumoniae* type 7F. Carbohyd. Res. 182:79-99.

Morelle, G. 1989. A plasmid extraction procedure on a miniprep scale. Focus. 11.1:7-8.

Morgenroth, J. and Kaufman, M. 1912. Arzneifestigkeit bei bakterien (pneumokokken). Zeitschrift für Imunitätsforschung und Experimentelle Therapie. 15:610-624.

Morona, JK., Guidolin, A., Morona, R., Hansman, D. and Paton, JC. 1994a. Isolation, Characterisation, and Nucleotide Sequence of IS*1202*, an Insertion Sequence of *Streptococcus pneumoniae.* J. Bacteriol. 176:4437-4443.

Morona, JK., Miller, DC., Coffey, TJ., Vindurampulle, CJ., Spratt, BG., Morona, R. and Paton, JC. 1998. Molecular and genetic characterisation of the capsule biosynthesis locus of *Streptococcus pneumoniae* type 23F. Microbiol. (Manuscript submitted).

Morona, R., Macpherson, DF., Van den Bosch, L., Carlin, NIA. and Manning, PA. 1995. Lipopolysaccharide with an altered O-antigen produced in *Escherichia coli* K-12 harbouring mutated, cloned *Shigella flexneri rfb* genes. Mol. Microbiol. 18:209-223.

Morona, R., Mavris, M., Fallarino, A. and Manning, PA. 1994b. Characterisation of the *rfc* region of *Shigella flexneri.* J. Bacteriol. 176: 733-747.

Mufson, MA., Krause, HE., Schiffman, G. and Hughey, DF. 1987. Pneumococcal antibody levels one decade after immunisation of healthy adults. Amer. J. Med. Sci. 293:279-284.

Mukasa, H. 1986. Properties of *Streptococcus mutans* glucosyl transferases, p 121-132. In Molecular microbiology and immunobiology of *Streptococcus mutans*. Hanada, S., Michalek, SM., Kiyono, H., Menaker, L. and McGhee, JR. Elsevier Science Publishers, New York.

Muñoz, R., Mollerach, M., López, R. and García, E. 1997. Molecular organisation 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.

Musher, DM. 1992. Infections caused by *Streptococcus pneumoniae*: Clinical spectrum, pathogenesis, Immunity, and treatment. Clin. Infect. Rev. 14:801-809.

Musher, DM., Luchi, MJ., Watson, DA., Hamilton, R. and Baughn, RE. 1990. Pneumococcal polysaccharide vaccine in young adults and older bronchitics: determination of IgG responses by ELISA and the effect of adsorption of serum with non-type-specific cell wall polysaccharide. J. Infect. Dis. 161:728-735.

Nesin M., Ramirez, M. and Tomasz, A. 1998. Capsular transformation of a multidrug-resistant *Streptococcus pneumoniae in vivo*. J. Infect. Dis. 177:707-713.

Neufeld, F. 1902. Über die agglutination der pneumokokken und über die theorien der agglutination. Zeitschrift für Hygiene und Infektionskrankheiten (Leipzig). 40:54-72.

Neufeld, F. and Rimpau, W. 1904. Über die antikorper des streptokokken- und pneumokokken-immunserums. Deutsche Med. Wochenschrift. 30:1458-1460.

Neuhard, J. and Thomassen, E. 1976. Altered deoxyribonucleotide pools in P2 educants of *Escherichia coli* K-12 due to deletion of the *dcd* gene. J. Bacteriol. 126:999-1001.

Nielsen, SV. and Henrichsen, J. 1992. Capsular types of *Streptococcus pneumoniae* isolated from blood and CSF during 1982-1987. Clin. Infect. Dis. 15:794-798.

Obaro, SK., Adegbola, RA., Banya, WAS. and Greenwood, BM. 1996. Carriage of pneumococci after pneumococcal vaccination. Lancet. 348:271-272.

Ochman, H., Gerber, AS. and Hartl, DL. 1988. Genetic applications of an inverse polymerase chain reaction. Genetics. 120:621-623.

Odani, T., Shimma, Y., Tanaka, A. and Jigami, Y. 1996. Cloning and analysis of the *MNN4* gene required for phosphorylation of *N*-linked oligosaccharides in *Saccharomyces cerevisiae*. Glycobiol. 6:805-810.

Ohno, N., Yadomae, T. and Miyazaki, T. 1980. The structure of the type specific polysaccharide of pneumococcus type XIX. Carbohyd. Res. 80:297-304.

Paton, JC. 1996. The contribution of pneumolysin to the pathogenicity of *Streptococcus pneumoniae*. Trends Microbiol. 4:103-106.

Paton, JC. 1998. Novel pneumococcal surface proteins: role in virulence and vaccine potential. Trends Microbiol. 6:85-87.

Paton, JC., Andrew, PW., Boulnois, GJ. and Mitchell, TJ. 1993. Molecular analysis of the pathogenicity of *Streptococcus pneumoniae*: The role of pneumococcal proteins. Ann.. Rev. Microbiol. 47:89-115.

Paton, JC. and Ferrante, A. 1983. Inhibition of human polymorphonuclear leucocyte respiratory burst, migration and bactericidal activity by the pneumococcal toxin, pneumolysin. Infect. Immun. 41:1212-1216.

Paton, JC., Lock, RA. and Hansman, D. 1983. Effect of immunisation with pneumolysin on survival time of mice challenged with *Streptococcus pneumoniae*. Infect. Immun. 40:584-552.

Paton, JC., Rowan-Kelly, B. and Ferrante, A. 1984. Activation of human complement by the pneumococcal toxin, pneumolysin. Infect. Immun. 43:1085-1087.

Paton, JC., Toogood, IR., Cockington, RA. and Hansman, D. 1986. Antibody response to pneumococcal vaccine in children aged 5 to 15 years. Amer. J. Dis. Child. 140:135-138.

Paulsen, IT., Beness, AM. and Saier, MH. Jr. 1997. Computer-based analyses of the protein constituents of transport systems catalysing export of complex carbohydrates in bacteria. Microbiol. 143:2685-2699.

Pearce, BJ., Naughton, AM. and Masure, HR. 1994. Peptide permeases modulate transformation in *Streptococcus pneumoniae*. Mol. Microbiol. 12:881-892.

Pearce, R. and Roberts, IS. 1995. Cloning and analysis of the gene clusters for the production of the *Escherichia coli* K10 and K54 antigens: identification of a new group of *serA*-linked capsule gene clusters. J. Bacteriol. 177:3992-3997.

Perry, MB., Daoust, V. and Carlo, DJ. 1981. The specific capsular polysaccharide of *Streptococcus pneumoniae* type 9V. Can. J. Biochem. 59:524-533.

Petit, C., Rigg, GP., Pazzani, C., Smith, A., Sieberth, V., Stevens, M., Boulnois, G., Jann, K. and Roberts, IS. 1995. Region 2 of the *Escherichia coli* K5 capsule gene cluster encoding proteins for the biosynthesis of the K5 polysaccharide. Mol. Microbiol. 17:611-620.

Phillips, NJ., Apicella, MA., Griffiss, JM. and Gibson, BW. 1993. Structural studies of the lipooligosaccharides from *Haemophilus influenzae* type b strain A2. Biochem. 32:2003-2012.

Pozzi, G., Masala, L., Iannelli, F., Manganelli, R., Håvarstein, LS., Piccoli, L., Simon, D. and Morrison, DA. 1996. Competence for genetic transformation in encapsulated strains of *Streptococcus pneumoniae*: two allelic variants of the peptide pheromone. J. Bacteriol. 178:6087-6090.

Presecan, E., Moszer, I., Boursier, L., Cruz Ramos, HC., de la Fuente, V., Hullo, MF., Lelong, C., Schleich, S., Sekowska, A., Song, BH., Villani, G., Kunst, F., Danchin, A. and Glaser, P. 1997. The Bacillus subtilis genome from *gerBC* (311 degrees) to *licR* (334 degrees). Microbiol. 143:3313-3328.

Ramirez, M. and Tomasz, A. 1998. Molecular characterisation of the complete 23F capsular polysaccharide locus of *Streptococcus pneumoniae*. J. Bacteriol. 180:5273-5278.

Ratcliffe, SW., Luh, J., Ganesan, AT., Behrens, B., Thompson, R., Montenegro, MA., Morelli, G. and Trautner, TA. 1979. The genome of Bacillus subtilis phage SPP1: the arrangement of restriction endonuclease generated fragments. Mol. Gen. Genet. 168:165-172.

Rebers, PA. and Heidelberger, M. 1961. The specific polysaccharide of type VI pneumococcus. 2. The repeating unit. J. Amer. Chem. Soc. 83:3056-3059.

Reeves, RE. and Goebel, WF. 1941. Chemoimmunological studies on the soluble specific substance of pneumococcus. V. The structure of the type III polysaccharide. J. Biol. Chem. 139:511-519.

Reizer, J., Reizer, A. and Saier, MH. Jr. 1992. A new sub-family of bacterial ABC-type transport systems catalysing export of drugs and carbohydrates. Prot. Sci. 1:1326-1332.

Reuber, TL. and Walker, GC. 1993. Biosynthesis of succinoglycan, a symbiotically important exopolysaccharide of *Rhizobium meliloti*. Cell 74:269-280.

Richards, JC. and Perry, MB. 1988. Structure of the specific capsular polysaccharide of *Streptococcus pneumoniae* type 23F (American type 23). Biochem. Cell Biol. 66:758-771.

Riley, ID. and Douglas, RM. 1981. An epidemiologic approach to pneumococcal disease. Rev. Infect. Dis. 3:233-245.

Robbins, JB., Austrian, R., Lee, C-J., Rastogi, SC., Schiffman, G., Henrichsen, J., Mäkelä, PH., Broome, CV., Facklam, RR., Tiesjema, RH. and Parke Jr., JC. 1983. Considerations for formulating the second-generation pneumococcal capsular polysaccharide vaccine with emphasis on the cross-reactive types within groups. J. Infect. Dis. 148:1136-1159.

Roberts, IS. 1996. The biochemistry and genetics of capsular polysaccharide production in bacteria. Ann. Rev. Microbiol. 50:285-315.

Roberts, IS., Mountford, R., High, N., Bitter-Suermann, D., Jann, K., Timmis, K. and Boulnois, G. 1986. Molecular cloning and analysis of the genes for the production of the K5, K7, K12 and K92 capsular polysaccharides in *Escherichia coli*. J. Bacteriol. 168:1228-1233.

Robertson, BD., Frosch, M. and van Putten, JP. 1994. The identification of cryptic rhamnose biosynthesis genes in *Neisseria gonorrhoeae* and their relationship to lipopolysaccharide biosynthesis. J. Bacteriol. 176:6915-6920.

Rosenow, C., Ryan, P., Weiser, JN., Johnson, S., Fontan, P., Ortqvist, A. and Masure, HR. 1997. Contribution of novel choline-binding proteins to adherence, colonisation and immunogenicity of *Streptococcus pneumoniae*. Mol. Microbiol. 25:819-829.

Rubens, CE., Heggen, LM., Haft, RF. and Wessels, MR. 1993. Identification of *cpsD*, a gene essential for type III capsule expression in group B streptococci. Mol. Microbiol. 8: 843-855.

Rubins, JB., Duane, PG., Charboneau, D. and Janoff, EN. 1992. Toxicity of pneumolysin to pulmonary endothelial cells in vitro. Infect. Immun. 60:1740-1746.

Rubins, JB., Duane, PG., Clawson, D., Charboneau, D., Young, J. and Niewoehner, DE. 1993. Toxicity of pneumolysin to pulmonary alveolar epithelial cells. Infect. Immun. 61:1352-1358.

Sampson, JS., O'Connor, SP., Stinson, AR., Tharpe, JA. and Russell, H. 1994. Cloning and nucleotide sequence of *psaA*, the *Streptococcus pneumoniae* gene encoding a 37-kilodalton protein homologous to previously reported *Streptococcus* sp. adhesins. Infect. Immun. 62:319-324.

Sanger, F., Coulson, AR., Hong, GF., Hill, DF. and Petersen, GB. 1983. Nucleotide sequence of bacteriophage lambda DNA. J. Mol. Biol. 162:729-773.

Sankilampi, U. 1997. *Streptococcus pneumoniae* and the elderly: an epidemiological and serological study focusing on the potential of the polysaccharide vaccine. PhD. Thesis. University of Oulu, Finland.

Sau, S. Bhasin, N., Wann, ER., Lee, JC., Foster, TJ. and Lee, CY. 1997. The *Staphylococcus aureus* allelic genetic loci for serotype 5 and 8 capsule expression contain the type-specific genes flanked by common genes. Microbiol. 143:2395-2405.

Sau, S. and Lee, CY. 1996. Cloning of type 8 capsule genes and analysis of gene clusters for the production of different capsular polysaccharides in *Staphylococcus aureus*. J. Bacteriol. 178:2118-2126.

Saxena, IM., Brown, RM. Jr., Fevre, M., Geremia, RA. and Henrissat, B. 1995. Multidomain architecture of beta-glycosyl transferases: implications for mechanism of action. J. Bacteriol. 177:1419-1424.

Saxena, IM., Kudlicka, K., Okuda, K. and Brown, OM. Jr. 1994. Characterisation of genes in the cellulose-synthesising operon (*acs* operon) of *Acetobacter xylinum*: implications for cellulose crystallisation. J. Bacteriol. 176:5735-5752.

Scott, JAG., Hall, AJ., Dagan, R., Dixon, JMS., Eykyn, SJ., Fenoll, A., Hortal, M., Jetté, LP., Jorgensen, JH., Lamothe, F., Latorre, C., Macfarlane, JT., Shlaes, DM., Smart, LE. and Taunay, A. 1996. Serogroup-specific epidemiology of *Streptococcus pneumoniae*: associations with age, sex, and geography in 7,000 episodes of invasive disease. Clin. Infect. Dis. 22:973-981.

Severin, A. and Tomasz, A. 1996. Naturally occurring peptidoglycan variants of *Streptococcus pneumoniae*. J. Bacteriol. 178:168-174.

Schreiber, JR. and Jacobs, MR. 1995. Antibiotic-resistant pneumococci. Pediatr. Clin. N. Amer. 42:519-537.

Shibaev, VN. 1986. Biosynthesis of bacterial polysaccharide chains composed of repeating units. Adv. Carbohyd. Chem. Biochem. 44:277-339.

Shoemaker, NB. and Guild, WR. 1974. Destruction of low efficiency markers is a slow process occurring at a heteroduplex stage of transformation. Mol. Gen. Genet. 128:283-290.

Silhavy, TJ., Berman, ML. and Enquist, LW. 1984. Experiments with gene fusions. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Silver, RP., Aaronson, W. and Vann, WF. 1984. Genetic and molecular analysis of *Escherichia coli* K1 antigen genes. J. Bacteriol. 157:568-575.

Skull, S., Leach, J. and Currie, BJ. 1996. *Streptococcus pneumoniae* carriage and penicillin/ceftriaxone resistance in hospitalised children in Darwin. Aust. NZ. J. Med. 26:391-395.

Skurnik, M., Venho, R., Toivanen, P. and al-Hendy, A. 1995. A novel locus of *Yersinia enterocolitica* serotype O:3 involved in lipopolysaccharide outer core biosynthesis. Mol. Microbiol. 17:575-594.

Smith, CJ. 1987. Nucleotide sequence analysis of Tn*4551*: use of *ermFS* operon fusions to detect promoter activity in *Bacteroides fragilis*. J. Bacteriol. 169:4589-4596.

Snaideck, DH., Schwartz, B., Lipman, H., Bogaerts, J., Butler, JC., Dagan, R., Echaniz-Aviles, G., LLoyd-Evans, N., Fenoll, A., Girgis, NI., Henrichsen, J., Klugman, K., Lehmann, D., Takala, AK., Vandepitte, J., Gove, S. and Breiman, RF. 1995. Potential interventions of childhood pneumonia: Geographic and temporal differences in serotype and serogroup distribution of sterile site pneumococcal isolates from children - implications for vaccine strategies. Pediatr. Infect. Dis. J. 14:503-510.

Soldo, B., Lazarevic, V., Margot, P. and Karamata, D. 1993. Sequencing and analysis of the divergon comprising *gtaB*, the structural gene of UDP-glucose pyrophosphorylase of *Bacillus subtilis* 168. J. Gen. Microbiol. 139:3185-3195.

Sorensen, UBS., Blum, J., Birch-Andersen, A. and Henrichsen, J. 1988. Ultrastructural localisation of capsules, cell wall polysaccharide, cell wall proteins, and F antigen in pneumococci. Infect. Immun. 43:876-878.

Sorensen, UBS., Henrichsen, J., Chen, H-C. and Szu, SC. 1990. Covalent linkage between the capsular polysaccharide and the cell wall peptidoglycan of *Streptococcus pneumoniae* revealed by immunochemical methods. Microb. Pathog. 8:325-334.

Southern, E. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98:503-517.

Steinfort, C., Wilson, R., Mitchell, T., Feldman, C., Rutman, A., Todd, H., Sykes, D., Walker, J., Saunders, K., Andrew, PW., Boulnois, GJ. and Cole, PJ. 1989. Effect of *Streptococcus pneumoniae* on human respiratory epithelium in vitro. Infect. Immun. 57:2006-2013.

Stevenson, G., Neal, B., Liu, D., Hobbs, M., Packer, NH., Batley, M., Redmond, JW., Lindquist, L. and Reeves, P. 1994. Structure of the O antigen of *Escherichia coli* and the sequence of its *rfb* gene cluster. J. Bacteriol. 176:4144-4156.

Stingele, F. and Mollet, B. 1996. Disruption of the gene encoding penicillin-binding protein 2b (*pbp2b*) causes altered cell morphology and cease in exoploysaccharide production in *Streptococcus thermophilus* Sfi6. Mol. Microbiol. 22:357-366.

Tabor, S. and Richardson, CC. 1985. A bacteriophage T7 RNA polymerase/ promoter system for controlled exclusive expression of specific genes. Proc. Natl. Acad. Sci. USA. 82:1074-1078.

Takala, AK., Eskola, J., Leinonen, M., Kayhty, H., Nissinen, A., Pekkanen, E. and Makela, PH. 1991. Reduction in oropharyngeal carriage of *Haemophilus influenzae* type b (Hib) in children immunised with an Hib conjugate vaccine. J. Infect. Dis. 164:982-986.

Takala, AK., Jero, J., Kela, E., Rönnberg, P-R, Koskenniemi, E. and Eskola, J. 1995. Risk factors for primary pneumococcal disease among children in Finland. JAMA 273:859-864.

Talbot, U., Paton, AW. and Paton, JC. 1996. Uptake of *Streptococcus pneumoniae* by respiratory epithelial cells. Infect. Immun. 64: 3772-3777.

Talkington, DF., Brown, BG., Tharpe, JA., Koenig, A. and Russell, H. 1996. Protection of mice against fatal pneumococcal challenge by immunisation with pneumococcal surface adhesin A (PsaA). Microb. Pathogen. 21:17-22.

Tamara, K. and Nei. M. 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.

Taunay, AE., Austrian, R., Landgraf, IM., Vieira, MFP. and Melles, CEA. 1990. Sorotipos de *Streptococcus pneumoniae* isolados de liquido cefalorraquidiano no periodo de 1977-1988 na cidade de Sao Paulo, Brasil. Rev. Inst. Med. Trop. Sao Paulo. 32:11-15.

Teele, DW., Klein, JO., Rosner, B. and the Greater Boston Otitis Media Study Group. 1989. Epidemiology of otitis media during the first seven years of life in children in Greater Boston: A prospective, cohort study. J. Infect. Dis. 160:83-94.

Thompson, JD., Higgins, DG. and Gibson, TJ. 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.

Tillet, WS. and Francis, T. Jr. 1930. Serological reaction in pneumonia with a non-protein somatic fraction of pneumococcus. J. Exp. Med. 52:561-571.

Tomasz, A. 1981. Surface components of *Streptococcus pneumoniae*. Rev. Infect. Dis. 3:190-211.

Towbin, H., Staehelin, T. and Gordon, J. 1979. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets : procedure and some applications. Proc. Natl. Acad. Sci. USA. 76:4350-4354.

Troy, FA., Frerman, FE. and Heath, EC. 1971. The biosynthesis of capsular polysaccharide in *Aerobacter aerogenes*. J. Biol. Chem. 246:118-.

Tuomanen, EI., Austrian, R. and Masure, HR. 1995. Pathogenesis of pneumococcal infection. New Engl. J. Med. 332:1280-1284.

Tuomanen, EI., Liu, H., Hengstler, B., Zak, O. and Tomasz, A. 1985. The induction of meningeal inflammation by components of the pneumococcal cell wall. J. Infect. Dis. 151:859-868.

Tuomanen, EI. and Masure, HR. 1997. Molecular and cellular biology of pneumococcal infection. Microb. Drug Resist. 3:297-308.

van Dam, JEG., Fleer, A. and Snippe, H. 1990. Immunogenicity and immunochemistry of *Streptococcus pneumoniae* capsular polysaccharide. Antonie van Leeuwenhoek. 58:1-47.

Van Eldere, J., Brophy, L., Loynds, B., Celis, P., Hancock, I., Carman, S., Kroll, JS. and Moxon. ER. 1995. Region II of the *Haemophilus influenzae* Type b capsulation locus is involved in serotype-specific polysaccharide synthesis. Mol. Microbiol. 15:107-118.

van Kranenburg, R., Marugg, JD. van Swam, II. Willem, NJ. and de Vos, WM. 1997. Molecular characterisation of the plasmid-encoded *eps* gene cluster essential for exopolysaccharide biosynthesis in *Lactococcus lactis*. Mol. Microbiol. 24:387-397.

Vaughan, EE. and de Vos, WM. 1995. Identification and characterisation of the insertion element IS*1070* from *Leuconostic lactis* NZ6009. Gene. 155:95-100.

Vázquez, M., Santana, O. and Quinto, C. 1993. The NodI and NodJ proteins from Rhizobium and Bradyrhizobium strains are similar to capsular polysaccharide secretion proteins from Gram-negative bacteria. Mol. Microbiol. 8:369-377.

Waltman, WD. II, McDaniel, LS., Gray, BM. and Briles, DE. 1990. Variation in the molecular weight of PspA (Pneumococcal surface protein A) among *Streptococcus pneumoniae*. Microb. Pathogen. 8:61-69.

Wang, L., Lui, D. and Reeves, PR. 1996. C-terminal half of *Salmonella enterica* WbaP (RfbP) is the galactosyl-1-phosphate transferase domain catalysing the first step of O-antigen synthesis. J. Bacteriol. 178:2598-2604.

Wang, L., Romana, LK. and Reeves, PR. 1992. Molecular analysis of a *Salmonella enterica* Group E1 *rfb* gene cluster: O Antigen and the genetic basis of the major polymorphism. Genetics. 130:429-443.

Watson, DA., Kapur, V., Musher, DM., Jacobson, JW. and Musser, JM. 1995. Identification, cloning and sequencing of DNA essential for encapsulation of *Streptococcus pneumoniae*. Curr. Microbiol. 31:251-259.

Watson, DA. and Musher, DM. 1990. Interruption of capsule production in *Streptococcus pneumoniae* serotype 3 by insertion of transposon Tn*916*. Infect. Immun. 58:3135-3138.

Watson, DA. and Musher, DM. 1996. The pneumococcus: Sugar-coated killer. Infect. Med. 13:373-381.

Watson, DA., Musher, DM., Jacobson, JW. and Verhoef, J. 1993. A brief history of the pneumococcus in biomedical research: A panoply of scientific discovery. Clin. Infect. Dis. 17:913-924.

Wenger, JD., Hightower, AW., Facklam, RR., Gaventa, S., Broome, CV. and the Bacterial Meningitis Study Group. 1990. Bacterial meningitis in the United States, 1986: Report of a multistate surveillance study. J. Infect. Dis. 162:1316-1323.

Weiser JN., Austrian, R., Sreenivasan, PK. and Masure, HR. 1994. Phase variation in pneumococcal opacity: relationship between colonial morphology and nasopharyngeal colonisation. Infect. Immun. 62:2582-2589.

Weiser, JN., Love, JM. and Moxon, ER. 1989. The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. Cell 59: 657-665.

Whitfield, C. 1995. Biosynthesis of lipopolysaccharide O antigens. Trends Microbiol. 3:178-185.

Whitfield, C., Amor, PA. and Koplin, R. 1997. Modulation of the surface architecture of Gram-negative bacteria by the action of surface polymer: lipid A-core ligase and by determinants of polymer chain length. Mol. Microbiol. 23:629-638.

Wren, BW., Russell, RR. and Tabaqchali, S. 1991. Antigenic cross-reactivity and functional inhibition by antibodies to *Clostridium difficile* toxin A, *Streptococcus mutans* glucan-binding protein, and a synthetic peptide. Infect. Immun. 59:3151-3155.

Yother, J. and Briles, DE. 1992. Structural properties and evolutionary relationships of PspA, a surface protein of *Streptococcus pneumoniae*, as revealed by sequence analysis. J. Bacteriol. 174:601-609.

Yother, J., Leopold, K., White, J. and Fischer, W. 1998. Generation and properties of a *Streptococcus pneumoniae* mutant which does not require choline or analogues for growth. J. Bacteriol. 180:2093-2101.

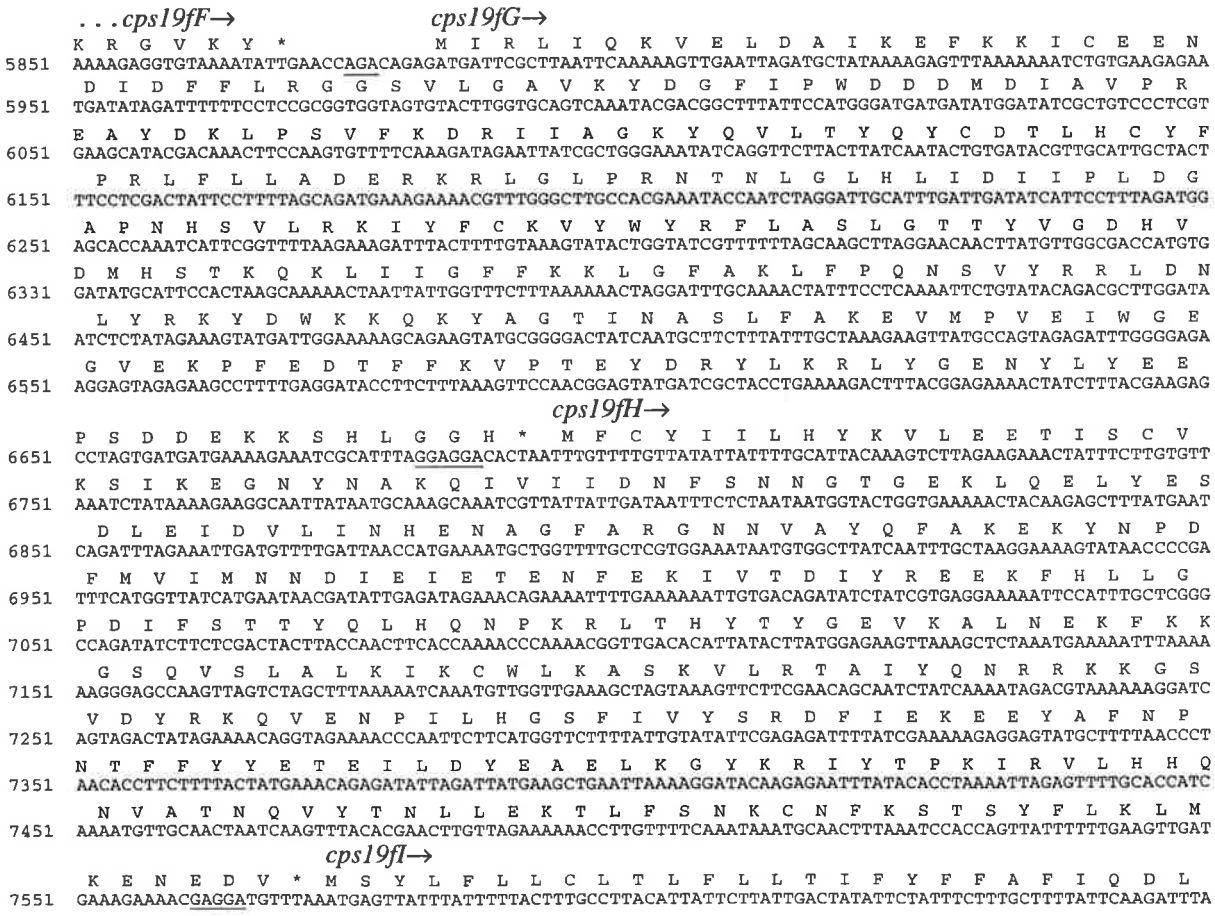
Yother, J., McDaniel, LS. and Briles, DE. 1986. Transformation of encapsulated *Streptococcus pneumoniae*. J. Bacteriol. 168:1463-1465.

Zhang, L., al-Hendy, A., Toivanen, P. and Skurnik, M. 1993. Genetic organisation and sequence of the *rfb* gene cluster of *Yersinia enterocolitica* serotype O:3: similarities to the dTDP-L-rhamnose biosynthesis pathway of *Salmonella* and to the bacterial polysaccharide transport systems. Mol. Microbiol. 9:309-321.

APPENDIX I

The nucleotide and deduced amino acid sequence of the distal portion of the *cps19f* locus.

The nucleotide sequence is numbered in accordance with GenBank accession number U09239 and is shown from nucleotide 5,851, within the distal portion of *cps19fF* to the *Hin*dIII site within *aliA*. The amino acid translation for each ORF is represented by single letter code above the first nucleotide of each codon. Possible ribosome binding sites are underlined.



```
      I  A  P  P  V  V  M  S  V  M  F  L  I  S  S  V  F  A  L  V  N  S  K  N  W  N  I  E  Y  S  G  I  A  Y
7651  ATTGCTCCTCCAGTAGTTATGTCTGTAATGTTTCTAATTAGTTCAGTATTTGCACTGGTTAATTCAAAAAACTGGAATATTGAATATAGTGGAATAGCCT
      I  L  I  I  S  G  I  I  I  F  S  I  P  L  M  A  L  K  S  P  N  F  N  T  E  V  K  I  A  D  R  L  I
7751  ATATTCTCATAATTAGTGGTATTATTATATTTTCGATTCCTTTAATGGCATTAAAATCACCTAATTTTAATACTGAGGTTAAGATTGCTGATCGATTAAT
      D  I  Q  F  W  K  I  A  L  T  I  I  I  D  L  F  I  L  Y  L  Y  R  K  E  I  Y  N  L  V  L  S  N  G
7851  TGATATCCAATTTTGGAAAATTGCTCTAACTATTATAATTGATCTCTTTATTTTGTATCTTTACAGGAAGGAAATATACAACCTTGTTCTTAGTAATGGA
      Y  T  G  S  N  I  Q  W  F  F  R  N  A  T  S  Y  E  G  E  L  T  V  R  T  F  I  R  V  L  I  R  V  I  D
7951  TATACGGGGTCAAATATTCAGTGGTTTTTTAGAAATGCAACGAGTTATGAAGGTGAATTGACAGTGCGAACTTTTATTCGAGTTCTCATTCGTGTTATTG
      V  S  A  Y  I  F  G  Y  T  F  I  N  N  F  L  I  Y  R  H  K  R  P  K  D  I  L  L  L  V  P  L  L  I
8051  ACGTATCTGCTTATATTTTTGGATATACTTTTATTAATAATTTTCTTATCTATCGCCATAAACGCCCTAAAGACATATTACTTTTAGTACCTTTATTAAT
      F  I  S  K  T  L  I  S  G  G  R  Q  D  I  I  K  I  L  I  A  Y  V  I  M  M  Y  I  Q  Q  K  R  K  V
8151  ATTTATTTCAAAAACTTTAATATCAGGAGGCCGGCAAGATATTATTAAAATTCTGATTGCCTATGTAATCATGATGTATATCCAACAAAAACGGAAAGTT
      G  W  N  R  V  I  S  H  K  Y  I  H  L  G  F  V  G  L  I  A  G  I  P  A  F  Y  Y  S  L  F  L  A  G  R
8251  GGATGGAATAGAGTCATATCTCATAAATATATTCACCTTGGATTTGTTGGTTTAATAGCAGGTATTCCAGCATTTTACTACTCTTTGTTTTTAGCCGGTC
      S  T  T  R  T  L  F  E  S  V  S  T  Y  L  G  G  S  I  Q  H  F  N  Q  Y  I  E  N  P  L  D  P  G  E
8351  GTTCAACGACTAGGACGCTATTTGAGAGTGTTTCGACCTATCTAGGAGGCTCAATTCAGCATTTTAATCAGTATATTGAAAATCCATTAGATCCTGGTGA
      V  F  G  S  E  T  L  V  P  I  L  N  I  L  G  E  M  G  L  V  N  Y  R  S  T  I  H  L  E  F  R  T  L
8451  AGTTTTTGGCAGTGAAACATTGGTGCCTATATTAAATATATTAGGGGAAATGGGCCTAGTTAATTATCGTAGTACAATTCATTTAGAATTTCGGACACTA
      G  V  T  V  G  N  V  Y  T  F  F  R  R  P  L  H  D  F  G  L  V  G  M  Y  V  F  V  F  A  V  G  A  F  F
8551  GGAGTTACTGTAGGAAATGTTTATACTTTTTTTAGAAGACCCTTGCATGATTTTGGTCTAGTTGGTATGTATGTATTTGTCTTTGCTGTAGGTGCTTTTT
      A  I  Y  Y  L  V  L  R  K  K  Q  V  G  F  N  L  D  I  H  T  I  I  Y  S  Y  V  F  Y  W  I  F  L  S
8651  TTGCTATTTATTATTTAGTTCTGAGAAAGAAACAGGTTGGTTTTAATTTGGATATTCATACCATTATTTATTCTTATGTCTTTTATTGGATTTTTTTATC
      S  I  E  Q  Y  S  F  T  M  I  S  L  Y  T  L  V  F  I  V  L  V  Y  F  M  A  I  F  Y  W  C  T  D  F
8751  ATCAATCGAGCAATACTCGTTCACAATGATTAGTCTATATACACTTGTATTTATTGTGTTGGTTTACTTTATGGCTATCTTTTACTGGTGTACAGATTTT
                                                                                                  cps19fJ→
      K  R  G  K  L  I  F  K  I  S  D  S  S  I  K  L  K  E  E  *                                 M  N  T  K  I  K
8851  AAAAGAGGAAAACTGATTTTTAAAATTTCTGACTCAAGTATCAAATTAAAAGAAGAATAACAGAATGTATAGGAGAGGGTAGATGAATACTAAAATTAAA
      N  I  I  T  S  F  S  Y  V  I  S  S  N  L  L  I  V  L  T  S  S  L  V  V  L  I  V  P  K  I  M  G  V  T
8951  AATATAATAACTAGTTTTTCTTATGTTATTTCTTCAAATCTGCTCATAGTTTTAACCTCATCACTAGTTGTTTTGATTGTTCCTAAAATAATGGGGGTAA
      E  Y  S  Y  W  Q  L  Y  I  F  Y  L  T  Y  I  G  F  F  H  L  G  W  I  D  G  I  Y  L  K  Y  G  G  L
9091  CTGAGTACAGTTACTGGCAACTTTATATTTTTTATCTGACCTATATCGGTTTTTTCCACTTGGGTTGGATTGATGGGATTTATCTCAAATATGGTGGCTT
      E  Y  T  N  L  D  R  K  Q  F  Y  S  Q  M  I  L  F  S  S  F  L  M  L  I  S  L  V  L  F  T  L  N  L
9151  AGAATATACAAATTTAGATAGAAAACAGTTTTATTCTCAGATGATTCTATTTTCTAGTTTCTTAATGCTAATCTCGCTGGTATTATTTACTTTGAACCTA
      I  T  V  R  D  E  N  A  R  Y  I  Y  N  M  A  I  I  S  M  T  V  T  N  L  R  T  L  Y  V  Y  I  L  Q  M
9251  ATAACTGTAAGGGATGAAAACGCAAGATATATTTATAATATGGCTATCATCAGCATGACAGTCACAAACTTAAGAACACTCTATGTTTATATCTTGCAGA
      T  N  R  L  K  D  S  S  V  I  L  I  S  D  R  V  L  Y  V  L  L  L  F  M  F  I  V  F  G  W  H  E  Y
9351  TGACAAATCGCTTGAAGGATAGTTCAGTCATTCTAATTAGTGATCGCGTTTTATATGTACTCCTTTTATTCATGTTTATTGTATTTGGATGGCATGAGTA
      K  V  M  I  W  A  D  I  L  G  R  T  F  S  L  M  L  S  F  W  I  C  K  D  I  V  F  Q  P  L  S  K  F
9451  CAAGGTCATGATTTGGGCTGATATTCTAGGTCGAACATTTTCTCTCATGCTTTCCTTCTGGATTTGTAAAGATATTGTGTTTCAGCCTTTGTCAAAATTT
      I  L  D  F  K  E  S  L  D  N  I  R  V  G  I  N  L  M  L  S  N  I  A  S  S  M  I  I  G  I  V  R  M  G
9551  ATCTTAGATTTCAAGGAGTCCCTTGATAATATCCGTGTTGGAATCAATTTAATGCTATCTAACATTGCGAGTAGCATGATTATAGGCATTGTTCGTATGG
      I  Q  W  N  W  N  I  E  T  F  G  K  V  S  L  T  L  S  I  S  N  L  L  M  T  F  I  N  A  I  G  L  V
9651  GAATTCAATGGAATTGGAATATTGAAACATTCGGGAAAGTGTCATTAACTTTGAGTATATCTAATTTATTAATGACTTTTATTAATGCCATTGGATTAGT
      I  F  P  L  I  K  R  T  K  T  E  N  L  P  K  I  Y  S  N  L  R  N  A  L  M  L  V  M  F  A  I  L  L
9751  TATCTTTCCTTTGATAAAGCGAACAAAGACTGAGAATTTACCTAAAATTTATTCTAATTTAAGAAATGCTTTGATGTTGGTTATGTTCGCAATCTTGCTC
      F  Y  Y  P  L  K  F  I  L  D  I  W  L  P  A  Y  K  D  A  L  V  F  M  A  L  I  F  P  M  S  V  Y  E  G
9851  TTCTACTATCCTTTAAAATTTATTCTTGATATTTGGCTTCCTGCTTATAAGGATGCCTTAGTTTTCATGGCCCTAATTTTTCCTATGTCAGTTTATGAAG
      K  M  A  L  V  I  N  T  Y  L  K  A  M  R  M  E  K  D  I  L  K  I  N  A  L  V  M  L  T  S  I  V  V
9951  GGAAAATGGCTTTGGTGATAAATACATATTTAAAAGCAATGAGAATGGAAAAAGACATTCTCAAAATTAATGCTTTGGTTATGTTAACTAGTATAGTAGT
      T  L  V  T  T  L  L  N  N  L  G  L  T  V  V  S  I  V  I  L  L  A  L  R  S  I  I  A  E  L  I  L
10051 GACATTAGTGACTACTCTACTACTAAATAATTTGGGGCTGACAGTTGTATCTATAGTTATTTTACTTGCTTTAAGAAGTATAATAGCTGAATTAATTTTA
      S  K  K  L  K  I  S  V  K  Q  D  I  A  L  E  L  L  M  T  I  I  F  I  S  S  S  W  Y  L  S  I  W  I  A
10151 TCCAAAAAACTGAAGATATCAGTCAAGCAAGACATTGCTTTAGAGTTACTTATGACGATTATATTTATTTCTTCAAGTTGGTATCTCTCTATTTGGATTG
      V  I  I  Y  L  L  A  Y  T  L  Y  L  Y  L  K  H  K  D  I  R  M  Y  I  E  Y  F  K  N  H  K  K  I  S
10251 CAGTAATAATTTATTTATTGGCGTATACTTTATATTTGTATTTAAAGCACAAAGATATCAGAATGTATATAGAATACTTTAAAAATCATAAAAAAATATC
                                                                                    cps19fK→
      *                                                                             M  K  K  I  M  L  V  F  G  T  R
10351 ATAAAAATTATATATCAATGAAATGGTAGATTACATTTCTACCTTTTTATCCATTTAGGAGGAAACGATGAAAAAGATAATGCTAGTTTTCGGTACACGT
      P  E  A  I  K  M  C  S  L  V  N  E  L  R  K  Q  E  D  M  K  T  V  V  C  V  T  G  Q  H  K  E  M  V  S
10451 CCAGAAGCAATAAAAATGTGTTCATTAGTCAATGAGTTGAGAAAACAGGAAGATATGAAAACAGTTGTTTGTGTAACTGGTCAACACAAGGAGATGGTTA
      P  V  L  D  L  F  G  V  Q  P  D  Y  D  L  E  I  M  K  A  N  Q  N  L  F  S  I  T  I  S  I  L  E  K
10551 GCCCTGTTTTAGATTTATTTGGAGTTCAACCAGACTATGATTTAGAAATTATGAAGGCTAATCAAAACTTGTTCTCTATAACAATAAGTATTTTAGAAAA
      I  K  P  V  L  E  K  E  Q  P  D  I  V  L  V  H  G  D  T  T  T  T  Y  A  A  A  L  A  A  F  Y  L  G
10651 AATAAAACCTGTGTTAGAAAAAGAACAACCAGATATTGTTTTGGTTCACGGTGATACTACTACGACATATGCAGCAGCTTTAGCGGCATTTTATCTAGGA
      I  K  V  G  H  V  E  A  G  L  R  T  Y  N  L  Q  S  P  F  P  E  E  F  N  R  Q  S  T  S  I  I  A  N  Y
10751 ATTAAAGTTGGTCATGTTGAAGCTGGATTACGAACTTACAACCTGCAAAGTCCTTTCCCAGAAGAGTTCAACAGACAATCGACTTCTATTATTGCAAATT
      H  F  A  P  T  E  L  A  K  E  N  L  I  K  E  G  R  N  N  I  Y  V  T  G  N  T  V  I  D  A  L  T  T
10851 ATCATTTTGCTCCTACAGAATTAGCTAAAGAAAATCTAATAAAAGAGGGTAGAAATAATATCTATGTTACTGGAAATACAGTGATTGATGCACTTACAAC
      T  V  Q  K  D  Y  T  H  P  D  L  L  N  D  G  N  R  L  I  L  L  T  A  H  R  R  E  N  L  G  E  P
10951 TACAGTACAAAAGGATTATACACACCCCGATCTAGATTTAAACGATGGAAATCGCCTCATCTTATTGACTGCTCATAGACGCGAAAATCTCGGAGAACCT
      M  R  H  M  F  R  A  V  K  R  V  L  N  E  Y  D  D  V  K  V  I  Y  P  I  H  K  N  P  L  V  R  E  T  A
11051 ATGAGACATATGTTTAGAGCTGTTAAACGAGTTTTAAATGAATATGACGATGTTAAAGTAATTTATCCAATTCATAAAAATCCATTGGTACGGGAAACAG
      T  E  I  F  G  D  T  E  R  I  Q  I  I  E  P  L  D  V  L  D  F  H  N  F  M  N  H  S  Y  M  I  L  T
11151 CTACAGAAATTTTTGGAGATACAGAACGTATTCAGATCATTGAACCTTTAGATGTTCTTGATTTTCACAACTTTATGAATCATAGTTATATGATTCTAAC
      D  S  G  G  V  Q  E  E  A  P  S  L  G  K  P  V  L  V  M  R  D  T  T  E  R  P  E  G  V  A  A  G  T
11251 TGACTCAGGAGGGGTCCAAGAAGAGGCACCTTCGTTAGGAAAGCCTGTATTGGTCATGCGAGATACGACAGAAAGACCTGAAGGAGTAGCTGCTGGAACG
      L  K  L  V  G  T  D  E  E  T  I  Y  Q  N  F  K  M  L  L  D  D  P  E  E  Y  K  K  M  S  R  A  S  N  P
11351 TTGAAATTGGTTGGGACTGATGAGGAGACTATTTATCAAAACTTTAAGATGCTTTTAGACGATCCCGAAGAATATAAAAAAATGAGTCGAGCTAGTAATC
                                                                                                  cps19fL→
      Y  G  N  G  D  A  S  K  Q  I  V  R  I  L  R  G  I  *                                       M  K
11451 CTTATGGAAATGGTGATGCTAGTAAACAGATTGTTCGAATTTTACGTGGAATTTGAGTGTGTTCAGATAAAGTAATATAGAAAGGTACCCTACTATGAAA
      G  I  I  L  A  G  G  S  G  T  R  L  Y  P  L  T  R  A  A  S  K  Q  L  M  P  V  Y  D  K  P  M  I  Y  Y
11551 GGTATTATTCTAGCAGGTGGTTCGGGGACACGTTTATATCCTTTGACTCGCGCTGCATCAAAACAACTTATGCCGGTTTATGATAAACCGATGATTTACT
```

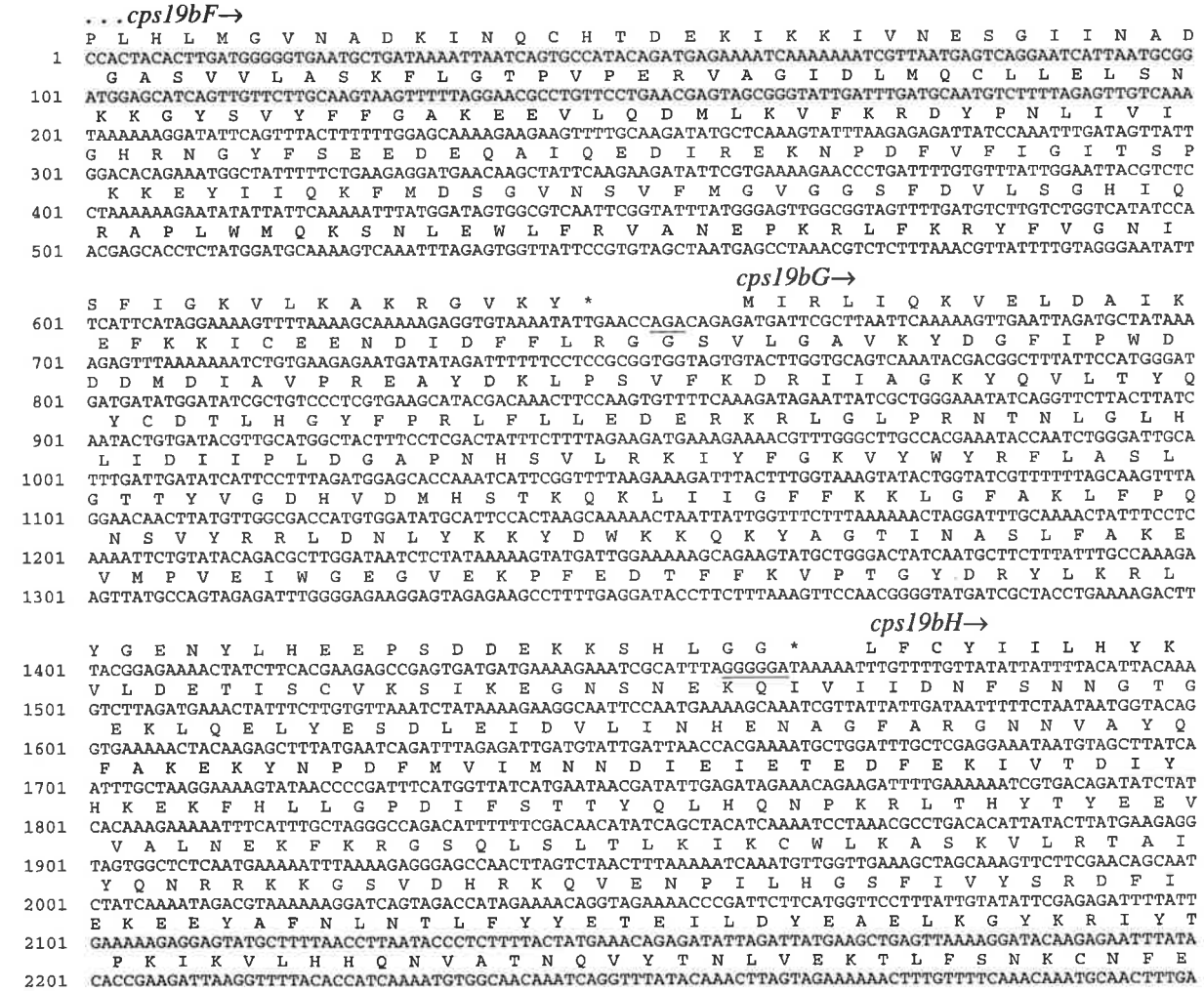
```
          P  L  S  T  L  M  L  A  G  I  R  D  I  L  I  I  S  T  P  Q  D  L  H  R  F  Q  E  L  L  Q  D  G  S
11651 ACCCACTTTCAACATTGATGTTGGCTGGGATTAGGGATATTTTGATTATTTCCACTCCACAGGATTTACATCGATTCCAAGAGCTTCTTCAAGACGGATC
          E  F  G  L  K  L  S  Y  A  E  Q  P  S  P  D  G  L  A  Q  A  F  I  I  G  E  E  F  I  S  D  D  S  V
11751 TGAGTTTGGGCTCAAACTTTCTTATGCAGAGCAACCAAGTCCAGATGGTTTGGCACAAGCCTTTATCATTGGGGAAGAGTTTATTTCTGATGATAGCGTT
          A  L  I  L  G  D  N  I  Y  H  G  S  G  L  S  K  M  L  Q  K  A  A  S  K  E  S  G  A  T  V  F  G  Y  H
11851 GCGCTGATCTTAGGTGACAATATCTATCACGGTTCTGGGCTTTCCAAGATGCTACAAAAGGCAGCGAGTAAGGAGTCGGGAGCAACTGTTTTTGGCTACC
          V  K  D  P  E  R  F  G  V  V  E  F  D  Q  D  M  K  A  I  S  I  E  E  K  P  E  Q  P  R  S  N  Y  A
11951 ATGTCAAGGATCCAGAGCGCTTTGGTGTGGTTGAGTTTGATCAGGATATGAAGGCTATTTCTATTGAAGAAAAGCCAGAGCAACCTCGTTCAAACTATGC
          V  T  G  L  Y  F  Y  D  N  D  V  V  E  I  A  K  S  I  K  P  S  P  R  G  E  L  E  I  T  D  V  N  K
12051 AGTTACAGGTCTCTATTTCTATGATAATGATGTAGTAGAGATTGCCAAGAGTATTAAACCAAGTCCTCGTGGTGAACTGGAAATTACAGATGTAAACAAG
          A  Y  L  D  R  G  D  L  S  V  E  L  M  G  R  G  F  A  W  L  D  T  G  T  H  E  S  L  L  E  A  S  Q  Y
12151 GCTTACCTAGATCGTGGTGATTTGTCTGTTGAGCTTATGGGACGTGGCTTTGCTTGGCTGGATACTGGAACTCATGAAAGTTTACTAGAGGCTTCACAGT
          I  E  T  V  Q  R  M  Q  N  V  Q  V  A  N  L  E  E  I  A  Y  R  M  G  Y  I  S  R  E  D  V  L  A  L
12251 ACATCGAAACAGTGCAACGGATGCAAAATGTTCAGGTAGCAAACTTAGAAGAAATTGCTTACCGTATGGGCTATATCAGTCGAGAAGATGTATTGGCCTT
                                                                                                  cps19fM→
          A  Q  P  L  K  K  N  E  Y  G  Q  Y  L  L  R  L  I  G  E  A  *  M  T  D  N  F  F  G  K  T  L  A  A
12351 AGCCCAACCACTTAAGAAAAATGAATACGGACAGTATCTGCTCCGTTTGATTGGAGAAGCATAGATGACAGATAATTTTTTCGGTAAGACGCTTGCGGCA
          R  K  V  E  A  I  P  G  M  L  E  F  D  I  P  V  H  G  D  N  R  G  W  F  K  E  N  F  Q  K  E  K  M  L
12451 CGCAAGGTTGAAGCTATTCCAGGCATGTTGGAGTTTGATATCCCCGTTCATGGAGATAATCGTGGCTGGTTTAAAGAAAATTTCCAAAAGGAAAAAATGC
          P  L  G  F  P  E  S  F  F  A  E  G  K  L  Q  N  N  V  S  F  S  R  K  N  V  L  R  G  L  H  A  E  P
12551 TTCCACTTGGATTTCCAGAGTCTTTCTTTGCAGAAGGAAAATTGCAAAACAATGTATCCTTCTCACGTAAAAATGTCCTTCGAGGTCTCCACGCAGAGCC
          W  D  K  Y  I  S  V  A  D  G  G  K  V  L  G  S  W  V  D  L  R  E  G  E  T  F  G  N  T  Y  Q  T  V
12651 TTGGGATAAGTACATCTCTGTAGCAGATGGAGGGAAAGTTCTGGGTTCTTGGGTTGATCTACGCGAGGGTGAAACCTTTGGGAATACCTATCAGACAGTG
          I  D  A  S  K  G  I  F  V  P  R  G  V  A  N  G  F  Q  V  L  S  D  T  V  S  Y  S  Y  L  V  N  D  Y  W
12751 ATTGATGCGAGTAAGGGAATCTTTGTTCCTCGAGGCGTAGCTAATGGCTTTCAAGTTTTATCAGATACAGTGTCATATAGCTATCTGGTCAATGATTACT
          A  L  E  L  K  P  K  Y  A  F  V  N  Y  A  D  P  S  L  G  I  E  W  E  N  I  A  E  A  E  V  S  E  A
12851 GGGCTCTTGAACTCAAACCCAAGTATGCCTTTGTGAACTACGCTGATCCAAGCCTTGGTATTGAATGGGAAAATATTGCAGAAGCAGAGGTTTCAGAAGC
                                                                                                  cps19fN→
          D  K  H  H  P  L  L  K  D  V  K  P  L  K  K  E  D  L  E  *                 M  T  E  Y  K  N  I  I  V  T
12951 AGACAAACATCATCCACTACTTAAGGATGTAAAACCTTTGAAAAAAGAAGATTTGGAATAAGGAAAGAATATGACTGAATACAAAAATATTATCGTGACA
          G  G  A  G  F  I  G  S  N  F  V  H  Y  V  Y  E  N  F  P  D  V  H  V  T  V  L  D  K  L  T  Y  A  G  N
13051 GGTGGAGCTGGCTTTATCGGTTCTAACTTTGTCCATTATGTTTACGAGAACTTTCCAGATGTTCATGTGACAGTCCTAGATAAGTTGACTTATGCTGGAA
          R  A  N  I  E  E  I  L  G  N  R  V  E  L  V  V  G  D  I  A  D  A  E  L  V  D  K  L  A  A  Q  A  D
13151 ACCGCGCGAATATTGAGGAAATTTTAGGTAATCGTGTTGAGTTAGTTGTTGGTGACATTGCTGATGCGGAGTTGGTAGACAAGTTGGCTGCTCAAGCAGA
          A  I  V  H  Y  A  A  E  S  H  N  D  N  S  L  N  D  P  S  P  F  I  H  T  N  F  I  G  T  Y  T  L  L
13251 TGCTATCGTTCATTATGCAGCGGAAAGCCACAATGATAATTCGCTCAATGATCCATCGCCATTTATTCATACTAACTTCATCGGAACCTATACTCTTTTA
          E  A  A  R  K  Y  D  I  R  F  H  H  V  S  T  D  E  V  Y  G  D  L  P  L  R  E  D  L  P  G  H  G  E  G
13351 GAAGCTGCTCGTAAGTATGATATTCGCTTCCACCATGTATCGACAGATGAAGTTTATGGGGATCTCCCTTTACGCGAAGATTTGCCAGGTCATGGTGAAG
          P  G  E  K  F  T  A  E  T  K  Y  N  P  S  S  P  Y  S  S  T  K  A  A  S  D  L  I  V  K  A  W  V  R
13451 GACCAGGTGAGAAATTTACTGCTGAAACAAAATACAACCCAAGCTCTCCGTACTCATCAACCAAGGCAGCCTCAGATTTGATTGTCAAAGCCTGGGTGCG
          S  F  G  V  K  A  T  I  S  N  C  S  N  N  Y  G  P  Y  Q  H  I  E  K  F  I  P  R  Q  I  T  N  I  L
13551 TTCTTTTGGAGTCAAGGCAACGATTTCCAACTGTTCAAATAACTACGGTCCTTATCAACATATCGAAAAATTCATCCCACGTCAGATTACTAACATCCTA
          S  G  I  K  P  K  L  Y  G  E  G  K  N  V  R  D  W  I  H  T  N  D  H  S  S  G  V  W  I  I  L  T  K  G
13651 AGTGGTATCAAGCCAAAACTTTACGGTGAAGGTAAGAACGTTCGTGATTGGATTCATACCAATGACCATTCTTCAGGAGTTTGGATAATCTTGACAAAAG
          Q  I  G  E  T  Y  L  I  G  A  D  G  E  K  N  N  K  E  V  L  E  L  I  L  K  E  M  G  Q  A  A  D  A
13751 GGCAAATCGGTGAAACCTACTTGATTGGGGCTGATGGTGAGAAGAACAATAAGGAAGTTTTGGAACTTATCCTTAAGGAAATGGGACAAGCTGCGGATGC
          Y  D  H  V  T  D  R  A  G  H  D  L  R  Y  A  I  D  A  S  K  L  R  D  E  L  G  W  K  P  E  F  T  N
13851 CTATGATCATGTGACTGACCGTGCAGGACATGACCTTCGCTATGCGATTGATGCTAGCAAGCTCCGTGATGAGTTGGGGTGGAAACCTGAATTTACCAAC
          F  E  A  G  L  K  A  T  I  K  W  Y  T  D  N  Q  E  W  W  K  A  E  K  E  A  V  E  A  N  Y  A  K  T  Q
13951 TTTGAAGCTGGGCTCAAGGCAACAATCAAGTGGTATACAGATAACCAAGAATGGTGGAAAGCAGAGAAAGAAGCTGTTGAAGCCAATTATGCTAAGACTC
                                                                                                  cps19fO→
          E  I  I  T  V  *                                                   M  I  L  I  T
14051 AGGAGATTATTACAGTATAAAAAGCAGGAAATAGCTGCTTTTTATTGCTATATTGGGAAGAGTTACATATTAGAAAGGTCTAGAGATGATTTTAATTACA
          G  A  N  G  Q  L  G  T  E  L  R  Y  L  L  D  E  R  N  E  E  Y  V  A  V  D  V  A  E  M  D  I  T  D  A
14151 GGGGCAAATGGCCAATTAGGAACGGAACTTCGCTATTTATTGGATGAACGTAATGAAGAATACGTGGCAGTAGATGTGGCTGAGATGGACATTACCGATG
          E  M  V  E  K  V  F  E  E  V  K  P  T  L  V  Y  H  C  A  A  Y  T  A  V  D  A  A  E  D  E  G  K  E
14251 CAGAAATGGTTGAGAAAGTTTTTGAAGAGGTGAAACCGACTTTAGTCTACCACTGTGCAGCCTACACCGCTGTTGATGCAGCAGAGGATGAAGGAAAAGA
          L  D  F  A  I  N  V  T  G  T  K  N  V  A  R  A  S  E  K  H  G  A  T  L  V  Y  I  S  T  D  Y  V  F
14351 GTTGGACTTCGCCATCAATGTGACGGGGACAAAAAATGTCGCAAGAGCATCTGAAAAGCATGGTGCAACTCTAGTTTATATTTCTACGGACTATGTCTTT
          D  G  K  K  P  V  G  Q  E  W  E  V  D  D  R  P  D  P  Q  T  E  Y  G  R  T  K  R  M  G  E  E  L  V  E
14451 GACGGTAAGAAACCAGTTGGACAAGAGTGGGAAGTTGATGACCGACCAGATCCACAGACAGAATATGGACGCACTAAGCGTATGGGGGAAGAGTTAGTTG
          K  H  V  S  N  F  Y  I  I  R  T  A  W  V  F  G  N  Y  G  K  N  F  V  F  T  M  Q  N  L  A  K  T  H
14551 AGAAGCATGTGTCTAATTTCTATATTATCCGTACTGCCTGGGTATTTGGAAATTATGGCAAAAACTTCGTTTTTACCATGCAAAATCTTGCGAAAACTCA
          K  T  L  T  V  V  N  D  Q  Y  G  R  P  T  W  T  R  T  L  A  E  F  M  T  Y  L  A  E  N  R  K  E  F
14651 TAAGACTTTAACAGTTGTAAATGACCAGTACGGTCGTCCGACTTGGACTCGTACCTTGGCTGAGTTCATGACCTACCTAGCTGAAAATCGTAAGGAATTT
          G  Y  Y  H  L  S  N  D  A  T  E  D  T  T  W  Y  D  F  A  V  E  I  L  K  D  T  D  V  E  V  K  P  V  D
14751 GGTTATTATCATTTGTCAAATGATGCGACAGAAGACACAACATGGTATGATTTTGCAGTTGAAATTTTGAAAGATACAGATGTCGAAGTCAAGCCAGTAG
          S  S  Q  F  P  A  K  A  K  R  P  L  N  S  T  M  S  L  A  K  A  K  A  T  G  F  V  I  P  T  W  Q  D
14851 ATTCCAGTCAATTTCCAGCCAAAGCTAAACGTCCGCTAAACTCAACGATGAGCCTGGCCAAAGCCAAAGCTACTGGATTTGTTATTCCAACTTGGCAAGA
          A  L  Q  E  F  Y  K  Q  E  V  R  *
14951 TGCATTGCAAGAATTTTACAAACAAGAAGTGAGATAAGTAGTAGAATGATTTTCTAGTCTAATAAAAGAGGCAGATAATGAACTCCAAAGGAGCTTAAGA
15051 TGTACGATTATCTTGTTGTTGGTGCTGGTCTCTTTGGCGCATAGCTTTGGCTCAGTTTCTATTATCGCTCACACCATCCATCAGAAGTTTAATCTGAAGG
15151 TACCCAATTATCGCCAAGAAGAAGATTGGGCTAGGATGGGTTTACCAATCACACGTAAGGAAATCTCTAATTGGCATATCAAGGCAAGTCAATACTATTT
15251 AGAGTCCCTTTATAACCTTTTACGAGAAAAGTTGTTAGAACAACCTCTTCTTCATGCGGATGAAACCTCTTATCGGGTCTTAGAAAGTGATAGTCAGCTG
15351 ACCTACTATTGGACCTTTTTGTCTGGGAAAGCTGAGAATCAAGTAATCACGCTTTACCACCATGATCAGTGTCGGAGTGGTTCGGTAGTGCAAGAATTCC
15451 TAGGAGATTATTCTGGCTATGTGCATTGTGATATGTTGCGGCAGTAACTTAGGACTTTAGTCCTCTAGTTCTGTCTATGCGATAGCAGTCCAAGGTTTAG
15551 GAGCAAGGCGACGCTAAGCTTGGTAAACTGCGAACCGCTAGAAGCTTATCGTCAACTGGAAGAAGCTGAACTTGTTGGATGTTGGGCGCATGTGAGAAGG
15651 AAATTTTTTGAAGTGCCCCCCAAGCAAGCAGATAAATCATCCTTAGGAGCTAAAGGTTTAGCTTATTGTGATCAGTTATTTTCCTTGGAAAGAGACTGGG
15751 AGGCTTTGCAGCTGATGAACGACTACAGAAACGTCAAGAACATCTCCAACCCCTAATGGAAGACTTCTTTGCTTAGTGCCGTCGTCAGTCAGTTTTATCG
15851 GGTTCAAAACTAGGAAGGGCAATTGAATACAGCCTCAAGTATGAAGAAACCTTTAAGACCATTTTAAAAGACGGACATCTGGTCCTTTCCAATAATCTAG
15951 CTGAACGCGCCATTAAATCATTGGTTATGGGACGGAGTAAAAGAGTCCAGTGGACTCTTTTAGCCTAAGCTAAATTTTAAAAAGCGAGGGTGGTTATTTT
```

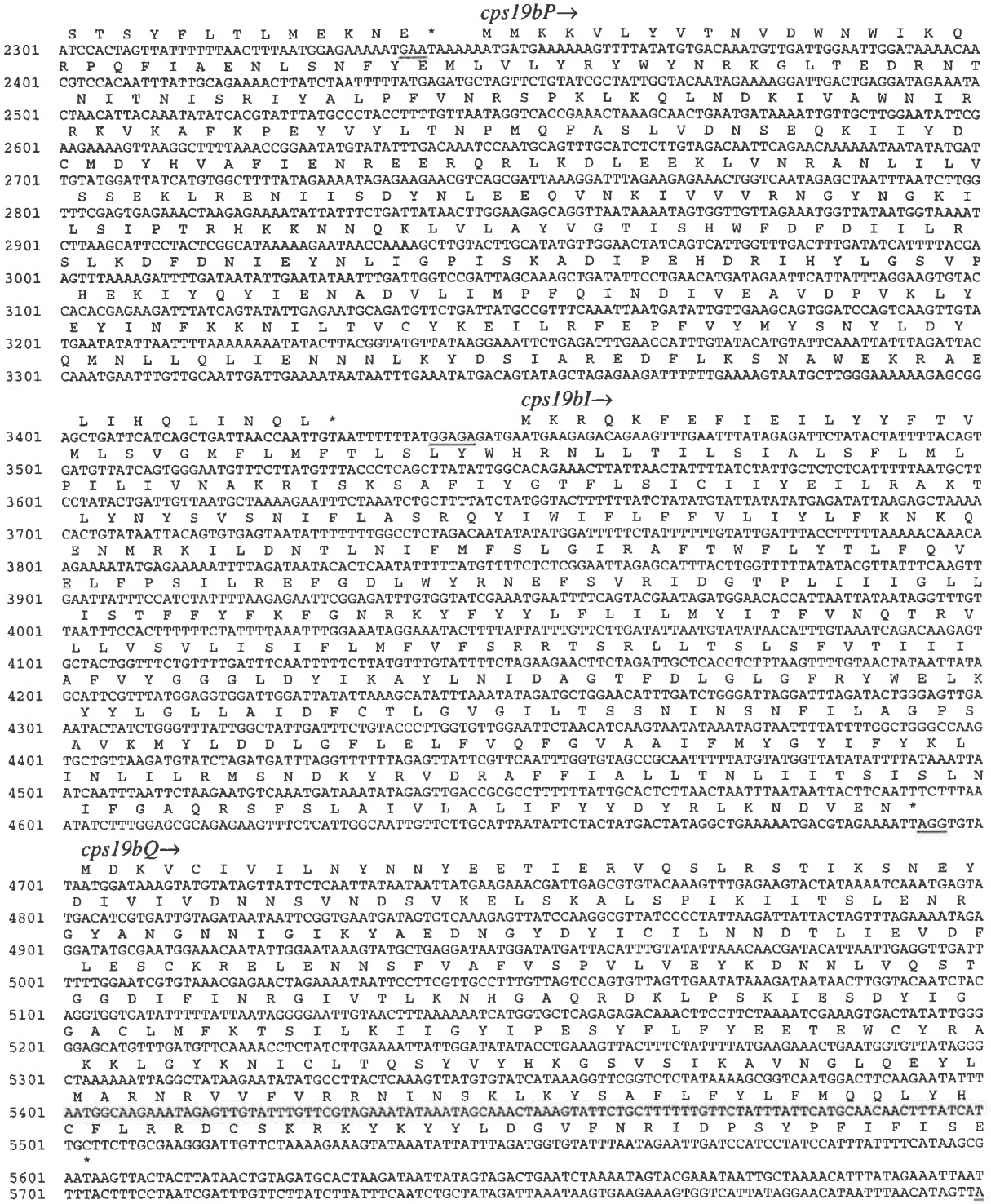
```
16051  CTCAAAGTTTTGAAGGAGCTAAAGCAAGAGCTATTATTATGAGCTTATTGGAAACAGCTAAACGTCATCAACTAAATAGTGAGAAATATCTATCCTATCT
16151  TCTAGAATGTCTTCCAAACGAGGAAACTCTCGTAAACAAAGAGGTTTTAGAGGCCTATTTACCATGGACTAAAGTTGTACAAGAAAAGTGCAAATAAGAA
16251  ATCTCCAGATTAGGAACTATCCGTGAGTTCACTAATCTGGAGATTTTTCAATAGACCTCGTTATTGGGCGGTTACGATATTCATATTTTTTGCAAAGATG
                                                                                                    aliA→
                                                                                                    M  K
16351  TTGTTTGAAAAATAATTTTCAAAAATTCTGAAAATTCTGTTGACATCTTTCTGAAAAGAGTCTATAATGGAGAGAAAGTTTTAAAGGAGAAAATGATGAA
         S  S  K  L  L  A  L  A  G  V  T  L  L  A  A  T  T  L  A  A  C  S  G  S  G  S  S  T  K  G  E  K  T
16451  AAGTTCAAAACTACTTGCCCTTGCGGGCGTGACATTATTGGCGGCGACTACTTTAGCTGCATGCTCTGGATCAGGTTCAAGCACTAAAGGTGAGAAGACA
         L  S  Y  I  Y  E  T  D  P  D  N  L  N  Y  L  T  T  A  K  A  A  T  A  N  I  T  S  D  V  V  D  G  L  L
16551  TTATCATACATTTATGAGACAGACCCTGATAACCTCAACTATTTGACAACTGCTAAGGCTGCGACAGCAAATATTACCAGTGACGTGGTTGATGGGTTGC
           E  N  D  R  Y  G  N  F  V  P  S  M  T  E  D  W  S  V  S  K  D  G  L  T  Y  T  Y  T  I  R  K  D  A
16651  TAGAAAATGATCGCTACGGGAACTTTGTGCCGTCTATGACTGAGGATTGGTCTGTATCCAAGGATGGGTTGACTTACACTTATACTATCCGTAAGGATGC
         K  W  Y  T  S  V  G  V  E  Y  A  A  V  K  A  Q  D  F  V  A  G  L  K  Y  A  A  D  K  K  S  D  A  L
16751  TAAATGGTATACTTCTGTGGGTGTAGAATACGCGGCAGTCAAGGCTCAGGACTTTGTGGCAGGGCTTAAGTATGCTGCTGATAAAAAATCAGATGCTCTT
         Y  L  V  Q  E  S  I  K  G  L  D  G  Y  V  K  G  E  I  K  D  F  S  Q  V  G  I  K  A  L  D  E  Q  T  V
16851  TACCTTGTTCAAGAATCAATCAAAGGATTGGATGGCTATGTCAAAGGGGAAATCAAAGATTTCTCTCAAGTAGGAATTAAGGCTCTGGATGAACAGACAG
           Q  Y  T  L  N  K  P  E  S  F
16951  TTCAGTACACTTTGAACAAACCAGAAAGCTT
```

APPENDIX II

The nucleotide and deduced amino acid sequence of the serotype specific region of the *cps19b* locus.

The amino acid translation for each ORF is represented by single letter code above the first nucleotide of each codon. Possible ribosome binding sites are underlined. The GenBank accession no. is AF004325.




cps19bP→
cps19bI→
cps19bQ→
cps19bR→

```
      Y  V  L  T  V  D  D  H  L  L  E  N  L  R  K  H  P  Q  F  I  K  K  F  N  Q  E  K  K  D  S  Q  L  L  R
6901  TATGTATTAACAGTAGACGATCATTTATTATAGAGAACCTAAGAAAACACCCTCAGTTTATTAAGAAGTTCAATCAAGAGAAAAAAGATAGCCAATTACTAC
        E  Y  I  K  H  L  T  K  N  R  N  E  A  Y  L  V  D  V  G  W  K  G  T  I  Q  D  S  I  K  K  A  L  P
7001  GTGAGTATATTAAGCATTTAACAAAAAACCGAAATGAGGCATATTTAGTTGATGTAGGTTGGAAAGGAACGATACAAGATAGTATCAAAAAGGCTCTTCC
        D  K  R  I  V  G  Y  Y  L  G  L  M  L  N  V  Y  S  V  E  N  K  T  D  K  T  G  L  L  F  S  D  Y  P
7101  AGACAAGAGAATAGTAGGATATTATCTAGGATTGATGCTCAATGTTTATTCAGTAGAAAATAAAACGGATAAAACTGGTTTATTATTTTCTGATTATCCA
        S  K  S  K  F  Y  D  I  V  S  R  N  F  G  F  Y  E  D  I  F  V  A  D  H  G  P  V  L  K  Y  K  K  E  S
7201  AGTAAATCGAAATTTTATGATATTGTGAGTAGAAATTTTGGTTTTTATGAAGATATTTTTGTAGCAGATCATGGTCCAGTTTTGAAATATAAAAAAGAGA
        D  I  I  P  I  I  D  D  D  K  K  H  V  S  I  Y  Q  A  V  K  D  Y  Q  E  E  L  V  L  G  F  S  E  I
7301  GTGATATTATCCCCATTATTGACGATGATAAGAAACATGTGAGTATTTATCAGGCAGTTAAAGATTATCAAGAAGAGTTAGTATTAGGATTTTCAGAAAT
        L  E  A  Y  K  K  M  K  F  L  P  F  E  Q  K  N  L  W  L  T  M  S  L  K  K  E  C  I  Y  I  P  K  L
7401  TTTGGAAGCTTATAAGAAAATGAAGTTTCTCCCTTTTGAACAAAAAAATCTGTGGCTAACGATGTCGCTAAAAAAAGAATGTATTTATATACCTAAGTTA
        Q  S  F  S  E  S  L  K  E  K  V  V  E  N  F  G  E  I  V  T  L  K  T  T  K  K  S  I  K  T  L  L  R  E
7501  CAATCATTTTCTGAATCTTTAAAAGAAAAAGTTGTTGAAAATTTTGGTGAGATAGTAACTCTTAAGACTACGAAAAAATCTATAAAAACGTTATTAAGAG
        K  S  D  L  L  W  V  D  F  V  Y  R  L  F  G  G  V  N  F  L  F  I  P  E  L  Y  T  R  V  I  F  L  L
7601  AGAAAAGCGATTTACTTTGGGTTGATTTCGTTTATAGATTGTTTGGTGGTGTCAATTTCCTATTTATTCCAGAATTATATACAAGAGTTATATTTTTATT
        K  Y  L  D  L  K  L  R  L  K  N  Y  G  E  *
```

cps19bJ→

```
                                                                  M  G  N  K  S  I  K  L  N  A  L  L  N  I  V  L  T  L  S  N  I  I
7701  GAAATATTTAGATTTGAAATTGAGGTTGAAAAATTATGGGGAATAAATCCATAAAGTTGAATGCATTATTAAATATTGTCCTGACGCTATCAAATATCAT
        F  P  L  I  T  F  P  Y  I  S  R  I  L  N  P  N  G  I  G  L  T  S  F  F  S  S  I  G  N  Y  G  I  L
7801  TTTCCCATTAATCACTTTTCCTTATATATCTAGAATATTGAATCCAAATGGTATAGGTTTAACTTCATTTTTTAGTTCAATAGGGAATTATGGTATTTTA
        L  A  S  L  G  I  S  T  Y  G  I  K  A  V  A  S  V  R  D  D  R  D  K  L  S  K  V  V  Q  E  L  M  I  I
7901  CTTGCTTCTCTGGGAATTTCAACTTATGGTATCAAAGCAGTAGCAAGTGTTAGAGATGATAGAGATAAGTTGTCAAAAGTAGTACAGGAGTTAATGATTA
        N  V  A  M  S  I  I  T  T  A  I  L  L  F  M  T  I  F  I  T  Q  L  N  R  E  F  S  L  L  L  I  T  C
8001  TAAACGTTGCTATGTCTATAATAACAACTGCAATACTATTATTTATGACTATATTTATAACACAATTGAATAGAGAATTTTCACTCCTATTGATCACATG
        G  T  I  L  S  S  P  F  A  L  N  W  L  Y  S  G  M  E  E  Y  T  Y  I  T  T  R  S  V  V  F  K  I  L
8101  TGGGACTATTTTATCTTCTCCTTTCGCCTTAAATTGGTTGTATAGTGGAATGGAAGAATATACGTATATTACTACTAGGTCAGTAGTGTTTAAAATTCTA
        S  L  I  L  I  F  L  L  V  K  R  P  E  D  Y  I  V  F  A  S  I  S  L  F  S  S  L  S  S  N  I  L  N  L
8201  TCATTAATATTGATTTTTCTACTTGTGAAAAGGCCAGAGGATTATATTGTTTTTGCTAGTATTTCATTGTTTTCTTCTCTAAGTTCAAATATCTTAAATC
        W  H  S  R  H  F  I  N  I  K  L  Y  K  N  L  Q  F  K  Y  H  F  K  P  M  W  Y  L  F  A  S  L  L  A
8301  TATGGCATAGCCGACATTTCATTAATATTAAATTATATAAAAATTTACAATTTAAATATCATTTTAAACCAATGTGGTATTTATTTGCCTCATTACTTGC
        V  N  I  Y  T  N  L  D  T  V  M  L  G  F  I  N  G  N  E  A  V  G  Y  Y  S  V  A  S  K  V  K  W  I
8401  AGTAAATATTTATACTAATTTAGATACAGTGATGCTCGGTTTTATTAATGGTAATGAGGCTGTGGGATACTATTCTGTGGCATCAAAGGTTAAGTGGATT
        L  L  S  L  I  T  S  I  S  A  V  L  L  L  R  L  S  F  Y  I  S  K  N  D  T  S  N  F  I  K  M  L  K  E
8501  TTGCTTTCTCTTATTACATCTATTAGTGCAGTTTTGCTACTGAGACTTTCATTTTATATTAGTAAAAATGACACCTCGAATTTTATAAAAATGTTAAAGG
        S  S  A  V  I  F  F  I  A  I  P  L  M  V  F  F  I  V  E  A  K  D  S  I  L  L  L  G  G  S  Q  Y  L
8601  AGTCATCTGCGGTTATATTTTTTATTGCGATTCCATTGATGGTATTCTTTATTGTAGAGGCGAAAGATAGTATCTTATTACTAGGAGGAAGTCAGTATCT
        P  A  T  L  A  M  Q  I  L  M  P  I  L  L  I  S  G  F  S  N  I  T  G  N  Q  I  L  I  P  M  N  R  E
8701  TCCTGCGACTTTAGCGATGCAAATACTTATGCCAATTTTACTTATTTCTGGTTTCTCGAATATTACAGGAAATCAAATATTGATTCCAATGAATAGAGAA
        K  Y  F  M  V  A  V  T  I  G  A  V  I  N  L  I  L  N  L  L  L  M  P  K  F  G  I  I  G  A  S  V  A  T
8801  AAATATTTTATGGTTGCAGTAACGATTGGTGCTGTGATTAATCTTATTTTGAATCTACTGTTAATGCCTAAGTTTGGAATTATTGGTGCTTCTGTCGCAA
        L  F  A  E  L  S  Q  M  T  V  Q  L  H  F  S  K  E  Y  L  V  S  N  I  S  I  K  S  L  V  N  V  I  I
8901  CTCTTTTTGCGGAATTGTCGCAGATGACGGTACAATTACATTTTTCAAAAGAATATTTAGTATCAAATATATCGATAAAGAGTTTGGTTAATGTGATAAT
        A  T  V  V  S  I  I  P  L  I  I  L  N  Q  L  I  T  I  T  I  P  F  Y  S  L  M  L  A  G  F  A  F  F
9001  TGCAACAGTTGTTTCTATAATACCACTAATCATTTTGAATCAGCTGATAACGATAACTATACCATTTTATTCTCTAATGCTAGCAGGTTTTGCTTTCTTT
        S  L  Y  L  V  I  L  L  L  L  K  E  E  V  T  I  Q  L  F  S  L  L  A  K  K  K  *
9101  TCATTATATTTAGTAATTCTGCTTTTATTAAAGGAGGAAGTGACGATTCAATTATTTTCTCTTCTTGCAAAGAAGAAGTAAATTGGTCAGAAATTGAAAT
```

cps19bK→

```
                                                                  M  K  I  M  L  V  F  G  T  R  P  E  A  I  K  M  C  P  L
9201  GTATAAACAAATAAATAATTTAATTATTAATTGAGGAGGAAATCATGAAGATAATGCTAGTTTTTGGTACACGTCCAGAAGCGATAAAAATGTGTCCATT
        V  N  E  L  K  K  Q  A  D  M  E  T  V  V  C  V  T  G  Q  H  K  E  M  V  S  P  V  L  E  L  F  G  V
9301  AGTGAATGAGTTGAAAAAACAGGCAGATATGGAAACAGTTGTTTGTGTAACTGGTCAACACAAGGAGATGGTTAGCCCTGTTTTGGAATTGTTTGGAGTT
        Q  P  D  Y  D  L  E  I  M  K  A  N  Q  T  L  F  S  I  T  T  S  I  L  E  K  I  K  P  V  L  E  E  E  Q
9401  CAACCAGATTATGATTTAGAAATTATGAAAGCTAATCAAACCTTGTTTTCTATAACAACAAGTATTTTAGAAAAAATTAAACCTGTTTTAGAAGAAGAAC
        P  D  I  V  L  V  H  G  D  T  T  T  T  Y  A  A  A  L  A  A  F  Y  L  G  I  K  V  G  H  V  E  A  G
9501  AACCAGATATTGTTCTAGTTCACGGTGACACTACTACGACATATGCAGCAGCTTTGGCAGCATTTTATCTAGGAATTAAGGTTGGTCATGTTGAAGCTGG
        L  R  T  Y  N  L  Q  S  P  F  P  E  E  F  N  R  Q  S  T  S  I  I  A  N  Y  H  F  A  P  T  E  L  A
9601  GTTACGAACTTACAACCTGCAAAGTCCCTTCCCAGAAGAATTTAATAGACAATCGACTTCTATTATTGCAAATTATCATTTTGCCCCTACAGAATTAGCT
        K  E  N  L  T  K  E  G  R  N  N  V  Y  V  T  G  N  T  V  I  D  A  L  K  T  T  V  Q  K  D  Y  T  H  P
9701  AAAGAAAATCTAACAAAAGAGGGTAGAAACAATGTTTATGTGACAGGTAATACGGTAATTGATGCACTTAAAACTACAGTACAAAAGGATTATACACACC
        D  L  D  L  N  A  D  N  R  L  I  L  L  T  A  H  R  R  E  N  L  G  E  P  M  K  H  M  F  R  A  V  K
9801  CTGATTTAGATTTAAACGCTGATAATCGTCTCATTCTATTGACTGCTCATAGACGTGAAAATCTCGGAGAACCTATGAAACACATGTTTAGAGCTGTTAA
        R  I  L  N  E  Y  D  D  V  K  V  I  Y  P  I  H  K  N  P  L  V  R  E  T  A  A  E  I  F  G  D  I  E
9901  ACGAATTTTAAATGAATATGACGATGTTAAGGTAATTTATCCAATTCATAAGAATCCTTTGGTTCGTGAAACAGCTGCGGAAATTTTTGGAGACATAGAA
        R  I  Q  I  I  E  P  L  D  V  L  D  F  H  N  F  M  N  N  S  Y  M  I  L  T  D  S  G  G  V  Q  E  E  A
10001 CGAATTCAGATTATTGAACCTTTAGATGTTCTGGATTTTCATAACTTTATGAATAATAGTTACATGATTTTAACTGACTCAGGAGGAGTTCAGGAAGAAG
        P  S  L  G  K  P  V  L  V  M  R  D  T  T  E  R  P  E  G  V  A  A  G  T  L  K  L  V  G  T  D  E  E
10101 CGCCTTCGTTAGGAAAGCCTGTATTGGTCATGCGAGATACGACAGAAAGACCTGAAGGAGTAGCTGCTGGAACGTTGAAATTGGTTGGGACTGATGAGGA
        T  I  Y  Q  N  F  K  M  L  L  D  D  S  E  E  Y  K  K  M  S  Q  A  S  N  P  Y  G  N  G  D  A  S  K
10201 GACTATTTATCAAAATTTTAAGATGCTTTTAGACGATTCCGAAGAATATAAAAAAATGAGTCAAGCTAGTAATCCTTATGGAAATGGTGATGCTAGTAAA
```

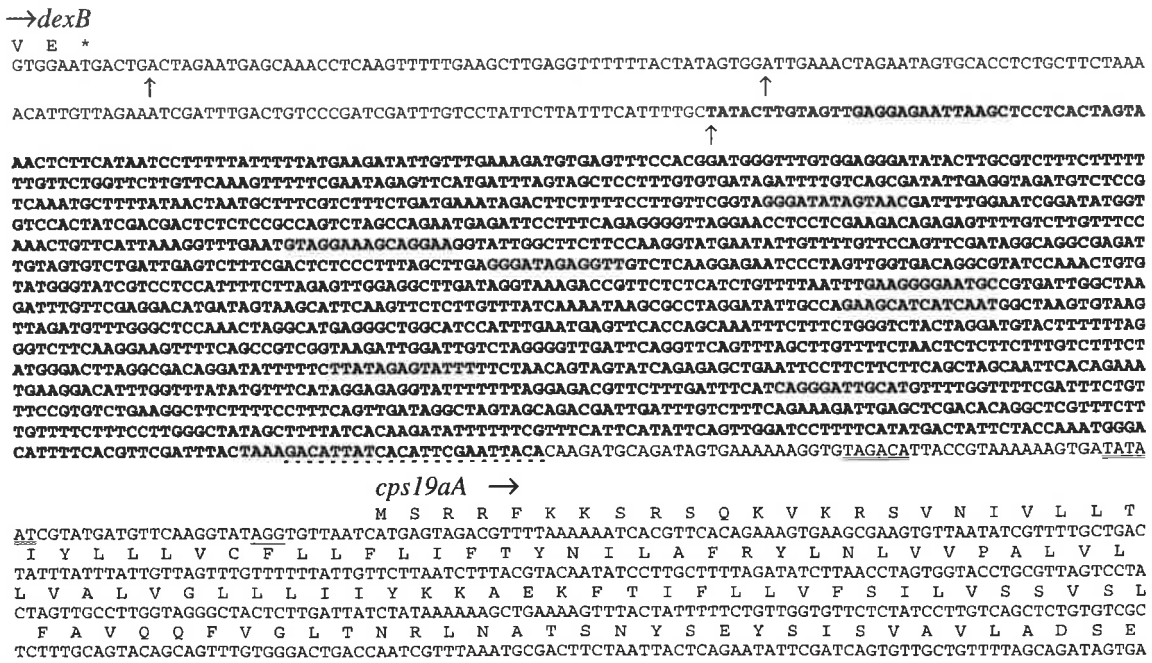
cps19bL→

```
      Q  I  V  R  I  L  R  G  I  *                                    M  K  G  I  I  L  A  G  G  S  G
10301 CAGATTGTTCGAATTTTACGTGGAATTTGAGTGTGTTTAGATAAAGTAATACAGAAAGGTACCCTACTATGAAAGGTATTATTCTAGCAGGTGGTTCGGG
        T  R  L  Y  P  L  T  R  A  A  S  K  Q  L  M  P  V  Y  D  K  P  M  I  Y  Y  P  L  S  T  L  M  L  A
10401 GACTCGCTTGTATCCTTTGACTCGCGCTGCATCAAAACAACTGATGCCGGTTTATGATAAACCCATGATTTACTACCCACTTTCAACATTGATGTTGGCT
        G  I  R  D  I  L  I  I  S  T  P  Q  D  L  H  R  ...
10501 GGAATTAGGGATATTTTGATTATTTCCACTCCACAGGATTTACATCGAT
```

APPENDIX III

The nucleotide and deduced amino acid sequence of the complete *cps19a* locus.

The nucleotide sequence is numbered in accordance with GenBank accession number AF094575 and is shown from the 3' end of *dexB* to the 5' region of *aliA*. The amino acid translation for each ORF is represented by single letter code above the first nucleotide of each codon. The putative -35 and -10 promoter sequences are double underlined and possible ribosome binding sites are singly underlined. The sequence which corresponds to the partial copy of IS*1202* is shown in bold and the inverted repeat is underlined with a dashed line. The three arrows in the 5' intergenic region indicate the point of three separate deletions when compared to the *cps19f* sequence.



```
      I  E  N  V  T  Q  L  M  S  V  T  A  P  T  G  T  D  N  E  N  I  Q  K  L  L  A  D  I  K  S  S  Q  N
2101  GATTGAGAATGTTACGCAACTGATGAGTGTGACAGCACCGACTGGGACTGATAATGAAAATATTCAAAAACTACTAGCTGATATCAAATCAAGTCAGAAT
      T  D  L  T  V  N  Q  S  S  S  Y  L  A  A  Y  K  S  L  I  A  G  E  T  K  A  I  V  L  N  S  V  F  E  N
2201  ACCGATTTGACGGTCAATCAGAGTTCGTCTTACTTGGCAGCTTATAAGAGTTTGATTGCAGGAGAGACTAAGGCCATTGTCTTAAATAGTGTCTTTGAAA
      I  I  E  S  E  H  P  D  Y  A  S  K  I  K  K  I  Y  T  K  G  F  T  K  K  V  E  A  P  K  T  S  K  D
2301  ATATCATCGAATCGGAGCATCCAGATTACGCATCGAAGATAAAAAAGATTTATACCAAGGGATTCACTAAAAAAGTAGAAGCTCCTAAGACGTCTAAGGA
      Q  S  F  N  I  Y  V  S  G  I  D  T  Y  G  P  I  S  S  V  S  R  S  Y  V  N  I  L  M  T  V  N  R  D
2401  TCAGTCTTTCAATATCTATGTTAGTGGAATTGACACCTATGGTCCTATTAGTTCGGTGTCGCGTTCATATGTCAATATCCTGATGACTGTCAATCGAGAT
      T  K  K  I  L  L  T  T  T  P  R  D  A  Y  V  P  I  A  D  G  G  N  N  Q  K  D  K  L  T  H  A  G  I  Y
2501  ACCAAGAAAATTCTCTTGACCACAACGCCACGTGATGCTTATGTACCAATCGCAGATGGTGGGAATAATCAAAAAGATAAATTGACCCATGCGGGCATCT
      G  V  D  S  S  I  H  T  L  E  N  L  Y  G  V  D  I  N  Y  Y  V  R  L  N  F  T  S  F  L  K  L  I  D
2601  ATGGAGTTGATTCGTCCATCCACACTTTAGAAAATCTCTATGGAGTGGATATCAATTACTATGTGCGATTGAACTTCACTTCTTTCTTGAAATTGATCGA
      L  L  G  G  V  D  V  Y  N  D  Q  D  F  T  S  L  H  G  K  F  H  F  P  V  G  N  V  H  L  D  S  E  Q
2701  TTTGTTGGGTGGGGTAGATGTTTATAATGACCAAGATTTTACATCTCTACATGGGAAGTTCCATTTCCCAGTTGGAAATGTCCACTTAGATTCTGAGCAG
      A  L  G  F  V  R  E  R  Y  S  L  A  D  G  D  H  D  R  G  R  N  Q  Q  K  V  I  A  A  I  L  Q  K  L  T
2801  GCTCTGGGATTTGTTCGTGAGCGCTATTCCCTAGCGGATGGTGATCATGACCGCGGGCGCAATCAACAAAAGGTGATTGCGGCCATCCTTCAAAAATTAA
      S  S  E  A  L  K  N  Y  S  R  I  I  D  S  L  Q  D  S  I  Q  T  N  M  P  L  E  T  M  I  N  L  V  N
2901  CTTCGAGCGAAGCGCTGAAAAATTATAGTAGGATCATTGATAGCTTGCAAGATTCTATCCAAACAAACATGCCACTTGAGACCATGATAAACTTGGTCAA
      A  Q  L  E  S  G  G  T  Y  K  V  N  S  Q  D  L  K  G  M  G  R  T  D  L  P  S  Y  A  M  P  D  S  N
3001  TGCTCAGTTAGAAAGTGGTGGAACGTACAAAGTGAATTCGCAAGACTTGAAGGGTATGGGACGGACGGATCTTCCTTCCTATGCGATGCCAGATAGTAAC
```

cps19aB →

```
      L  Y  M  M  E  I  N  D  S  S  L  A  S  V  K  T  A  I  Q  D  V  L  E  G  R  *     M  I  D  I  H  S  H
3101  CTCTATATGATGGAAATTAACGACAGTAGCCTTGCATCTGTCAAAACTGCTATTCAGGATGTGTTGGAGGGCAGATGAAATGATTGATATTCATTCGCAC
      I  V  F  D  V  D  D  G  P  K  S  R  E  E  S  K  A  L  L  A  E  S  Y  R  Q  G  V  R  T  I  V  S  T  S
3201  ATCGTCTTTGATGTAGATGATGGTCCCAAGTCAAGAGAGGAAAGCAAGGCTCTCTTGGCAGAATCCTACAGGCAGGGGGTGCGAACCATTGTCTCTACCT
      H  R  R  K  G  M  F  E  T  P  E  E  K  I  A  E  N  F  L  Q  V  R  E  I  A  K  E  V  A  S  D  L  V
3301  CTCACCGTCGCAAGGGCATGTTTGAAACTCCGGAAGAGAAGATAGCAGAAAACTTTCTTCAGGTTCGGGAAATAGCTAAGGAAGTGGCGAGTGACTTGGT
      I  A  Y  G  A  E  I  Y  Y  T  S  D  V  L  D  K  L  E  K  N  R  I  P  T  L  N  N  S  R  Y  A  L  I
3401  CATTGCTTACGGGGCTGAAATTTACTACACATCAGATGTTTTGGATAAGCTGGAAAAAAAATCGGATTCCGACCCTCAATAATAGTCGTTATGCCTTGATA
      E  F  S  M  N  T  P  Y  R  D  I  H  S  A  L  S  K  I  L  M  L  G  I  T  P  V  I  A  H  I  E  R  Y  D
3501  GAGTTTAGTATGAACACCCCTTATCGCGATATTCATAGTGCCTTGAGTAAAATATTGATGTTGGGAATTACTCCAGTCATTGCCCACATTGAGCGCTATG
      A  L  E  N  N  E  K  R  V  R  E  L  I  N  M  G  C  Y  T  Q  V  N  S  S  H  V  L  K  S  K  L  F  G
3601  ATGCACTTGAAAATAATGAAAAACGCGTTCGAGAGCTGATCAATATGGGCTGTTACACGCAAGTAAATAGTTCACATGTCCTCAAATCCAAACTTTTTGG
      E  R  Y  K  F  M  K  K  R  A  Q  Y  F  L  E  R  D  L  V  H  I  I  A  S  D  M  H  N  V  D  S  R  P
3701  AGAACGTTATAAATTCATGAAAAAAAGAGCTCAGTATTTTTTGGAGCGTGACTTGGTTCATATCATTGCAAGTGATATGCATAATGTGGACAGCAGACCT
      P  H  M  A  E  A  Y  D  L  V  S  Q  K  Y  G  E  A  K  A  Q  E  L  F  I  D  N  P  R  K  I  V  M  D  Q
3801  CCCCATATGGCAGAAGCATATGACCTTGTTTCCCAAAAATACGGAGAAGCGAAGGCTCAGGAACTTTTTATAGACAATCCTCGAAAAATTGTAATGGATC
```

cps19aC →

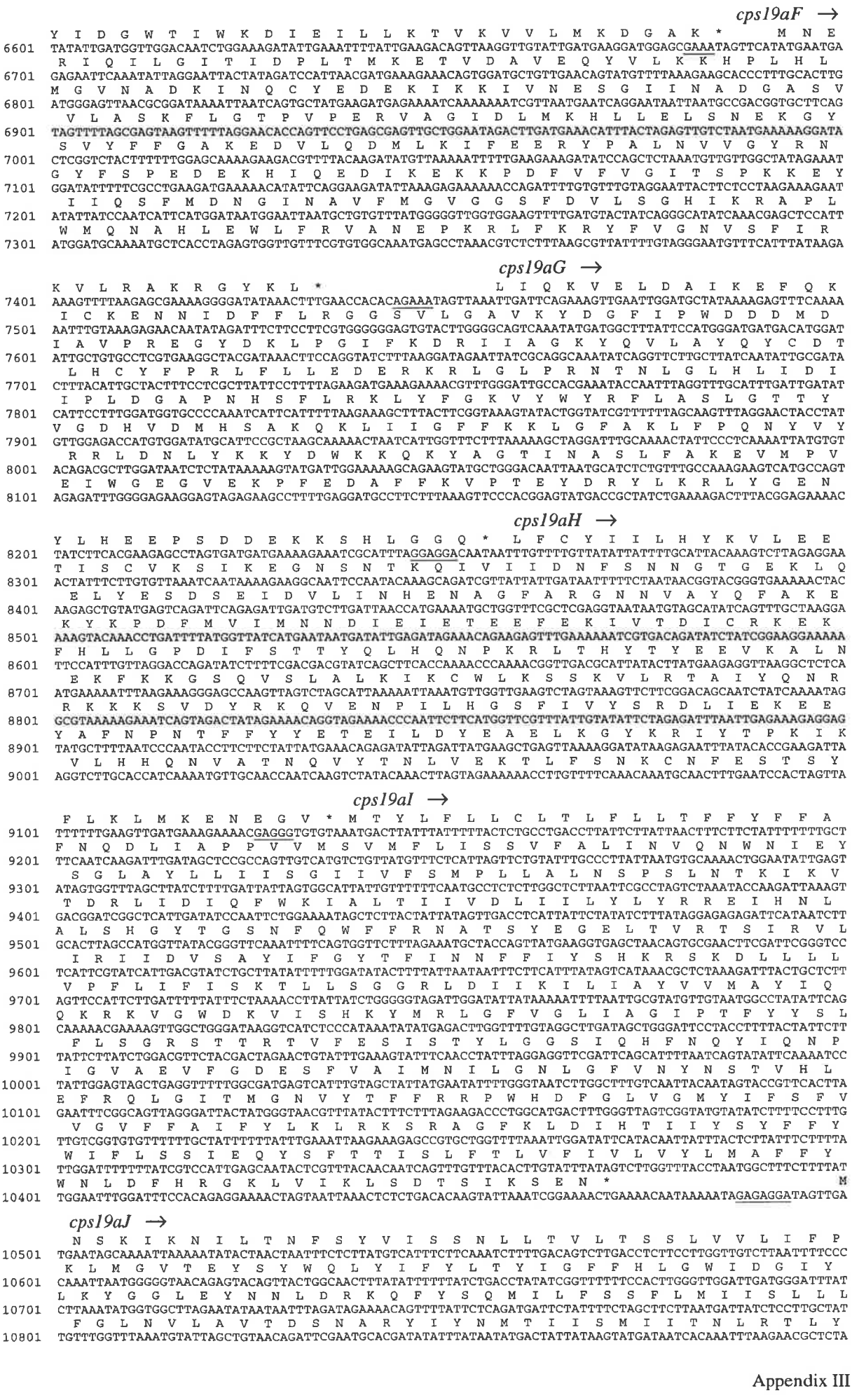
```
      L  I  *        M  K  E  Q  N  T  I  E  I  D  V  F  Q  L  L  K  T  L  W  K  H  K  L  I  I  L  L
3901  AACTAATTTAGGAGAAATGATGAAAGAACAAAACACGATAGAAATCGATGTATTTCAATTACTTAAAACCTTGTGGAAACACAAGCTAATAATTTTATTA
      V  A  L  V  T  G  A  G  A  F  A  Y  S  T  F  I  V  K  P  E  Y  T  S  T  T  R  I  Y  V  V  N  R  N  Q
4001  GTGGCACTTGTGACAGGGGCGGGAGCTTTTGCATATAGCACTTTTATTGTTAAGCCAGAATATACAAGCACCACGCGTATTTACGTAGTCAACCGTAATC
      G  D  K  P  G  L  T  N  Q  D  L  Q  A  G  T  Y  L  V  K  D  Y  R  E  I  I  L  S  Q  D  V  L  E  K
4101  AAGGAGATAAGCCGGGACTGACGAATCAGGACTTGCAGGCAGGAACTTATCTGGTAAAAGACTACCGCGAAATTATCCTTTCGCAAGATGTATTGGAAAA
      V  A  T  N  L  K  L  D  M  P  A  K  T  L  T  S  K  V  Q  V  T  V  P  A  D  T  R  I  V  S  I  S  V
4201  GGTAGCGACAAATTTGAAGTTGGATATGCCAGCAAAAACGTTAACTAGCAAAGTGCAAGTGACTGTACCAGCCGACACTCGTATCGTCTCAATCTCTGTC
      K  D  K  Q  P  E  E  A  S  R  I  A  N  S  L  R  E  V  A  A  E  K  I  I  A  V  T  R  V  S  D  V  T  T
4301  AAGGATAAACAACCAGAGGAAGCCAGTCGCATTGCTAATTCTCTACGAGAAGTTGCTGCAGAAAAGATCATCGCTGTAACGCGAGTATCTGATGTAACGA
      L  E  E  A  R  P  A  M  T  P  S  S  P  N  V  R  R  N  T  L  V  G  F  L  G  G  A  A  V  T  V  I  T
4401  CGCTTGAAGAAGCGCGACCAGCTATGACTCCCTCTTCTCCAAATGTTCGACGCAACACCTTAGTTGGTTTTCTTGGAGGAGCCGCCGTAACAGTAATTAC
      V  L  L  I  E  L  L  D  T  R  V  K  R  P  E  D  V  E  D  V  L  Q  M  P  L  L  G  V  V  P  D  F  D
4501  TGTTCTTTTGATTGAGTTGCTCGATACCCGTGTAAAACGTCCTGAAGATGTTGAAGATGTACTGCAAATGCCACTTCTAGGAGTCGTTCCAGATTTTGAT
```

cps19aD →

```
      K  M  K  *        M  P  T  L  E  I  S  Q  A  K  L  E  L  V  K  K  A  E  E  Y  Y  N  A  L  C  T  N
4601  AAAATGAAATAGGAGGAAGTTATGCCAACGTTAGAAATCTCACAGGCGAAATTAGAACTTGTGAAAAAGGCAGAGGAATATTATAATGCTTTGTGCACGA
      L  Q  L  S  G  D  D  L  K  V  F  S  I  T  S  V  K  A  G  E  G  K  T  T  T  S  T  N  I  A  W  A  F
4701  ACCTACAGTTAAGTGGAGATGATTTGAAAGTATTTTCTATCACTTCTGTGAAAGCAGGAGAAGGAAAAACAACGACTTCCACCAATATCGCTTGGGCTTT
      A  R  A  G  Y  K  T  L  L  I  D  A  D  M  R  N  S  V  M  S  G  V  F  K  S  R  E  I  T  G  L  T
4801  TGCGCGTGCAGGTTACAAAACGCTTCTGATTGATGCAGACATGCGTAACTCCGTTATGTCAGGTGTCTTTAAATCAAGGGAAAGGATTACTGGACTAACA
      E  F  L  S  G  T  T  D  L  S  Q  G  L  C  D  T  N  I  E  N  L  F  V  I  Q  A  G  S  V  S  P  N  P  T
4901  GAATTTCTATCAGGGACAACAGACCTGTCACAGGGGCTTTGCGATACCAATATTGAGAATCTCTTTGTGATTCAGGCTGGCTCTGTGTCACCAAATCCGA
      A  L  L  Q  S  K  N  F  T  T  M  L  E  T  L  R  K  Y  F  D  Y  I  V  V  D  T  A  P  V  G  V  V  I
5001  CAGCCCTTCTTCAAAGTAAGAATTTCACTACAATGCTTGAAACCTTGCGTAAATATTTTGACTATATCGTCGTAGACACTGCCCCTGTCGGTGTCGTGAT
      D  A  A  I  I  M  Q  K  C  D  A  S  I  L  V  T  K  A  G  E  T  K  R  R  D  L  Q  K  A  K  E  Q  L
5101  TGATGCGGCTATCATTATGCAGAAATGTGATGCTTCTATTTTAGTGACGAAGGCAGGTGAAACAAAACGACGGGATCTTCAAAAAGCGAAAGAACAGTTG
      E  Q  T  G  K  S  C  L  G  V  V  L  N  K  L  D  T  S  V  E  K  Y  G  S  Y  G  S  Y  G  N  Y  R  K  Q
5201  GAACAAACTGGGAAGTCGTGTTTAGGAGTTGTGTTGAATAAACTCGATACTTCAGTAGAGAAATACGGTTCTTATGGGAGTTATGGGAATTACAGGAAAC
```

cps19aE →

```
      K  K  *        M  S  E  K  L  A  K  P  S  V  A  I  V  Q  S  F  L  V  I  L  L  T  Y  L  L
5301  AGAAAAAATAGGTTGAGGGATAGAAATGAGTGAAAAACTAGCAAAGCCTTCAGTAGCCATAGTCCAGAGTTTTCTTGTTATTTTATTGACTTATCTACTT
      S  A  V  R  E  T  E  I  V  S  T  T  A  I  V  L  Y  I  L  H  Y  F  V  F  Y  I  S  D  Y  G  R  N  F  F
5401  AGTGCTGTGAGAGAAACGGAGATTGTTTCAACAACAGCTATTGTACTTTATATCCTCCACTATTTTGTCTTTTATATCAGTGATTATGGACGGAATTTCT
      K  R  G  Y  L  I  E  L  V  Q  T  L  K  Y  I  L  F  F  A  L  A  I  S  I  S  N  F  F  L  E  D  R  F
5501  TTAAAAGGGGATATTTGATTGAACTTGTCCAGACATTGAAATATATCCTATTCTTTGCGCTAGCGATTAGTATTTCTAATTTTTTCTTAGAAGATCGATT
      S  I  S  R  R  G  M  I  Y  F  L  L  L  H  V  F  L  V  Y  V  L  N  P  F  I  K  W  Y  W  K  R  A  Y
5601  TAGTATTTCCAGAAGAGGCATGATTTACTTCCTCTTATTACATGTTTTCTTAGTCTATGTGCTAAACCCATTTATCAAGTGGTATTGGAAGCGGGCTTAT
      P  N  F  K  G  S  K  K  I  L  L  L  T  A  T  S  R  V  E  K  V  L  D  R  L  I  E  S  D  D  V  V  G  E
5701  CCCAACTTTAAAGGAAGTAAGAAGATTCTCCTACTTACAGCAACTTCTCGTGTTGAAAAGGTACTGGATAGACTAATAGAATCAGATGATGTTGTTGGGG
      L  V  A  V  S  V  L  D  K  P  D  F  Q  H  D  D  L  K  V  V  A  E  G  E  I  V  N  F  A  T  H  E  V
5801  AGTTGGTAGCCGTCAGTGTTTTAGATAAACCAGATTTTCAGCATGATGATTTAAAGGTTGTAGCAGAGGGGGAGATAGTAAATTTTGCGACTCATGAGGT
      V  D  E  V  F  I  N  L  P  S  E  K  Y  N  I  G  E  L  V  S  Q  F  E  T  M  G  I  D  V  T  V  N  L
5901  GGTCGATGAAGTCTTTATCAATCTTCCAAGTGAAAAATACAATATTGGAGAGCTTGTCTCTCAGTTTGAAACGATGGGAATTGATGTAACAGTCAATCTA
      N  A  F  D  W  A  R  N  K  Q  I  C  E  M  A  G  L  N  V  V  T  F  S  T  T  F  Y  K  T  S  H  V  I  A
6001  AATGCTTTCGATTGGGCACGTAACAAGCAAATTTGTGAGATGGCAGGACTAAATGTTGTGACTTTTTCTACAACATTTTATAAGACTAGCCATGTGATTG
      K  R  I  I  D  I  V  G  S  L  V  G  L  I  L  C  G  L  V  S  I  V  L  V  P  L  I  R  K  D  G  G  S
6101  CTAAGCGGATTATTGATATTGTCGGTTCCCTGGTAGGTTTGATATTATGTGGTCTAGTCAGTATTGTGCTGGTTCCTTTGATTCGAAAGGATGGGGGTTC
      A  F  F  A  Q  T  R  I  G  K  N  G  R  H  F  T  F  Y  K  F  R  S  M  C  V  D  A  E  E  K  K  R  E
6201  TGCTTTTTTTGCTCAGACGCGTATAGGAAAAAATGGCCGCCATTTACTTTTTATAAGTTTCGCTCTATGTGTGGATGCTGAGGAGAAAAAAGAGAA
      L  M  E  Q  N  T  M  Q  G  G  M  F  K  V  D  D  P  R  I  T  K  I  G  H  F  I  R  K  T  S  L  D  E
6301  CTCATGGAACAAAATACCATGCAAGGTGGAATGTTTAAGGTGGACGATGATCCTCGTATCACGAAAATTGGTCATTTTATACGGAAGACGAGCTTGGACG
      L  P  Q  F  Y  N  V  L  K  G  D  M  S  L  V  G  T  R  P  P  T  V  D  E  Y  E  H  Y  T  P  E  Q  K
6401  AGCTACCACAGTTTTATAATGTTCTAAAGGGAGATATGAGTTTGGTAGGTACACGACCACCAACAGTGGACGAGTATGAGCACTATACCCCAGAACAAAA
      R  R  L  S  F  K  P  G  I  T  G  L  W  Q  V  S  G  R  S  E  I  K  N  F  D  E  V  V  K  L  D  V  A
6501  ACGTCGCCTAAGTTTTAAACCTGGCATAACAGGTTTATGGCAGGTCAGCGGACGAAGTGAAATCAAAAATTTCGATGAAGTTGTCAAATTAGATGTGGCC
```


cps19aF →
cps19aG →
cps19aH →
cps19aI →
cps19aJ →

```
       V  Y  V  L  Q  M  T  N  R  L  K  D  S  S  I  I  L  L  S  D  R  V  L  Y  V  L  L  L  F  L  F  I  V
10901  CGTGTATGTTTTGCAGATGACAAACCGCTTGAAGGATAGTTCCATCATTCTACTCAGTGATCGCGTTTTATATGTGCTCCTTTTATTCCTGTTTATTGTA
       F  G  W  H  E  Y  K  V  M  I  L  A  D  I  L  G  R  S  F  S  L  I  L  S  F  W  I  C  K  D  I  V  F  Q
11001  TTTGGTTGGCATGAATACAAGGTTATGATTTTGGCAGATATTCTAGGTCGATCATTTTCTCTCATACTTTCCTTCTGGATTTGTAAAGATATTGTGTTTC
        P  L  S  K  F  I  F  N  I  K  E  S  F  D  N  I  R  V  G  I  N  L  M  L  S  N  I  A  S  S  L  I  I
11101  AGCCTTTGTCAAAATTCATCTTTAATATAAAAGAGTCTTTTGATAATATCCGAGTTGGTATCAATTTAATGTTATCCAACATCGCAAGTAGTTTGATTAT
        G  I  V  R  L  G  I  Q  W  N  W  N  I  E  T  F  G  K  V  S  L  T  L  S  V  S  N  L  L  M  T  F  I
11201  CGGTATTGTTCGTCTAGGAATCCAATGGAACTGGAATATCGAAACATTCGGGAAAGTATCTCTGACCTTGAGTGTTTCCAATCTATTGATGACTTTCATT
       N  A  I  G  L  V  V  F  P  L  L  R  R  T  K  A  E  N  L  P  K  I  Y  S  N  L  R  N  V  L  M  L  I  M
11301  AATGCCATTGGACTAGTCGTCTTTCCGCTATTAAGACGAACAAAAGCAGAAAACTTACCTAAGATTTATTCTAATTTAAGAAACGTTTTGATGCTTATCA
        F  A  I  L  L  I  Y  Y  P  L  K  I  V  L  D  L  W  L  P  A  Y  Q  D  A  L  I  F  M  T  L  I  F  P
11401  TGTTCGCGATTTTGCTCATTTACTATCCTTTAAAAATTGTATTAGACCTCTGGTTGCCAGCCTATCAAGATGCCTTGATTTTCATGACCCTTATTTTCCC
        M  S  V  Y  E  G  K  M  A  L  V  I  N  T  Y  L  K  A  L  R  M  E  R  D  I  L  R  V  N  A  L  V  M
11501  TATGTCAGTCTATGAAGGGAAAATGGCATTGGTCATTAATACTTACTTAAAGGCATTAAGAATGGAAAGAGATATCCTTAGAGTCAATGCCTTAGTAATG
       L  I  S  M  G  V  T  L  V  T  T  Y  L  L  N  S  L  E  L  T  V  V  S  I  V  V  L  L  A  L  R  S  I  I
11601  TTGATCAGTATGGGAGTGACCCTGGTTACAACATACCTGTTAAATAGTTTGGAGCTGACTGTTGTATCGATAGTTGTTTTGCTAGCTTTGAGAAGTATCA
        A  E  L  I  L  S  K  K  L  D  V  S  V  K  K  D  I  V  L  E  F  L  L  T  L  V  F  I  S  S  S  W  Y
11701  TAGCCGAATTAATTCTGTCTAAAAAACTGGATGTTTCGGTTAAGAAGGATATTGTATTAGAATTTCTTTTGACGCTTGTCTTTATTTCTTCAAGTTGGTA
        L  P  I  G  L  A  V  I  V  Y  T  I  A  Y  G  L  Y  L  Y  L  K  H  E  D  I  K  T  Y  L  A  Y  F  K
11801  TTTGCCGATTGGGCTGGCAGTAATTGTCTACACAATAGCCTACGGTTTATATCTCTATTTGAAACACGAAGATATCAAAACCTATTTAGCTTACTTTAAA
```

cps19aK →

```
       A  S  K  K  T  S  N  *                                                              M  K  K  I
11901  GCTAGTAAAAAAACATCAAATTAAAAAATTATATAATGATTAAGTGGTAGATTCTATTTCTACCGTTTTAGATATTCGGGAGGTAATGATGAAAAAAATT
       M  L  V  F  G  T  R  P  E  A  I  K  M  C  P  L  V  N  E  L  K  K  H  E  D  M  E  T  I  V  C  V  T  G
12001  ATGTTAGTATTCGGTACACGTCCGGAAGCCATCAAAATGTGTCCATTAGTCAATGAGTTGAAAAAACACGAAGATATGGAAACAATTGTGTGTGTTACTG
        Q  H  K  E  M  V  S  P  V  L  D  L  F  G  V  V  P  D  Y  D  L  E  I  M  K  A  N  Q  T  L  F  S  I
12101  GACAACACAAAGAGATGGTTAGTCCTGTTTTAGATTTATTTGGTGTTGTACCAGATTATGATTTAGAAATTATGAAGGCTAACCAAACCTTGTTCTCTAT
        T  T  S  I  L  E  K  I  K  P  V  L  E  K  E  Q  P  D  I  V  L  V  H  G  D  T  T  T  T  Y  A  A  A
12201  CACAACTAGTATCTTGGAAAAGATAAAACCAGTTTTAGAGAAGGAACAACCAGATATTGTCCTAGTTCACGGTGACACTACGACAACTTATGCAGCAGCC
       L  A  A  F  Y  L  G  I  K  V  G  H  V  E  A  G  L  R  T  Y  N  L  Q  S  P  F  P  E  E  F  N  R  Q  S
12301  TTGGCAGCATTCTATTTGGGAATTAAAGTAGGACATGTTGAAGCTGGTTTGCGAACGTACAATTTACAAAGTCCATTTCCTGAAGAATTTAACAGGCAAT
        T  S  I  I  A  T  Y  H  F  A  P  T  E  L  A  K  E  N  L  L  K  E  G  R  E  N  V  Y  V  T  G  N  T
12401  CGACATCAATCATTGCAACTTACCATTTTGCTCCAACTGAGTTGGCTAAAGAAAATCTCTTAAAAGAAGGTAGAGAGAATGTTTATGTGACTGGAAATAC
        V  I  D  A  L  T  T  T  V  Q  E  D  Y  T  H  T  H  L  D  L  N  A  N  N  R  L  I  L  L  T  A  H  R
12501  TGTCATTGATGCTCTTACAACTACTGTTCAAGAGGATTATACACACACTCATTTAGATTTAAACGCTAACAATCGTCTCATCTTATTGACTGCTCATAGA
       R  E  N  L  G  E  P  M  R  H  M  F  R  A  V  K  R  V  L  N  E  Y  E  D  V  K  V  I  Y  P  I  H  K  N
12601  CGCGAAAATCTCGGCGAACCGATGAGACATATGTTTAGAGCAGTTAAACGAGTATTGAATGAATACGAAGATGTTAAAGTCATTTATCCAATTCATAAGA
        P  L  V  R  E  T  A  A  E  I  F  G  D  T  E  R  I  Q  I  I  E  P  L  D  V  L  D  F  H  N  F  M  N
12701  ATCCTTTGGTACGTGAAACAGCTGCAGAGATCTTTGGAGATACAGAACGGATTCAAATTATTGAACCTTTGGATGTTCTTGATTTCCACAACTTCATGAA
        Q  S  Y  M  I  L  T  D  S  G  G  V  Q  E  E  A  P  S  L  G  K  P  V  L  V  M  R  D  T  T  E  R  P
12801  CCAAAGTTATATGATTTTAACAGATTCTGGCGGAGTTCAGGAAGAAGCACCTTCTTTAGGAAAACCTGTATTGGTTATGCGCGATACTACTGAAAGACCT
       E  G  V  A  A  G  T  L  K  L  V  G  T  D  E  E  T  I  Y  Q  N  F  K  L  L  L  D  D  S  G  E  Y  K  K
12901  GAAGGAGTAGCGGCAGGAACCTTGAAATTAGTAGGAACTGACGAAGAAACAATTTATCAAAACTTCAAACTACTTCTAGATGATTCTGGAGAATACAAAA
        M  S  Q  A  S  N  P  Y  G  N  G  D  A  S  Q  Q  I  V  Q  I  L  R  G  I  *
13001  AAATGAGTCAGGCTAGCAATCCATACGGTAATGGTGATGCTAGTCAACAGATTGTTCAGATTTTGCGTGGGATTTAAGAAGATTCTCTAAAAGAGTCTAA
13101  GAGGAATCCACTCCACAAACATAAAACTCTTATGCTATCATTTATGATTGAGATAGACTCATAACTTAATATTTTTATCCTAAGAAAATTCGTTCATATT
```

cps19aL →

```
                                        M  K  G  I  I  L  A  G  G  S  G  T  R  L  Y  P  L  T  R  A  A
13201  CAATATGAATGAAAAACAACGAAGAAAGGTTATTTCATTATGAAAGGTATTATTCTTGCAGGCGGCTCAGGTACCCGCCTGTACCCACTTACTCGGGCTG
        S  K  Q  L  M  P  V  Y  D  K  P  M  I  Y  Y  P  L  S  T  L  M  L  A  G  I  K  D  I  L  I  I  S  T
13301  CGTCAAAACAGCTGATGCCGGTTTATGATAAACCTATGATTTATTATCCGTTGTCGACATTAATGTTGGCTGGAATTAAAGATATTTTGATTATCTCAAC
        P  Q  D  L  P  R  F  K  D  L  L  L  D  G  S  E  F  G  I  K  L  S  Y  A  E  Q  P  S  P  D  G  L  A
13401  TCCTCAAGATTTGCCCCGTTTTAAGGACTTGCTCTTGGATGGTTCCGAATTTGGGATCAAGCTTTCCTATGCGGAACAACCTAGTCCCGATGGACTTGCT
       Q  A  F  L  I  G  E  E  F  I  G  D  D  S  V  A  L  I  L  G  D  N  I  Y  H  G  P  G  L  S  T  M  L  Q
13501  CAGGCTTTTCTTATCGGTGAAGAATTTATCGGTGACGATAGCGTTGCCTTGATTTTAGGCGACAATATCTACCATGGTCCTGGTCTGAGCACAATGCTTC
        K  A  A  K  K  E  K  G  A  T  V  F  G  Y  Q  V  K  D  P  E  R  F  G  V  V  E  F  D  T  D  M  N  A
13601  AAAAAGCAGCCAAGAAAGAGAAAGGTGCGACTGTTTTTGGCTACCAAGTGAAGGATCCAGAGCGTTTTGGTGTGGTTGAGTTTGATACAGACATGAATGC
        I  S  I  E  E  K  P  E  Y  P  R  S  N  Y  A  V  T  G  L  Y  F  Y  D  N  D  V  V  E  I  A  K  Q  I
13701  TATTTCTATAGAAGAAAAACCGGAGTATCCTCGCTCCAACTATGCAGTGACAGGACTGTATTTCTATGATAATGATGTTGTGGAGATTGCTAAACAGATC
       K  P  S  A  R  G  E  L  E  I  T  D  V  N  K  A  Y  L  N  R  G  D  L  S  V  E  L  M  G  R  G  F  A  W
13801  AAACCTAGTGCTCGTGGCGAGTTAGAAATTACAGACGTTAACAAGGCTTACCTAAATCGTGGTGACCTTTCTGTTGAGCTGATGGGGCGTGGTTTTGCCT
        L  D  T  G  T  H  E  S  L  L  E  A  S  Q  Y  I  E  T  V  Q  R  M  Q  N  V  Q  V  A  N  L  E  E  I
13901  GGTTGGATACGGGAACCCATGAAAGCTTGCTAGAAGCTTCTCAGTATATTGAAACGGTTCAGCGTATGCAGAACGTTCAAGTGGCAAATCTAGAAGAAAT
        S  Y  R  M  G  Y  I  S  R  E  D  V  L  E  L  A  Q  L  L  K  K  N  E  Y  G  Q  Y  L  L  R  L  I  G
14001  TTCCTATCGTATGGGCTACATTAGTCGTGAAGACGTGCTGGAATTGGCTCAGCTTCTTAAGAAAAATGAATACGGACAATATTTGCTCCGTTTGATTGGA
```

cps19aM →

```
       E  A  *  M  S  D  N  F  F  G  K  T  L  V  V  R  K  I  D  A  I  P  G  L  L  E  F  D  I  P  V  H  G  D
14101  GAAGCATAGATGTCAGATAATTTTTTTGGAAAGACACTTGTGGTACGCAAGATTGATGCTATACCAGGACTGCTAGAGTTTGATATTCCCGTTCATGGAG
        N  R  G  W  F  K  E  N  F  Q  K  E  K  M  L  P  L  G  F  P  E  S  F  F  A  A  G  K  L  Q  N  N  V
14201  ACAATCGTGGTTGGTTTAAGGAAAATTTCCAGAAGGAAAAGATGCTACCGCTTGGTTTTCCTGAAAGCTTCTTTGCTGCAGGGAAACTGCAAAATAACGT
        S  F  S  R  K  N  V  L  R  G  L  H  A  E  P  W  D  K  Y  I  S  V  A  D  D  G  K  V  L  G  S  W  V
14301  CAGCTTTTCTCGCAAAAAATGTTCTTCGAGGACTCCATGCTGAACCTTGGGACAAGTATATCTCTGTTGCAGACGATGGGAAGGTTTTAGGATCTTGGGTA
       D  L  R  E  G  E  T  F  G  N  V  Y  Q  T  E  I  D  A  S  K  G  I  F  V  P  R  G  V  A  N  G  F  Q  V
14401  GATTTGCGTGAAGGCGAGACCTTTGGAAATGTTTACCAGACAGAGATTGATGCAAGCAAGGGAATCTTTGTTCCTCGAGGCGTAGCTAATGGCTTCCAAG
        L  S  D  T  V  S  Y  S  Y  L  V  N  D  Y  W  A  L  E  L  K  P  K  Y  A  F  V  N  Y  A  D  P  S  L
14501  TTCTATCAGATACAGTGTCATATAGCTATCTGGTCAATGATTACTGGGCGCTTGAACTCAAACCCAAGTATGCCTTTGTGAACTACGCTGATCCAAGCCT
        G  I  E  W  E  N  I  A  E  A  E  V  S  E  A  D  K  N  H  P  L  L  K  D  V  K  P  L  K  K  E  D  L
14601  TGGTATTGAATGGGAAAATATTGCAGAAGCAGAGGTTTCAGAAGCAGATAAAAATCATCCACTACTTAAGGATGTAAAACCTTTGAAAAAAGAAGATTTG
```

cps19aN →

```
       E  *              M  T  E  Y  K  N  I  I  V  T  G  G  A  G  F  I  G  S  N  F  V  H  Y  V  Y  E  N  F  P
14701  GAATAAGGAAAGAATATGACTGAATACAAAAATATTATCGTGACAGGTGGAGCTGGCTTTATCGGTTCTAACTTTGTCCATTATGTTTACGAGAACTTTC
        D  V  H  V  T  V  L  D  K  L  T  Y  A  G  N  R  A  N  I  E  E  I  L  G  N  R  V  E  L  V  V  G  D
14801  CAGATGTTCATGTGACAGTCCTAGATAAGTTGACTTATGCTGGAAACCGCGCGAATATTGAGGAAATTTTAGGTAATCGTGTTGAGTTAGTTGTTGGTGA
        I  A  D  A  E  L  V  D  K  L  A  A  Q  A  D  A  I  V  H  Y  A  A  E  S  H  N  D  N  S  L  N  D  P
14901  CATTGCTGATGCGGAGTTGGTAGACAAGTTGGCTGCTCAAGCAGATGCTATCGTTCATTATGCAGCGGAAAGCCACAATGATAATTCGCTCAATGATCCA
       S  P  F  I  H  T  N  F  I  G  T  Y  T  L  L  E  A  A  R  K  Y  D  I  R  F  H  H  V  S  T  D  E  V  Y
15001  TCGCCATTTATTCATACTAACTTCATTGGAACCTATACTCTTTTAGAAGCTGCTCGTAAGTATGATATTCGCTTCCACCATGTATCGACAGATGAAGTTT
        G  D  L  P  L  R  E  D  L  P  G  H  G  E  G  P  G  E  K  F  T  A  E  T  K  Y  N  P  S  S  P  Y  S
15101  ATGGGGATCTCCCTTTACGCGAAGATTTGCCAGGTCATGGAGAAGGGCCGGGTGAGAAATTTACGGCTGAAACCAAGTACAATCCAAGCTCGCCTTACTC
        S  T  K  A  A  S  D  L  I  V  K  A  W  V  R  S  F  G  V  K  A  T  I  S  N  C  S  N  N  Y  G  P  Y
15201  ATCAACCAAGGCAGCCTCAGATTTGATTGTCAAAGCCTGGGTGCGTTCTTTTGGAGTCAAGGCAACGATTTCCAACTGTTCAAATAACTACGGTCCTTAT
       Q  H  I  E  K  F  I  P  R  Q  I  T  N  I  L  S  G  I  K  P  K  L  Y  G  E  G  K  N  V  R  D  W  I  H
15301  CAACATATCGAAAAATTCATCCCACGTCAGATTACTAACATCCTAAGTGGAATTAAGCCAAAACTTTACGGTGAAGGTAAGAATGTTCGTGACTGGATTC
```

```
        T  N  D  H  S  S  G  V  W  T  I  L  T  K  G  Q  I  G  E  T  Y  L  I  G  A  D  G  E  K  N  N  K  E
15401   ATACCAATGACCATTCTTCAGGAGTTTGGACAATCTTGACAAAAGGGCAAATCGGTGAAACCTACTTGATTGGGGCTGATGGTGAGAAGAACAATAAGGA
          V  L  E  L  I  L  K  E  M  G  Q  A  T  D  A  Y  D  H  V  T  D  R  A  G  H  D  L  R  Y  A  I  D  A
15501   AGTTTTGGAACTTATCCTTAAGGAAATGGGACAAGCTACGGATGCCTATGATCATGTGACTGACCGTGCAGGACATGACCTTCGCTATGCGATTGATGCC
        S  K  L  R  E  E  L  G  W  K  P  E  F  T  N  F  E  A  G  L  K  A  T  I  K  W  Y  T  D  N  Q  E  W  W
15601   AGCAAGCTCCGTGAGGAGTTGGGGTGGAAACCTGAATTTACCAACTTTGAAGCTGGGCTCAAGGCAACAATCAAGTGGTATACAGATAACCAAGAATGGT
          K  A  E  K  E  A  V  E  A  N  Y  A  K  T  Q  E  I  I  T  V  *
15701   GGAAAGCAGAAAAAGAAGCTGTTGAAGCCAATTATGCTAAGACTCAGGAGATTATTACAGTATAAAAAGCAGGAAATAGCTGCTTTTTATTGCTATATTG

                                                cps19aO  →
                                                        M  I  L  I  T  G  A  N  G  Q  L  G  T  E  L  R  Y  L  L  D  E  R  N  E
15801   GGAAGAGTTACATATTAGAAAGGTCTAGAGATGATTTTAATTACAGGGGCAAATGGCCAATTAGGAACGGAACTTCGCTATTTATTGGATGAACGTAATG
          E  Y  V  A  V  D  V  A  E  M  D  I  T  D  A  E  M  V  E  K  V  F  E  E  V  K  P  T  L  V  Y  H  C
15901   AAGAATACGTGGCAGTAGATGTGGCTGAGATGGACATTACCGATGCAGAAATGGTTGAGAAAGTTTTTGAAGAGGTGAAACCGACTTTAGTCTACCACTG
          A  A  Y  T  A  V  D  A  A  E  D  E  G  R  E  L  D  F  A  I  N  V  T  G  T  K  N  V  A  K  A  S  E
16001   TGCAGCCTACACCGCTGTTGATGCAGCAGAGGATGAAGGAAGAGAGTTGGACTTCGCCATCAATGTGACGGGGACAAAAAATGTCGCAAAAGCATCTGAA
        K  H  G  A  T  L  V  Y  I  S  T  D  Y  V  F  D  G  K  K  P  V  G  Q  E  W  E  V  D  D  R  P  D  P  Q
16101   AAGCATGGTGCAACTCTAGTTTATATTTCTACGGACTATGTCTTTGATGGTAAGAAACCAGTTGGACAAGAGTGGGAAGTTGATGACCGACCAGATCCAC
          T  E  Y  G  R  T  K  R  M  G  E  E  L  V  E  K  H  V  S  N  F  Y  I  I  R  T  A  W  V  F  G  N  Y
16201   AGACAGAATATGGCCGTACTAAGCGTATGGGGGAAGAGTTAGTTGAGAAGCATGTGTCTAATTTCTATATTATCCGTACTGCCTGGGTATTTGGAAATTA
          G  K  N  F  V  F  T  M  Q  N  L  A  K  T  H  K  T  L  T  V  V  N  D  Q  Y  G  R  P  T  W  T  R  T
16301   TGGCAAAAACTTCGTTTTTACCATGCAAAATCTTGCGAAAACTCATAAGACTTTAACAGTTGTAAATGACCAGTACGGTCGTCCGACTTGGACTCGTACC
        L  A  E  F  M  T  Y  L  A  E  N  R  K  E  F  G  Y  Y  H  L  S  N  D  A  T  E  D  T  T  W  Y  D  F  A
16401   TTGGCTGAGTTCATGACCTACCTAGCTGAAAATCGTAAGGAATTTGGTTATTATCATTTGTCAAATGATGCGACAGAAGACACAACATGGTATGATTTTG
          V  E  I  L  K  D  T  D  V  E  V  K  P  V  D  S  S  Q  F  P  A  K  A  K  R  P  L  N  S  T  M  S  L
16501   CAGTTGAAATTTTGAAAGATACAGATGTCGAAGTCAAGCCAGTAGATTCCAGTCAATTTCCAGCCAAAGCTAAACGTCCGCTAAACTCAACGATGAGCCT
          A  K  A  K  A  T  G  F  V  I  P  T  W  Q  D  A  L  Q  E  F  Y  K  Q  E  V  R  *
16601   GGCCAAAGCCAAAGCTACTGGATTTGTTATTCCAACTTGGCAAGATGCATTGCAAGAATTTTACAAACAAGAAGTGAGATAAGTAGTAGAATGATCTTCT
16701   AGTCTAATAAAAGAGGCAGATAATGAACTCCAAAGGAGCTTAAGATGTACGATTATCTTGTTGTTGGTGCTGGTCTCTTTGGCGCATAGCTTTGGCTCAG
16801   CTTCTATTATCGCTCACACCATCCATCAGAAGTTTAATCTGAAGGTACCCAATTATCGCCAAGAAGAAGATTGGGCTAGGATGGGTTTACCAATCACACG
16901   TAAGGAAATCTCTAATTGGCATATCAAGGCAAGTCAATACTATTTAGAGTCCCTTTATAACCTTTTACGAGAAAAGTTGTTAGAACAACCTCTTCTTCAT
17001   GCGGATGAAACCTCTTATCGGGTGCTAGAGAGTGATAGCCATCTGACCTACTATTGGACCTTTTTGTCTGGGAAAGCTGAGAATCAAGCAATCACGCTGT
17101   ACCATCATGATCAGCGTCGGAGTGGTTTAGTAGTACAAGAATTCCTAGGAGATTATTCTGGCTATGTGCATTGTGATATGTTGCGGCAGTAACTTAGGAC
17201   TTTAGTCCTCTAGTTCTGCCTATGCGATAGCAGTCCAAGGTTTAGGAGCAAGGCGACGCTAAGCTTGGTAAACTACGAACCGCTAGAAGCTTATCGTCAA
17301   CTGGAAAGAAGCTGAACTTGTTGGATGTTGGGCGCATGTGAAGAAGGAAGTTTTTTGAAGCGCCCCCAAGCAAGCGGATAAATCATCGTTAGGAGCTAAA
17401   GGTTTAGCTTATTGTGATCAGTTATTTGCCTTGGAAAGAGACTGGGAGGCTTTGCTAGCTGATGAACGACTACAGAAACGTCAAGAAGAGCTCCAACCCC
17501   TAATGGAAGATTTCTTTGCTTGGTGCCGGCGTCAGTCAGTTTTATCGGGTTCAAAACTAGGAAGGGCAATTGAATACAGCCTCAAGTATAAAGAAACCTT
17601   TAAGACCATTTTAAAAGACGGACATCTGGTCCTTTCCAATAATCTAGCTGAACGCGCCATTAAATCATTGGTTATGGGACGGAGTAAAAGAGTCCAGTGG
17701   ACTCTTTTAGCCTAAGCTCAGTTTAAAAAAGTGAGGGTGGTTATTTTCTCAAAATTTTGAAGGAGCTAAAGCAAGAGCTATTATTATGAGCTTATTGGAA
17801   ACAGCTAAACGTCATCAACTAAATAGCGAGAAATATCTATCCTATCTTCTAGAATGTCTTCCAAACGAGGAAACTCTCGTAAACAAAGAGGTTTTAGAGG
17901   CTTATTTACCATGTACTAAAGTTGTACAAGAAAAGTGCAAATAAGAAATCTCCAGATTAGGAACTATATATGAGTTCTCTAGTCTGGAGATTTTTCAATA
18001   TACTTCGTTATTGGACGGTTACGATATTCATATTTTTTGCAAAGATGTTGTTTGAAAAATAATTTTCAAAAATTCTGAAAATTCTGTTGACAACTTTCTG

                                                aliA  →
                                                        M  K  S  S  R  L  F  A  L  A  G  V  T  L  L  A  A  T  T  L
18101   AAAAGAGTCTATAATGGAGAGAAAGTTTTAAAGGAGAAAATGATGAAAAGTTCAAGACTATTTGCCCTTGCGGGCGTGACATTATTGGCGGCGACTACTT
          A  A  C  S  G  S  G  S  S  T  K  G  E  K  T  F  S  Y  I  Y  E  T  D  P  D  N  L  N  Y  L  T  T  A
18201   TAGCTGCATGCTCTGGATCAGGTTCAAGCACTAAAGGTGAGAAAACATTCTCATACATTTATGAGACAGACCCTGATAACCTCAACTATTTGACAACTGC
          K  A  A  T  A  N  I  T  S  N  V  V  D  G  L  L  E  N  D  R  Y  G  N  F  V  P  S  M  A  E  D  W  S
18301   TAAGGCTGCGACAGCAAATATTACCAGTAACGTGGTTGATGGTTTGCTAGAAAATGATCGCTACGGGAACTTTGTGCCGTCTATGGCTGAGGATTGGTCT
        V  S  K  D  G  L  T  Y  T  Y  T  I  R  K  D  A  K  W  Y  T  S  E  G  E  E  Y  A  A  V  K  A  Q  D  F
18401   GTATCCAAGGATGGATTGACTTACACTTATACTATCCGTAAGGATGCAAAATGGTATACTTCTGAAGGTGAAGAATACGCGGCAGTCAAAGCTCAAGACT
          V  T  G  L  K  Y  A  A  D  K  K  S  D  A  L  Y  L  V  Q  E  S  I  K  G  V  D  A  Y  V  K  G  E  I
18501   TTGTAACAGGACTAAAATATGCTGCTGATAAAAAATCAGATGCTCTTTACCTTGTTCAAGAATCAATCAAAGGGGTGGATGCCTATGTAAAAGGGGAAAT
          K  D  F  S  Q  V  G  I  K  A  L  D  D  Q  T  V  Q  Y  T  L  N  K  P  E  S  F  W  N  S  K  T  T  M
18601   CAAAGATTTCTCACAAGTAGGAATTAAGGCTTTGGATGATCAGACAGTTCAGTACACTTTGAACAAACCAGAAAGTTTTTGGAACTCAAAAACAACCATG
        G  V  L  A  P  V  N  E  E  L  L  N  S  K  G  D  D  S
18701   GGTGTGCTTGCGCCAGTTAATGAAGAGTTGTTGAACTCAAAAGGGGATGATTCT
```

APPENDIX IV

The nucleotide and amino acid sequence of the *cps19a* locus from *cps19aJ-aliA* from *S. pneumoniae* strain 19A2.

The nucleotide sequence is numbered in accordance with GenBank accession number AF105113 and starts from the centre of *cps19aJ* to the beginning of *aliA*. This sequence has been referred to as *cps19a*$_2$ throughout this thesis to distinguish it from the *cps19a* locus from *S. pneumoniae* strain 19A1. The amino acid translation for each ORF is represented by single letter code above the first nucleotide of each codon, except when translation is in the complimentary DNA strand, when it is shown below the nucleotide sequence. Possible ribosome binding sites are underlined. The promoters for *aliA* and *cps19a*$_2$*O* are indicated and the stem-loop termination sequence is indicated with arrows.

```
      ...cps19a2J→
        S  K  F  I  F  N  I  K  E  S  F  D  N  I  R  V  G  I  N  L  M  L  S  N  I  A  S  S  L  I  I  G  I
1     TGTCAAAATTCATCTTTAATATAAAAGAGTCTTTTGATAATATCCGAGTTGGTATCAATTTAATGTTATCCAACATCGCAAGTAGTTTGATTATCGGTAT

        V  R  L  G  I  Q  W  N  W  N  I  E  T  F  G  K  V  S  L  T  L  S  V  S  N  L  L  M  T  F  I  N  A
101   TGTTCGTCTAGGAATCCAATGGAACTGGAATATCGAAACATTCGGGAAAGTATCTCTGACCTTGAGTGTTTCCAATCTATTGATGACTTTCATTAATGCC

        I  G  L  V  V  F  P  L  L  R  R  T  K  A  E  N  L  P  K  I  Y  S  N  L  R  N  V  L  M  L  I  M  F  A
201   ATTGGACTAGTCGTCTTTCCGCTATTAAGACGAACAAAAGCAGAAAACTTACCTAAGATTTATTCTAATTTAAGAAACGTTTTGATGCTTATCATGTTCG

        I  L  L  I  Y  Y  P  L  K  I  V  L  D  L  W  L  P  A  Y  Q  D  A  L  I  F  M  T  L  I  F  P  M  S
301   CGATTTTGCTCATTTACTATCCTTTAAAAATTGTATTAGACCTCTGGTTGCCAGCCTATCAAGATGCCTTGATTTTCATGACCCTTATTTTCCCTATGTC

        V  Y  E  G  K  M  A  L  V  I  N  T  Y  L  K  A  L  R  M  E  R  D  I  L  R  V  N  A  L  V  M  L  I
401   AGTCTATGAAGGGAAAATGGCATTGGTCATTAATACTTACTTAAAGGCATTAAGAATGGAAAGAGATATCCTTAGAGTCAATGCCTTAGTAATGTTGATC

        S  M  G  V  T  L  V  T  T  Y  L  L  N  S  L  E  L  T  V  V  S  I  V  V  L  L  A  L  R  S  I  I  A  E
501   AGTATGGGAGTGACCCTGGTTACAACATACCTGTTAAATAGTTTGGAGCTGACTGTTGTATCGATAGTTGTTTTGCTAGCTTTGAGAAGTATCATAGCCG

        L  I  L  S  K  K  L  D  V  S  V  K  K  D  I  V  L  E  F  L  L  T  L  V  F  I  S  S  S  W  Y  L  P
601   AATTAATTCTGTCTAAAAAACTGGATGTTTCGGTTAAGAAGGATATTGTATTAGAATTTCTTTTGACGCTTGTCTTTATTTCTTCAAGTTGGTATTTGCC

        I  G  L  A  V  I  V  Y  T  I  A  Y  G  L  Y  L  Y  L  K  H  E  D  I  K  T  Y  L  A  Y  F  K  A  S
701   GATTGGGCTGGCAGTAATTGTCTACACAATAGCCTACGGTTTATATCTCTATTTGAAACACGAAGATATCAAAACCTATTTAGCTTACTTTAAAGCTAGT

                                                                                                  cps19a2K→
      K  K  T  S  N  *                                                                          M  K  K  I  M  L
801   AAAAAAACATCAAATTAAAAAATTATATAATGATTAAGTGGTAGATTCTATTTCTACCGTTTTAGATATTCGGGAGGTAATGATGAAAAAAATTATGTTA

        V  F  G  T  R  P  E  A  I  K  M  C  P  L  V  N  E  L  K  K  H  E  D  M  E  T  I  V  C  V  T  G  Q  H
901   GTATTCGGTACACGTCCGGAAGCCATCAAAATGTGTCCATTAGTCAATGAGTTGAAAAAACACGAAGATATGGAAACAATTGTGTGTGTTACTGGACAAC
```

```
      K  E  M  V  S  P  V  L  D  L  F  G  V  V  P  D  Y  D  L  E  I  M  K  A  N  Q  T  L  F  S  I  T  T
1001  ACAAAGAGATGGTTAGTCCTGTTTTAGATTTATTTGGTGTTGTACCAGATTATGATTTAGAAATTATGAAGGCTAACCAAACCTTGTTCTCTATCACAAC

      S  I  L  E  K  I  K  P  V  L  E  K  E  Q  P  D  I  V  L  I  H  G  D  T  T  T  T  Y  A  A  A  L  A
1101  TAGTATCTTGGAAAAGATAAAACCAGTTTTAGAGAAGGAACAACCAGATATTGTCCTAATTCACGGTGACACTACGACAACTTATGCAGCAGCCTTGGCA

      A  F  Y  L  G  I  K  V  G  H  V  E  A  G  L  R  T  Y  N  L  Q  S  P  F  P  E  E  F  N  R  Q  S  T  S
1201  GCATTCTATTTGGGAATTAAAGTAGGACATGTTGAAGCTGGTTTGCGAACGTACAATTTACAAAGTCCATTTCCTGAAGAATTTAACAGGCAATCGACAT

      I  I  A  T  Y  H  F  A  P  T  E  L  A  K  E  N  L  L  K  E  G  R  E  N  V  Y  V  T  G  N  T  V  I
1301  CAATCATTGCAACTTACCATTTTGCTCCAACTGAGTTGGCTAAAGAAAATCTCTTAAAAGAAGGTAGAGAGAATGTTTATGTGACTGGAAATACTGTCAT

      D  A  L  T  T  T  V  Q  E  D  Y  T  H  T  H  L  D  L  N  A  N  N  R  L  I  L  L  T  A  H  R  R  E
1401  TGATGCTCTTACAACTACTGTTCAAGAGGATTATACACACACTCATTTAGATTTAAACGCTAACAATCGTCTCATCTTATTGACTGCTCATAGACGCGAA

      N  L  G  E  P  M  R  H  M  F  R  A  V  K  R  V  L  N  E  Y  E  D  V  K  V  I  Y  P  I  H  K  N  P  L
1501  AATCTCGGCGAACCGATGAGACATATGTTTAGAGCAGTTAAACGAGTATTGAATGAATACGAAGATGTTAAAGTCATTTATCCAATTCATAAGAATCCTT

      V  R  E  T  A  A  E  I  F  G  D  T  E  R  I  Q  I  I  E  P  L  D  V  L  D  F  H  N  F  M  N  Q  S
1601  TGGTACGTGAAACAGCTGCAGAGATCTTTGGAGATACAGAACGGATTCAAATTATTGAACCTTTGGATGTTCTTGATTTCCACAACTTCATGAACCAAAG

      Y  M  I  L  T  D  S  G  G  V  Q  E  E  A  P  S  L  G  K  P  V  L  V  M  R  D  T  T  E  R  P  E  G
1701  TTATATGATTTTAACAGATTCTGGCGGAGTTCAGGAAGAAGCACCTTCTTTAGGAAAACCTGTATTGGTTATGCGCGATACTACTGAAAGACCTGAAGGA

      V  A  A  G  T  L  K  L  V  G  T  D  E  E  T  I  Y  Q  N  F  K  L  L  L  D  D  S  G  E  Y  K  K  M  S
1801  GTAGCGGCAGGAACCTTGAAATTAGTAGGAACTGACGAAGAAACAATTTATCAAAACTTCAAACTACTTCTAGATGATTCTGGAGAATACAAAAAAATGA

      Q  A  S  N  P  Y  G  N  G  D  A  S  Q  Q  I  V  Q  I  L  R  G  I  *
1901  GTCAGGCTAGCAATCCATACGGTAATGGTGATGCTAGTCAACAGATTGTTCAGATTTTGCGTGGGATTTAAGAAGATTCTCTAAAAGAGTCTAAGAGGAA

2001  TCCACTCCACAAACATAAAACTCTTATGCTATCATTTATGATTGAGATAGACTCATAACTTAATATTTTTATCCTAAGAAAATTCGTTCATATTCAATAT
```

cps19a$_2$L→

```
                                  M  K  G  I  I  L  A  G  G  S  G  T  R  L  Y  P  L  T  R  A  A  S  K
2101  GAATGAAAAACAACGAAGAAAGGTTATTTCATTATGAAAGGTATTATTCTTGCAGGCGGCTCAGGTACCCGCCTGTACCCACTTACTCGGGCTGCGTCAA

      Q  L  M  P  V  Y  D  K  P  M  I  Y  Y  P  L  S  T  L  M  L  A  G  I  K  D  I  L  I  I  S  T  P  Q
2201  AACAGCTGATGCCGGTTTATGATAAACCTATGATTTATTATCCGTTGTCGACATTAATGTTGGCTGGAATTAAAGATATTTTGATTATCTCAACTCCTCA

      D  L  P  R  F  K  D  L  L  L  D  G  S  E  F  G  I  K  L  S  Y  A  E  Q  P  S  P  D  G  L  A  Q  A
2301  AGATTTGCCCCGTTTTAAGGACTTGCTCTTGGATGGTTCCGAATTTGGGATCAAGCTTTCCTATGCGGAACAACCTAGTCCCGATGGACTTGCTCAGGCT

      F  L  I  G  E  E  F  I  G  D  D  S  V  A  L  I  L  G  D  N  I  Y  H  G  P  G  L  S  T  M  L  Q  K  A
2401  TTTCTTATCGGTGAAGAATTTATCGGTGACGATAGCGTTGCCTTGATTTTAGGCGACAATATCTACCATGGTCCTGGTCTGAGCACAATGCTTCAAAAAG

      A  K  K  E  K  G  A  T  V  F  G  Y  Q  V  K  D  P  E  R  F  G  V  V  E  F  D  T  D  M  N  A  I  S
2501  CAGCCAAGAAAGAGAAAGGTGCGACTGTTTTTGGCTACCAAGTGAAGGATCCAGAGCGTTTTGGTGTGGTTGAGTTTGATACAGACATGAATGCTATTTC

      I  E  E  K  P  E  Y  P  R  S  N  Y  A  V  T  G  L  Y  F  Y  D  N  D  V  V  E  I  A  K  Q  I  K  P
2601  TATAGAAGAAAAGCCGGAGTATCCTCGCTCCAACTATGCAGTGACAGGACTGTATTTCTATGATAATGATGTTGTGGAGATTGCTAAACAGATCAAACCT

      S  A  R  G  E  L  E  I  T  D  V  N  K  A  Y  L  N  R  G  D  L  S  V  E  L  M  G  R  G  F  A  W  L  D
2701  AGTGCTCGTGGCGAGTTAGAAATTACAGACGTTAACAAGGCTTACCTAAATCGTGGTGACCTTTCTGTTGAGCTGATGGGGCGTGGTTTTGCCTGGTTGG

      T  G  T  H  E  S  L  L  E  A  S  Q  Y  I  E  T  V  Q  R  M  Q  N  V  Q  V  A  N  L  E  E  I  S  Y
2801  ATACGGGAACCCATGAAAGCTTGCTAGAAGCTTCTCAGTATATTGAAACGGTTCAGCGTATGCAGAACGTTCAAGTGGCAAATCTAGAAGAAATTTCCTA

      R  M  G  Y  I  S  R  E  D  V  L  E  L  A  Q  P  L  K  K  N  E  Y  G  R  Y  L  L  R  L  I  G  E  A
2901  TCGTATGGGCTACATTAGTCGTGAAGACGTGCTGGAATTGGCTCAGCCTCTTAAGAAAAATGAATACGGACGATATTTGCTCCGTTTGATTGGAGAAGCA
```

cps19a$_2$M→

```
      *  M  S  D  N  F  F  G  K  T  L  V  V  R  K  I  D  A  I  P  G  L  L  E  F  D  I  P  V  H  G  D  N  R
3001  TAGATGTCAGATAATTTTTTTGGAAAGACACTTGTGGTACGCAAGATTGATGCTATACCAGGACTGCTAGAGTTTGACATTCCCGTTCATGGAGACAATC

      G  W  F  K  E  N  F  Q  K  E  K  M  L  P  L  G  F  P  E  S  F  F  A  A  G  K  L  Q  N  N  V  S  F
3101  GTGGTTGGTTTAAGGAAAATTTCCAGAAGGAAAAGATGCTACCGCTTGGTTTTCCTGAAAGCTTCTTTGCTGCAGGGAAACTGCAAAATAACGTCAGCTT

      S  R  K  N  V  L  R  G  L  H  A  E  P  W  D  K  Y  I  S  V  A  D  D  G  K  V  L  G  S  W  V  D  L
3201  TTCTCGCAAAAATGTTCTTCGAGGACTCCATGCTGAACCTTGGGACAAGTATATCTCTGTTGCAGACGATGGGAAGGTTTTAGGATCTTGGGTAGATTTG

      R  E  G  E  T  F  G  N  V  Y  Q  T  E  I  D  A  S  K  G  I  F  V  P  R  G  V  A  N  G  F  Q  V  L  S
3301  CGTGAAGGCGAGACCTTTGGAAATGTTTACCAGACAGAGATTGATGCTAGCAAGGGAATTTTTGTCCCTCGTGGCGTAGCCAACGGTTTTCAAGTCCTCT

      D  T  V  S  Y  S  Y  L  V  N  D  Y  W  A  L  E  L  K  P  K  Y  A  F  V  N  Y  A  D  P  A  L  G  I
3401  CAGATACAGTTTCCTATAGCTATCTGGTCAATGACTACTGGGCTCTTGAACTCAAACCCAAATATGCCTTTGTTAACTATGCTGACCCAGCATTGGGAAT

      E  W  E  N  L  P  E  A  E  V  S  E  A  D  K  H  H  P  L  L  R  D  V  K  P  L  T  K  D  E  L  *
3501  TGAGTGGGAAAACCTACCGGAAGCTGAGGTTTCAGAGGCAGACAAACACCATCCTCTATTAAGGGATGTCAAACCACTTACGAAAGACGAGTTGTAAAAG
```

cps19a$_2$N→

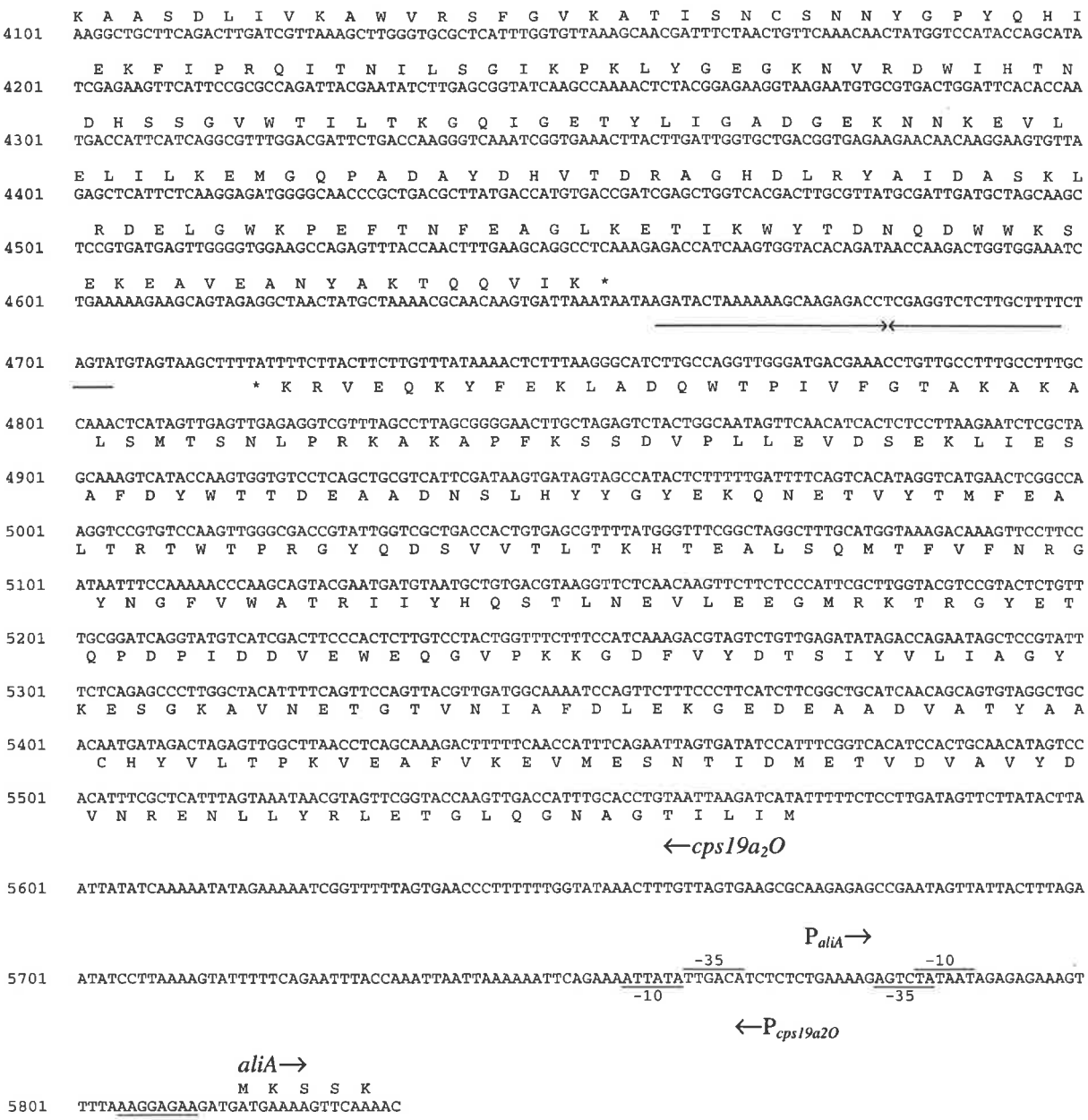
```
            M  T  E  Y  K  K  I  I  V  T  G  G  A  G  F  I  G  S  N  F  V  H  Y  V  Y  N  N  F  P  D  V
3601  GAAACATTATGACTGAATACAAAAAAATTATCGTGACAGGTGGAGCTGGTTTTATCGGTTCTAACTTTGTCCACTATGTTTACAATAACTTTCCAGATGT

      H  V  T  V  L  D  K  L  T  Y  A  G  N  R  A  N  I  E  E  I  L  G  D  R  V  E  L  V  V  G  D  I  A
3701  CCATGTGACAGTGCTGGACAAGCTGACTTATGCAGGTAATCGTGCCAATATCGAGGAAATTTTAGGCGACCGTGTTGAGTTGGTTGTTGGAGATATTGCT

      D  A  A  L  V  D  K  L  A  A  E  A  D  A  I  V  H  Y  A  A  E  S  H  N  D  N  S  L  N  D  P  S  P  F
3801  GATGCAGCCTTGGTAGACAAGTTGGCGGCTGAAGCGGATGCTATCGTTCACTATGCGGCAGAAAGCCACAATGATAACTCGCTCAATGACCCGAGTCCAT

      I  H  T  N  F  I  G  T  Y  T  L  L  E  A  A  R  K  Y  D  I  R  F  H  H  V  S  T  D  E  V  Y  G  D
3901  TTATCCACACCAACTTTATCGGAACTTACACACTTTTAGAAGCGGCTCGTAAATACGACATTCGTTTCCACCATGTATCGACTGATGAAGTCTATGGTGA

      L  P  L  R  E  D  L  P  G  H  G  E  G  L  G  E  K  F  T  A  E  T  K  Y  N  P  S  S  P  Y  S  S  T
4001  CCTGCCTCTGCGTGAAGATTTGCCAGGTCATGGGGAAGGCCTAGGTGAGAAATTTACCGCTGAAACCAAGTACAATCCAAGCTCGCCTTACTCATCAACC
```

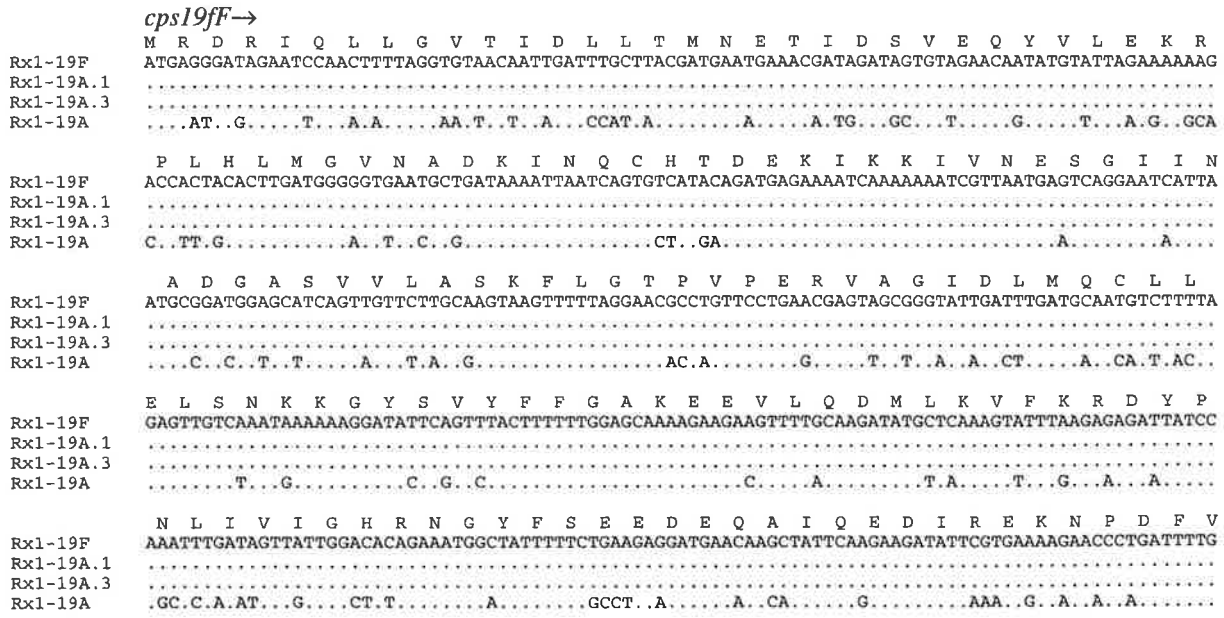

K A A S D L I V K A W V R S F G V K A T I S N C S N N Y G P Y Q H I
4101 AAGGCTGCTTCAGACTTGATCGTTAAAGCTTGGGTGCGCTCATTTGGTGTTAAAGCAACGATTTCTAACTGTTCAAACAACTATGGTCCATACCAGCATA
E K F I P R Q I T N I L S G I K P K L Y G E G K N V R D W I H T N
4201 TCGAGAAGTTCATTCCGCGCCAGATTACGAATATCTTGAGCGGTATCAAGCCAAAACTCTACGGAGAAGGTAAGAATGTGCGTGACTGGATTCACACCAA
D H S S G V W T I L T K G Q I G E T Y L I G A D G E K N N K E V L
4301 TGACCATTCATCAGGCGTTTGGACGATTCTGACCAAGGGTCAAATCGGTGAAACTTACTTGATTGGTGCTGACGGTGAGAAGAACAACAAGGAAGTGTTA
E L I L K E M G Q P A D A Y D H V T D R A G H D L R Y A I D A S K L
4401 GAGCTCATTCTCAAGGAGATGGGGCAACCCGCTGACGCTTATGACCATGTGACCGATCGAGCTGGTCACGACTTGCGTTATGCGATTGATGCTAGCAAGC
R D E L G W K P E F T N F E A G L K E T I K W Y T D N Q D W W K S
4501 TCCGTGATGAGTTGGGGTGGAAGCCAGAGTTTACCAACTTTGAAGCAGGCCTCAAAGAGACCATCAAGTGGTACACAGATAACCAAGACTGGTGGAAATC
E K E A V E A N Y A K T Q Q V I K *
4601 TGAAAAAGAAGCAGTAGAGGCTAACTATGCTAAAACGCAACAAGTGATTAAATAATAAGATACTAAAAAAGCAAGAGACCTCGAGGTCTCTTGCTTTTCT
4701 AGTATGTAGTAAGCTTTTATTTTCTTACTTCTTGTTTATAAAACTCTTTAAGGGCATCTTGCCAGGTTGGGATGACGAAACCTGTTGCCTTTGCCTTTGC
* K R V E Q K Y F E K L A D Q W T P I V F G T A K A K A
4801 CAAACTCATAGTTGAGTTGAGAGGTCGTTTAGCCTTAGCGGGGAACTTGCTAGAGTCTACTGGCAATAGTTCAACATCACTCTCCTTAAGAATCTCGCTA
L S M T S N L P R K A K A P F K S S D V P L L E V D S E K L I E S
4901 GCAAAGTCATACCAAGTGGTGTCCTCAGCTGCGTCATTCGATAAGTGATAGTAGCCATACTCTTTTTGATTTTCAGTCACATAGGTCATGAACTCGGCCA
A F D Y W T T D E A A D N S L H Y Y G Y E K Q N E T V Y T M F E A
5001 AGGTCCGTGTCCAAGTTGGGCGACCGTATTGGTCGCTGACCACTGTGAGCGTTTTATGGGTTTCGGCTAGGCTTTGCATGGTAAAGACAAAGTTCCTTCC
L T R T W T P R G Y Q D S V V T L T K H T E A L S Q M T F V F N R G
5101 ATAATTTCCAAAAACCCAAGCAGTACGAATGATGTAATGCTGTGACGTAAGGTTCTCAACAAGTTCTTCTCCCATTCGCTTGGTACGTCCGTACTCTGTT
Y N G F V W A T R I I Y H Q S T L N E V L E E G M R K T R G Y E T
5201 TGCGGATCAGGTATGTCATCGACTTCCCACTCTTGTCCTACTGGTTTCTTTCCATCAAAGACGTAGTCTGTTGAGATATAGACCAGAATAGCTCCGTATT
Q P D P I D D V E W E Q G V P K K G D F V Y D T S I Y V L I A G Y
5301 TCTCAGAGCCCTTGGCTACATTTTCAGTTCCAGTTACGTTGATGGCAAAATCCAGTTCTTTCCCTTCATCTTCGGCTGCATCAACAGCAGTGTAGGCTGC
K E S G K A V N E T G T V N I A F D L E K G E D E A A D V A T Y A A
5401 ACAATGATAGACTAGAGTTGGCTTAACCTCAGCAAAGACTTTTTCAACCATTTCAGAATTAGTGATATCCATTTCGGTCACATCCACTGCAACATAGTCC
C H Y V L T P K V E A F V K E V M E S N T I D M E T V D V A V Y D
5501 ACATTTCGCTCATTTAGTAAATAACGTAGTTCGGTACCAAGTTGACCATTTGCACCTGTAATTAAGATCATATTTTTCTCCTTGATAGTTCTTATACTTA
V N R E N L L Y R L E T G L Q G N A G T I L I M
←cps19a2O
5601 ATTATATCAAAAATATAGAAAAATCGGTTTTTAGTGAACCCTTTTTTGGTATAAACTTTGTTAGTGAAGCGCAAGAGAGCCGAATAGTTATTACTTTAGA
PaliA→
-35
-10
5701 ATATCCTTAAAAGTATTTTTCAGAATTTACCAAATTAATTAAAAAATTCAGAAAATTATATTGACATCTCTCTGAAAAGAGTCTATAATAGAGAGAAAGT
-10
-35
←Pcps19a2O
aliA→
M K S S K
5801 TTTAAAGGAGAAGATGATGAAAAGTTCAAAAC

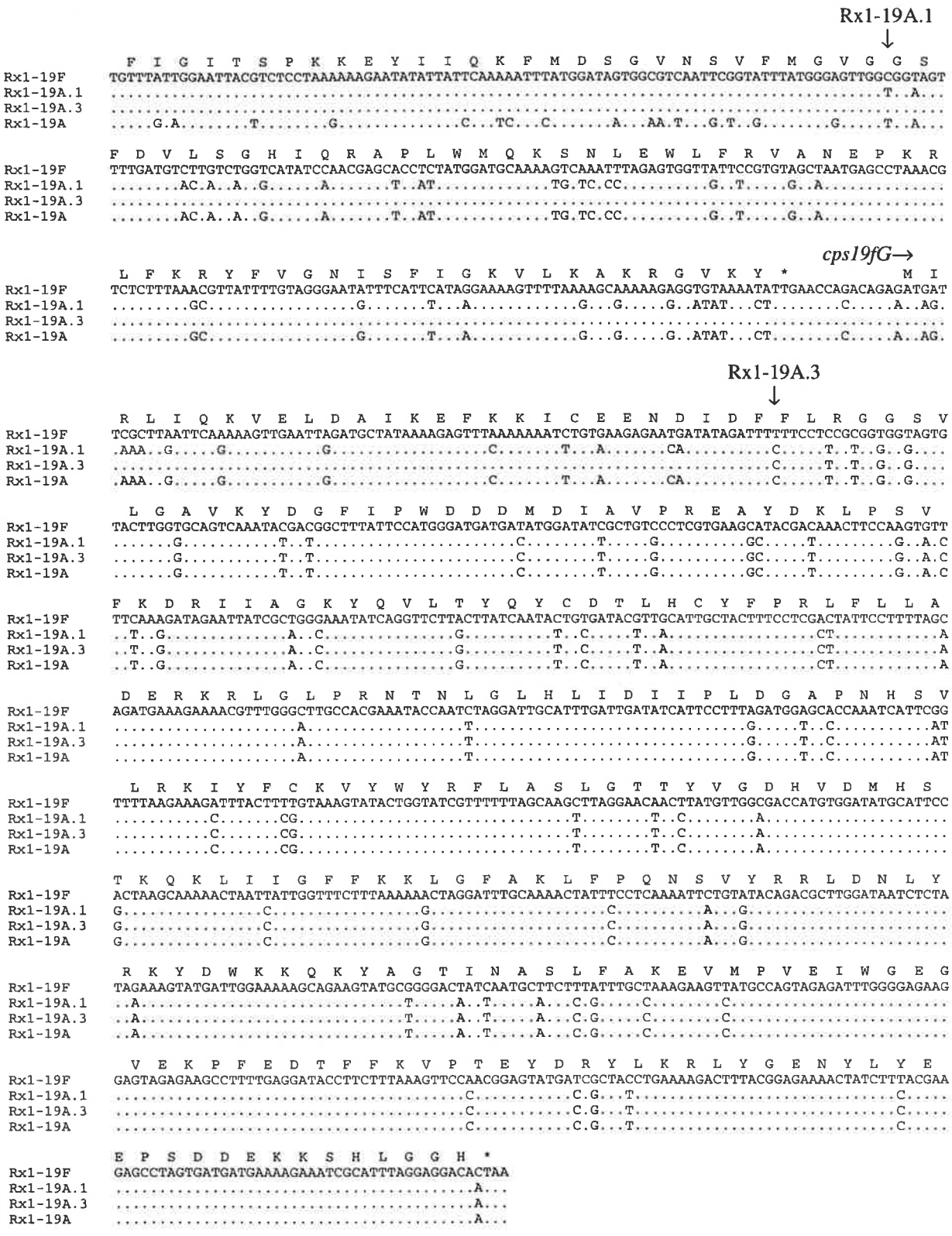
APPENDIX V

The nucleotide sequence of Rx1-19F, Rx1-19A.1, Rx1-19A.2, Rx1-19A.3 and Rx1-19A in the regions where recombination between the *cps19f* and *cps19a* loci has occurred.

A.

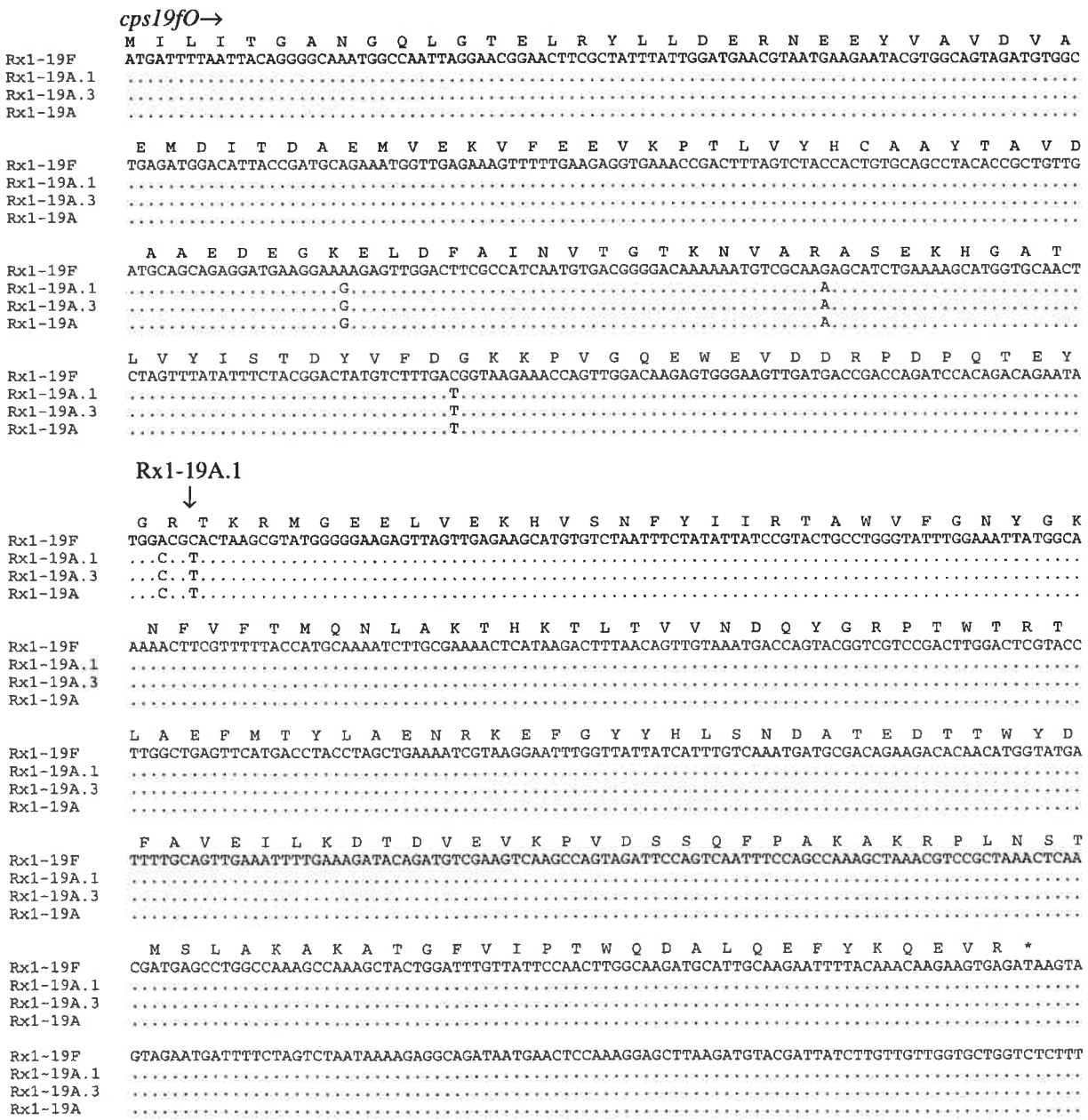
The nucleotide sequence of *cps19F* and *cps19G* for Rx1-19F, Rx1-19A.1, Rx1-19A.3 and Rx1-19A indicating the regions where recombination at the 5' end of the *cps* locus has occurred in Rx1-19A.1, Rx1-19A.3. The amino acid translation for Rx1-19F is represented by single letter code above the first nucleotide of each codon. The vertical arrows indicate the first nucleotide which is type 19A-specific in each transformant.

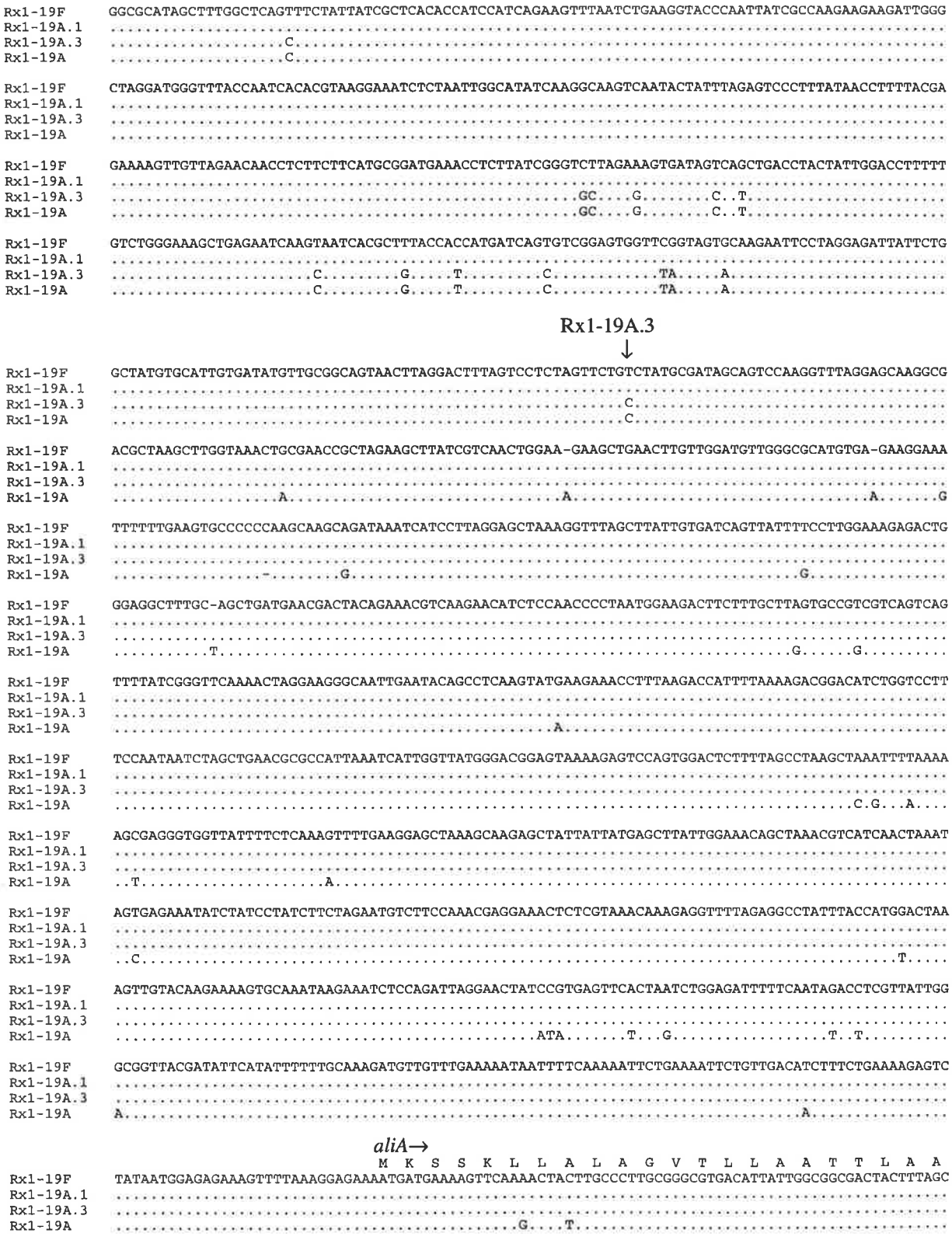



Rx1-19A.1
cps19fG→
Rx1-19A.3
Rx1-19F
Rx1-19A.1
Rx1-19A.3
Rx1-19A

B.

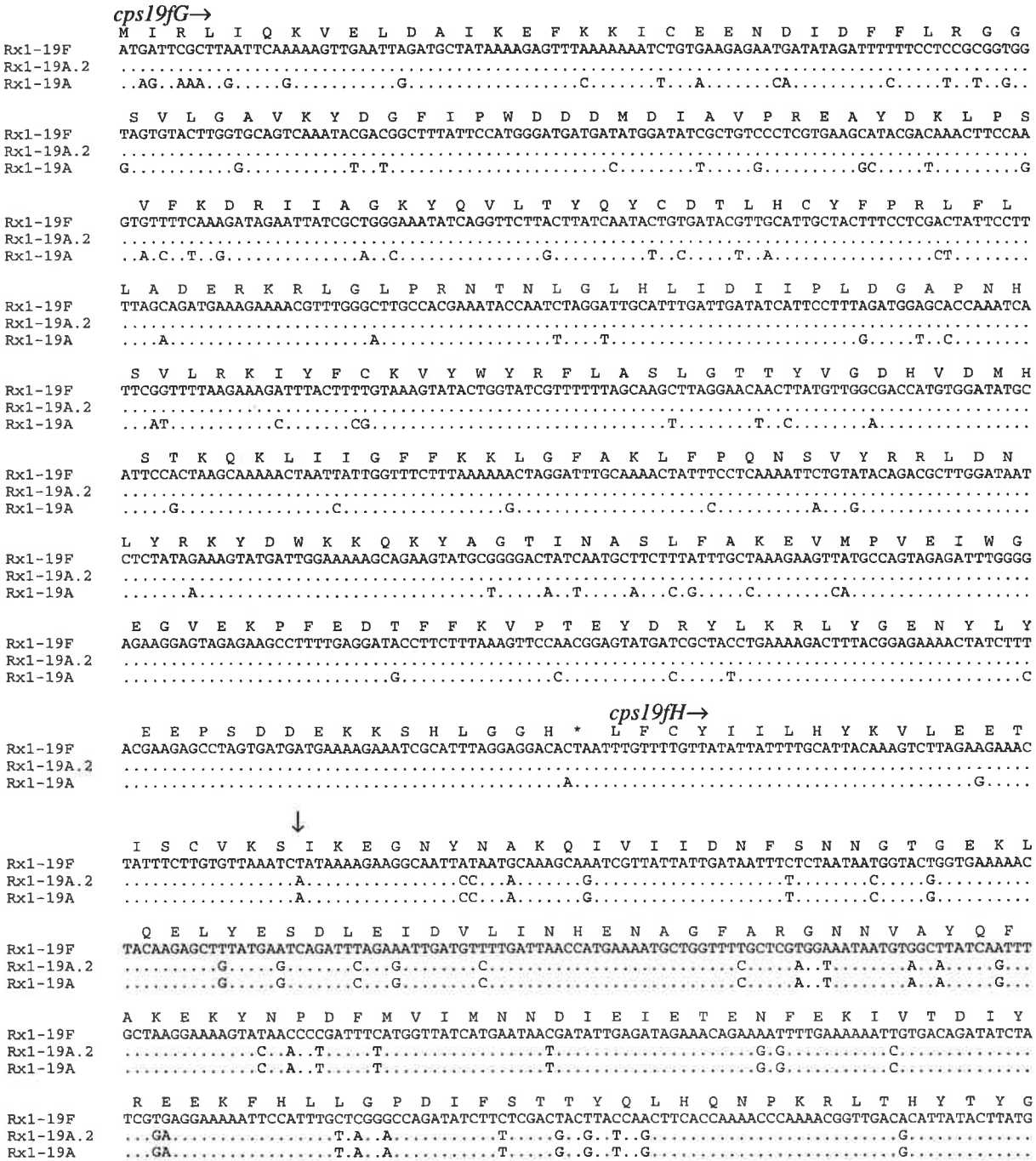
The nucleotide sequence of *cps19O* and the 3' intergenic region for Rx1-19F, Rx1-19A.1, Rx1-19A.3 and Rx1-19A indicating the regions where recombination at the 3' end of the *cps* locus has occurred in Rx1-19A.1, Rx1-19A.3. The amino acid translation for Rx1-19F is represented by single letter code above the first nucleotide of each codon. The vertical arrows indicate the last nucleotide which is type 19A-specific in each transformant.

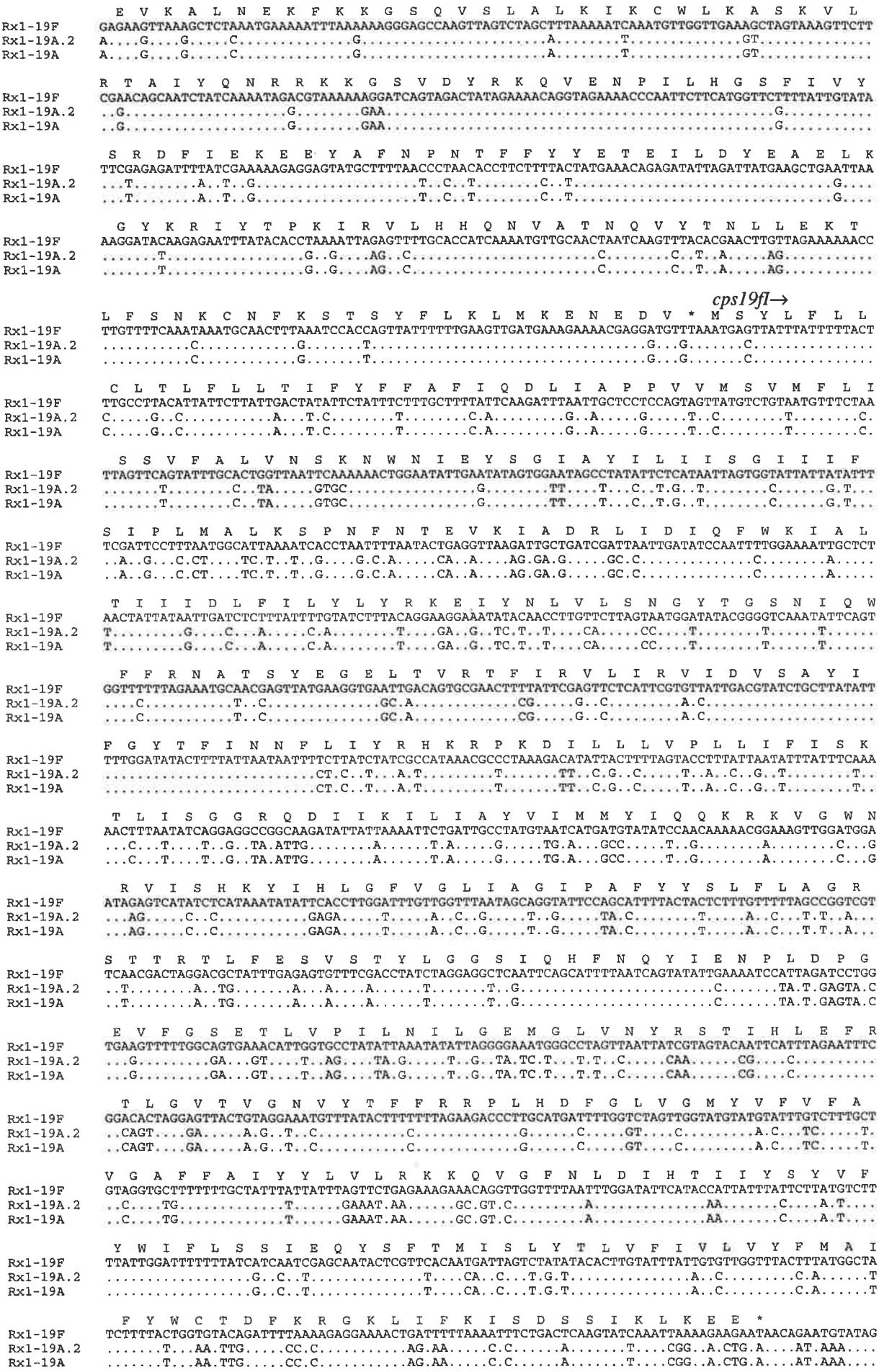


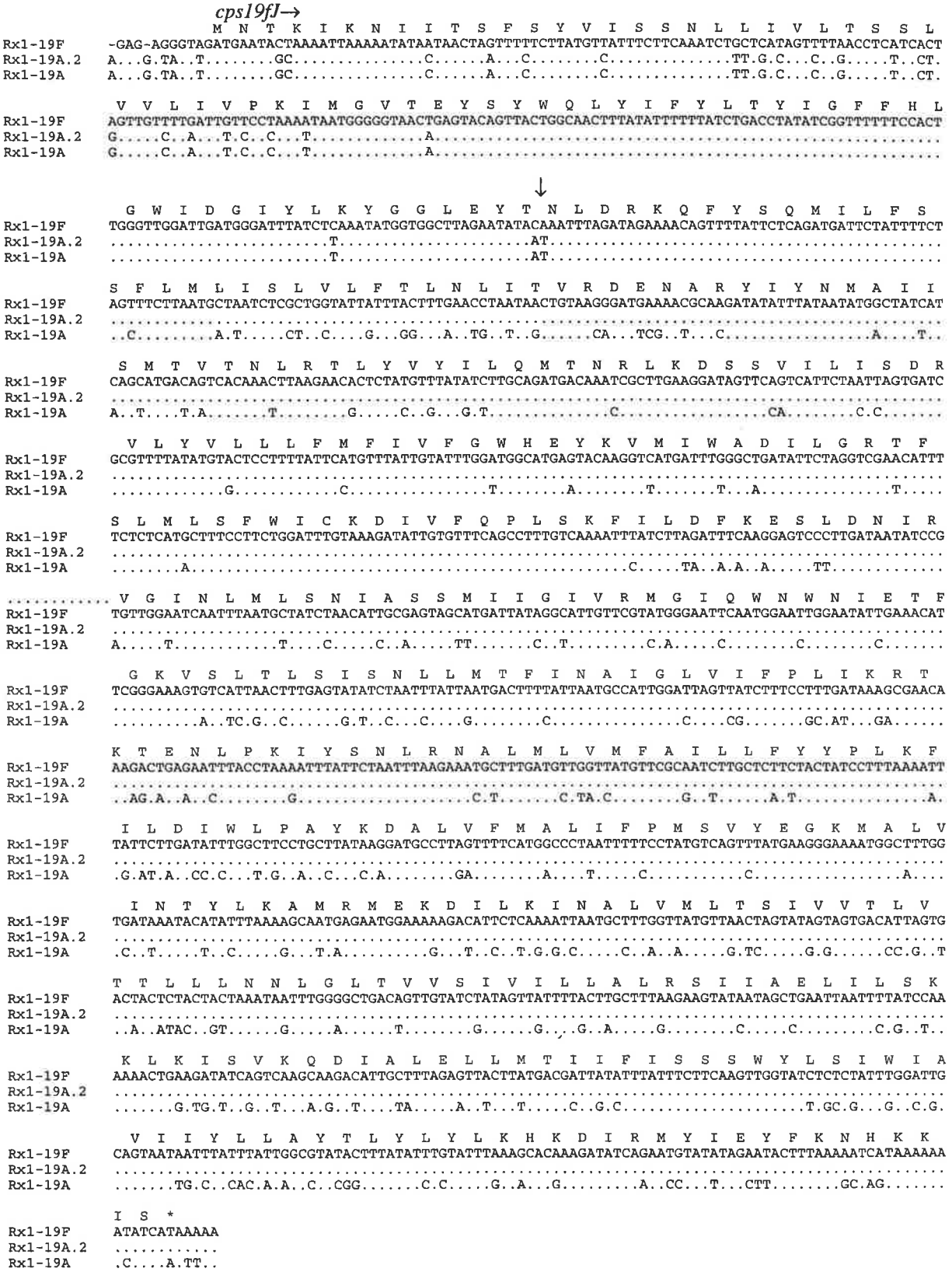

Rx1-19A.3
aliA→
M K S S K L L A L A G V T L L A A T T L A A

C.

The nucleotide sequence of *cps19G*, *cps19H*, *cps19I* and *cps19J* for Rx1-19F, Rx1-19A.2 and Rx1-19A The amino acid translation for Rx1-19F is represented by single letter code above the first nucleotide of each codon. The region where recombination has occurred is indicated by vertical arrows marking the first and last nucleotides which are type 19A-specific.




cps19fl→


cps19fJ→
Rx1-19F
Rx1-19A.2
Rx1-19A

APPENDIX VI

The nucleotide and deduced amino acid sequence of the central (type-specific) portion of *cps19c* locus.

The nucleotide sequence is numbered in accordance with GenBank accession number AF105116 and starts from the distal portion of *cps19cR* to the beginning of *cps19cL*. The amino acid translation for each ORF is represented by single letter code above the first nucleotide of each codon. Possible ribosome binding sites are underlined.

```
      . . . cps19cR→
        A  Y  K  K  M  E  F  L  P  F  E  Q  K  N  L  W  L  M  M  S  L  K  K  E  C  I  Y  I  P  K  L  Q  S
  1   AAGCTTATAAGAAAATGGAGTTTCTCCCTTTTGAACAAAAAAATCTGTGGCTAATGATGTCGCTAAAAAAAGAATGTATTTATATACCTAAGTTACAATC
        F  S  E  S  L  K  E  K  V  V  E  N  L  G  E  I  V  T  L  K  T  T  K  K  S  I  K  T  L  L  R  K  N
 101  ATTTTCTGAATCTTTAAAAGAAAAAGTTGTTGAAAATTTGGGTGAGATAGTAACTCTTAAGACTACGAAAAAATCTATAAAAACGTTATTAAGAAAGAAT
        S  D  L  L  W  V  D  F  V  Y  R  L  F  G  G  V  N  F  L  F  I  P  E  L  Y  T  R  V  I  F  L  L  K  Y
 201  AGCGATTTACTTTGGGTTGATTTCGTTTATAGATTGTTTGGTGGTGTCAATTTCCTATTTATTCCAGAATTATATACAAGAGTTATATTTTTATTGAAAT
        L  D  L  K  L  R  L  K  N  Y  G  E  *

                                                  cps19cJ→
                                                          M  G  N  K  S  I  K  L  N  A  L  L  N  I  V  L  T  L  S  N  I  I  F  P
 301  ATTTAGATTTGAAATTGAGGTTGAAAAATTATGGGGAATAAATCCATAAAGTTGAATGCATTATTAAATATTGTCCTGACGCTATCAAATATCATTTTCC
          L  I  T  F  P  Y  I  S  R  I  L  N  P  N  G  I  G  L  T  S  F  F  S  S  I  G  N  Y  G  I  L  L  A
 401  CATTAATCACTTTCCCTTATATATCTAGAATATTGAATCCAAATGGTATAGGTTTAACTTCATTTTTTAGTTCAATAGGGAATTATGGTATTTTACTTGC
        S  L  G  I  S  T  Y  G  I  K  A  V  A  S  V  R  D  D  R  D  K  L  S  K  V  V  Q  E  L  M  I  I  N
 501  TTCTCTGGGAATTTCAACTTACGGTATCAAAGCAGTAGCAAGTGTTAGAGATGATAGAGATAAGTTGTCAAAAGTAGTACAGGAGTTAATGATTATAAAC
        V  A  M  S  I  I  T  T  A  I  L  L  F  M  T  I  F  I  T  Q  L  N  R  E  F  S  L  L  L  I  T  C  G  T
 601  GTTGCTATGTCTATAATAACAACTGCAATACTATTATTTATGACTATATTTATAACACAATTGAATAGAGAATTTTCACTCCTATTGATCACATGTGGGA
          I  L  S  S  P  F  A  L  N  W  L  Y  S  G  M  E  E  Y  T  Y  I  T  T  R  S  V  V  F  T  I  L  S  L
 701  CTATTTTATCTTCTCCTTTCGCCTTAAATTGGTTGTATAGTGGAATGGAAGAATATACGTATATTACTACTAGGTCAGTAGTGTTTACAATTCTATCATT
          I  L  I  F  L  L  V  K  R  P  E  D  Y  I  V  F  A  S  I  S  L  F  S  S  L  S  S  N  I  L  N  L  W
 801  AATATTGATTTTTCTACTTGTGAAAAGGCCAGAGGATTATATTGTTTTTGCTAGTATTTCATTGTTTTTCTTCTCTAAGTTCAAATATCTTAAATCTATGG
        H  S  R  H  F  I  N  I  K  L  Y  K  N  L  Q  F  K  Y  H  F  K  P  M  W  Y  L  F  A  S  L  L  A  V  N
 901  CATAGCCGACATTTCATTAATATTAAATTATATAAAAATTTACAATTTAAATATCATTTTAAACCAATGTGGTATTTATTTGCCTCATTACTTGCAGTAA
          I  Y  T  N  L  D  T  V  M  L  G  F  I  N  G  N  E  A  V  G  Y  Y  S  V  A  S  K  V  K  W  I  L  L
1001  ATATTTATACTAATTTAGATACAGTGATGCTCGGTTTTATTAATGGTAATGAGGCTGTGGGATACTATTCTGTGGCATCAAAGGTTAAGTGGATTTTGCT
        S  L  I  T  S  I  S  A  V  L  L  P  R  L  S  F  Y  I  S  K  N  D  T  S  N  F  I  K  M  L  K  E  S
1101  TTCTCTTATTACATCTATTAGTGCAGTTTTGCTACCGAGACTTTCATTTTATATTAGTAAAAATGACACCTCGAATTTTATAAAAATGTTAAAGGAGTCA
        S  A  V  I  F  F  I  A  I  P  L  M  V  F  F  I  V  E  A  K  D  S  I  L  L  L  G  G  S  Q  Y  L  P  A
1201  TCTGCGGTTATATTTTTTATTGCGATTCCATTGATGGTATTCTTTATTGTAGAGGCGAAAGATAGTATCTTATTACTAGGAGGAAGTCAGTATCTTCCTG
          T  L  A  M  Q  I  L  M  P  I  L  L  I  S  G  F  S  N  I  T  G  N  Q  I  L  I  P  M  N  R  E  K  Y
1301  CGACTTTAGCGATGCAAATACTTATGCCAATTTTACTTATTTCTGGTTTCTCGAATATTACAGGAAATCAAATATTGATTCCAATGAATAGAGAAAAATA
        F  M  V  A  V  T  I  G  A  V  I  N  L  I  L  N  L  L  L  M  P  K  F  G  I  I  G  A  S  V  A  T  L
1401  TTTTATGGTTGCAGTAACGATTGGTGCTGTGATTAATCTTATTTTGAATCTACTGTTAATGCCTAAGTTTGGAATTATTGGTGCTTCTGTCGCAACTCTT
        F  A  E  L  S  Q  M  T  V  Q  L  H  F  S  K  E  Y  L  V  S  N  I  S  I  K  S  L  V  N  V  I  I  A  T
1501  TTTGCGGAATTGTCGCAGATGACGGTACAATTACATTTTTCAAAAGAATATTTAGTATCAAATATATCGATAAAGAGTTTGGTTAATGTGATAATTGCAA
          V  V  S  T  I  P  L  I  I  L  N  Q  L  I  T  I  T  I  P  F  Y  S  L  M  L  A  G  F  A  F  F  S  L
1601  CAGTTGTTTCTACAATACCACTAATCATTTTGAATCAGCTGATAACGATAACTATACCATTTTATTCTCTAATGCTAGCAGGTTTTGCTTTCTTTTCATT
          Y  L  V  I  L  L  L  L  K  E  E  V  T  I  Q  L  F  S  L  L  A  K  K  K  *
1701  ATATTTAGTAATTCTGCTTTTATTAAAGGAGGAAGTGACGATTCAATTATTTTCTCTTCTTGCAAAGAAGAAGTAAATTGGTTAGAAATTGAAATGTATA

                                                  cps19cK→
                                                          M  K  I  M  L  V  F  G  T  R  P  E  A  I  K  M  C  P  L  V  N
1801  AACAAATAAAGAATTTAATTATTGATTTAGGAGGAAATCATGAAGATAATGCTAGTTTTTGGTACACGTCCAGAAGCGATAAAAATGTGTCCATTAGTGA
        E  L  K  K  Q  A  D  M  E  T  V  V  C  V  T  G  Q  H  K  E  M  V  S  P  V  L  E  L  F  G  V  Q  P
1901  ATGAGTTGAAAAAACAGGCAGATATGGAAACAGTTGTTTGTGTAACTGGTCAACACAAGGAGATGGTTAGCCCTGTTTTGGAATTGTTTGGAGTTCAACC
          D  Y  D  L  E  I  M  K  A  N  Q  T  L  F  S  I  T  T  S  I  L  E  K  I  K  P  V  L  E  E  E  Q  P
2001  AGACTATGATTTAGAAATAATGAAAGCTAATCAAACCTTGTTCTCTATAACAACAAGTATTTTAGAAAAAATTAAACCTGTTTTAGAAGAAGAACAACCA
```

```
      D  I  V  L  V  H  G  D  T  T  T  T  Y  A  A  A  L  A  A  F  Y  L  G  I  K  V  G  H  V  E  A  G  L  R
2101  GATATTGTTCTAGTTCATGGTGACACTACTACAACATATGCAGCAGCTTTGGCAGCATTTTATCTTGGAATTAAGGTTGGCCATGTTGAAGCTGGGTTAC
        T  Y  N  L  Q  S  P  F  P  E  E  F  N  R  Q  S  T  S  I  I  A  N  Y  H  F  A  P  T  E  L  A  K  E
2201  GAACTTACAACCTGCAAAGTCCCTTCCCAGAAGAATTTAATAGACAATCGACTTCTATTATTGCAAATTATCATTTTGCTCCAACTGAATTGGCTAAAGA
        N  L  L  K  E  G  R  E  N  V  Y  V  T  G  N  T  V  I  D  A  L  T  T  T  V  Q  K  D  Y  T  H  P  D
2301  AAATCTCTTAAAAGAGGGTAGAGAGAATGTTTATGTAACTGGAAATACAGTTATTGATGCACTTACAACTACAGTACAAAAGGATTATACACACCCTGAT
        L  D  L  N  V  N  T  R  L  I  L  L  T  A  H  R  R  E  N  L  G  E  P  M  K  H  M  F  R  A  V  K  R  V
2401  TTAGATTTAAACGTTAACACTCGTCTTATTCTACTGACTGCTCATAGACGTGAAAATCTCGGAGAACCTATGAAACACATGTTTAGAGCTGTTAAACGAG
        L  N  E  Y  D  D  V  K  V  I  Y  P  I  H  K  N  P  L  V  R  E  T  A  T  E  I  F  G  D  T  E  R  I
2501  TCTTAAATGAATATGACGATGTTAAGGTAATTTATCCAATTCATAAGAATCCCTTGGTGCGTGAAACAGCTACAGAAATTTTTGGAGACACAGAACGTAT
        Q  I  I  E  P  L  D  V  L  D  F  H  N  F  M  N  H  S  Y  M  I  L  T  D  S  G  G  V  Q  E  E  A  P
2601  TCAGATTATTGAACCTTTGGATGTTCTTGATTTTCATAATTTCATGAATCATAGTTACATGATTTTAACTGATTCAGGAGGAGTTCAGGAAGAGGCTCCT
        S  L  G  K  P  V  L  V  M  R  D  T  T  E  R  P  E  G  V  A  A  G  T  L  K  L  V  G  T  D  E  E  T  I
2701  TCTTTAGGAAAACCTGTATTGGTCATGCAGATACGACAGAAAGACCTGAAGGAGTAGCTGCCGGAACGTTGAAATTGGTTGGAACTGATGAGGAGACTA
        Y  Q  N  F  K  M  L  L  D  D  S  E  E  Y  K  K  M  S  Q  A  S  N  P  Y  G  K  G  D  A  S  K  Q  I
2801  TTTATCAAAACTTTAAGATGCTTTTAGACGATTCCGAAGAATATAAAAAAATGAGTCAAGCTAGTAATCCTTATGGAAAAGGTGATGCTAGTAAACAGAT
        V  H  I  L  S  G  I  *
2901  TGTTCATATTTTAAGCGGAATTTAAGCGAGGCCAAATAAAGTAATAAAAAACACTATCTTATAAAAGGTATTGATCTTGTAGTTGATTCGGGAACATGTT
3001  CATATCCTTTGACTCGAGCTACATAAAAATAACTTGTGCTGATTTATGATAGATCGATAATTTACTACTACTTTGGACATTGATGTTAGCAGTTATTAGG
3101  GATGTTTTGATTAACTCAACTTTTCAGGATTCGCCTTGCTTTCTAGATTTTCTTCTGATTTTTATCATTACTTCAGTAAGTTAAATCGTCTATTTACTAG
```

cps19cS→

```
                                                                                                  L  K  I  V  I  P  R  I  I
3201  AATTAAAACTTTTAAAAAATCCATTGTTAGCTCATGTTTTAGATATAGGTAAACATAAATTTAGGAGGTGTTAGTTTGAAGATTGTAATTCCAAGAATTA
        H  N  K  E  Q  L  T  W  D  W  S  G  T  I  T  N  I  K  K  F  L  G  K  Y  E  I  V  E  E  Q  N  I  F
3301  TTCATAACAAAGAACAACTGACCTGGGATTGGTCCGGGACAATAACTAATATAAAAAAATTTTTAGGGAAATACGAGATTGTTGAGGAACAGAATATTTT
        Y  T  F  R  M  N  V  H  K  V  L  V  R  L  G  I  K  K  S  D  M  S  M  T  Y  I  K  Y  A  E  N  Q  V
3401  CTATACTTTTAGAATGAATGTGCACAAAGTGCTTGTTCGTTTAGGTATTAAAAAATCTGATATGAGCATGACGTATATAAAATATGCTGAAAATCAAGTT
        H  L  S  P  E  D  V  C  L  T  F  D  E  F  P  L  S  F  P  D  N  P  V  Y  I  Y  Q  D  L  N  L  H  Y  L
3501  CATCTATCGCCAGAGGATGTTTGTCTCACGTTTGATGAATTTCCTTTATCTTTTCCTGATAATCCAGTTTATATCTATCAAGACTTAAATCTTCATTATT
        I  E  S  S  Q  N  N  S  Q  S  F  K  Y  S  G  F  Q  N  V  P  A  D  I  L  D  R  R  M  R  K  Q  E  I
3601  TGATAGAGAGTTCTCAAAACAATAGTCAATCGTTCAAATATAGTGGTTTTCAAAACGTCCCGGCTGATATTCTAGATAGACGAATGAGAAAACAGGAAAT
        F  Y  N  Q  A  T  G  I  F  T  M  S  K  W  F  S  D  Y  L  I  A  Q  Q  G  L  P  V  E  K  V  H  Y  V
3701  ATTTTATAATCAAGCTACTGGAATATTTACTATGAGTAAATGGTTTTCAGATTACTTGATAGCTCAACAAGGACTTCCAGTTGAAAAAGTTCATTATGTG
        G  A  G  T  N  M  N  N  L  F  L  D  H  S  H  K  E  R  N  K  F  L  F  I  G  K  D  F  F  R  K  G  G  D
3801  GGGGCAGGAACAAATATGAATAATCTATTTCTTGACCACTCTCATAAGGAACGTAATAAGTTTTTATTTATTGGTAAAGATTTTTTTCGTAAAGGAGGAG
        L  V  Y  N  A  F  V  Y  L  Q  N  N  L  M  P  E  A  E  L  Y  I  I  G  P  S  D  V  P  M  E  F  N  N
3901  ATCTTGTTTATAATGCTTTTGTCTATTTGCAAAATAATCTCATGCCGGAGGCAGAGTTGTACATTATAGGTCCTTCAGATGTTCCGATGGAATTTAACAA
        P  N  V  Y  F  L  G  N  L  S  A  D  K  V  Q  Y  F  Y  N  L  C  D  V  F  V  L  P  S  R  F  E  A  F
4001  TCCGAATGTTTATTTTTTAGGTAATCTATCAGCCGATAAGGTGCAATATTTTTATAATCTTTGTGATGTATTTGTTTTACCTTCCCGATTTGAGGCATTT
        G  I  V  F  V  E  A  L  C  Y  G  L  P  C  I  G  R  D  L  M  E  M  P  N  L  I  Q  N  N  E  T  G  L  L
4101  GGAATTGTATTTGTTGAAGCTCTTTGCTATGGTTTACCATGTATCGGTCGTGATTTAATGGAAATGCCAAACCTAATTCAAAATAATGAAACTGGATTAT
        L  P  T  E  E  E  N  P  Q  V  L  A  D  V  M  Y  N  L  I  K  D  E  S  F  F  K  N  V  Q  T  K  Q  D
4201  TATTACCTACTGAAGAGGAAAATCCACAGGTTTTAGCTGACGTAATGTATAATTTGATAAAAGATGAAAGCTTTTTTAAAAATGTTCAGACTAAACAAGA
        Y  Y  K  A  E  Y  S  W  D  T  V  A  K  R  M  I  S  I  M  K  Q  D  M  N  N  N  L  *
4301  TTATTATAAAGCAGAATATTCGTGGGACACAGTTGCCAAAAGAATGATTTCAATTATGAAGCAAGATATGAACAACAATCTATAAGTGAAAAATTAAAAG
4401  ATATTCTAGTCTAACGGTCTATTCCTACTAACAGTAATACTGAAAGAAATTTTAAAATACAAGTTATACGTATAGTGACTTCTCCTAATTCTGACATGGT
4501  GAAAAGATTATTTAAAACCTAATTTAAAAGATTAAAGGAAGTTAAATTCAGCTTATTGAATAAAATTAATTTAGTCTAGTTATTTTCTTGTCAGTCTTAT
4601  TCATTTTTTTTCCGAACCTTGTGAGTTGTTTGCAACTACCAGAATACACAAAGAACAGAATCTCTTAGCATCATTGAATAGAACAATATCTGCATAAATTT
4701  CAGATATAACTAGTGGTGCAGAAAGTGGATCTTTGAAAATAAGAATTCTTTGTAGAGCTATTTCACTAGCTGGTGAAGAATTGTTTTTCAATTTTTCGTA
4801  CAGATTACAATAATGTTATGGTGGGTTTAAATTTCTTTGTTGGTTGTGTTTGATGAATAGAGATGCACGTATCGAAATTCAAGATAGTGGTGATATCGCT
4901  TGCTTAATCTCTTCTCCAGTATTTGTAGTAGAATTAAGTGTTCTTGATTAAAATAAAGAACACAACACTTTTTATTCAGTGTTGTGTATGGGTGAAACGA
```

cps19cL→

```
                  M  K  G  I  I  L  A  G  G  S  G  T  R  L  Y  P  L  T  R  A  A  S  K  Q  L  M  P  V
5001  AAGGAACGATTGTACTTATGAAAGGTATTATTCTCGCGGGTGGTTCGGGGACACGTTTATATCCTTTGACTCGAGCTGCATCAAAGCAACTGATGCCGGT
        Y  D  K  P  M  I  Y  Y  P  L  S  T  L  M  L  A  G  I  R  D  I  L  I  I  S  T  P  Q  D  L  P  R  F
5101  TTATGATAAACCGATGATTTACTACCCACTTTCAACTTTGATGTTGGCTGGGATTAGGGATATTTTGATTATCTCAACTCCTCAAGATTTGCCTCGTTTT
        K  E  L  L  Q  D  G  S  E  F  G  I  Q  L  S  Y  A  E  Q  P  S  P  D  G  L  D
5201  AAAGAGCTCCTTCAAGATGGCTCTGAGTTTGGGATTCAATTGTCTTATGCAGAGCAACCAAGTCCAGATGGTTTGGATCC
```

APPENDIX VII

Publications

Morona, JK., Morona, R. and Paton, JC. 1997. Characterization of the locus encoding the *Streptococcus pneumoniae* type 19F capsular polysaccharide biosynthetic pathway. Mol. Microbiol. 23:751-763.

Morona, JK., Morona, R. and Paton, JC. 1997. Molecular and genetic characterization of the capsule biosynthesis locus of *Streptococcus pneumoniae* type 19B. J. Bacteriol. 179:4953-4958.

Morona, JK., Morona, R. and Paton, JC. 1999. Analysis of capsule loci from various *Streptococcus pneumoniae* serotypes using long-range PCR identifies two classes of *cpsC*. Microbiol. (manuscript submitted).

Morona, JK., Morona, R. and Paton, JC. 1999. Genetics of capsular polysaccharide biosynthesis in *Streptococcus pneumoniae* types belonging to serogroup 19. (manuscript in preparation).

Paton, JC. and Morona, JK. 1999. *Streptococcus pneumoniae* capsular polysaccharide. In: Fischetti, V., Novick, R., Ferretti, J., Portnoy, D. and Rood, J. (eds.). Gram-positive pathogens. ASM Press, Washington DC, USA. (in press).