Sequence and structural analysis of the rfb (O antigen) gene cluster from a group C1 Salmonella enterica strain

Sang J. Lee,† Lajwant K. Romana‡ and Peter R. Reeves*

Department of Microbiology, The University of Sydney, NSW 2006, Australia

(Received 15 January 1992; revised 22 May 1992; accepted 8 June 1992)

The rfb (O antigen) gene cluster of a group C1 Salmonella enterica strain was sequenced; it comprised seven open reading frames which precisely replaced the 16 open reading frames of a group B strain. Two genes of the mannose biosynthetic pathway were present: rfbK (phosphomannomutase) had a G+C content of 0.61 and had only 40% identity to rfbK of group B but was very similar to cpsG of the capsular polysaccharide pathway with 96% identity, whereas rfbM (guanosine diphosphomannose (GDP-Man) pyrophosphorylase) had a G+C content of 0.39. Other genes had G+C contents ranging from 0.24 to 0.28. rfbM(C1) and rfbM(B) had 60% identity, which is much less than expected within a species, but nonetheless indicates a much more recent common ancestor than for rfbK. The other genes showed much lower or no similarity to rfb genes of other S. enterica strains. It appears that the gene cluster evolved outside of Salmonella in a species with low G+C content: the rfbM gene presumably derives from that period whereas the rfbK gene appears to have arisen after transfer of the cluster to S. enterica by duplication of the S. enterica cpsG gene, presumably replacing an rfbK gene of low G+C content.

Introduction

The O antigen of Gram-negative bacteria is a cell surface polysaccharide which comprises many repeats of an oligosaccharide. This repeat unit varies among strains and in Salmonella enterica there are about 60 different forms (Ewing, 1986) which have been used to divide the species serologically into groups A, B, C1 etc. In group C1 strains the repeat unit contains four mannose residues, one N-acetylglucosamine residue and a glucose branch (Lindberg et al., 1988). The genes responsible for biosynthesis and assembly of the O antigen repeat unit are located within the rfb gene cluster which maps at 42 min on the S. enterica chromosome (Sanderson & Roth, 1989). The genetics of the rfb region has been reviewed by Mäkelä & Stocker (1984), and more recently we have cloned the rfb regions from groups B (Brahmbhatt et al., 1988; Jiang et al., 1991; Wyk & Reeves, 1989), A and D (Liu et al., 1991; Verma et al., 1988; Verma & Reeves, 1989), C1 (Lee et al., 1992), C2 (Brown et al., 1991, 1992) and E1 (Wang et al., 1992).

In S. enterica of groups A, B and D polymerization of the repeat units to give complete O antigen is dependent on the rfc gene which maps at 32 min (Mäkelä & Stocker, 1984; Sanderson & Roth, 1989). The O antigens of E. coli groups O8 and O9 have been shown to be synthesized by single sugar chain extension and not by synthesis and subsequent polymerization of an oligomeric O unit. It has been suggested that group C1 S. enterica resembles groups O8 and O9 E. coli in this regard as all three have in common dependence on a functional rfe gene (Jann & Jann, 1984).

In group C2 the rfc gene is within the rfb region (Brown et al., 1991, 1992); in group E1 the polymerization function is again within the rfb region and a potential rfc gene has been identified (Wang et al., 1992). Some studies have indicated that in group C1 the polymerization function is located within the rfb region (Mäkelä, 1966): however, if group C1 O antigen synthesis, like that of E. coli groups O8 and O9, is by single sugar extension, then there will not be an O antigen polymerase at all.

In our report on the cloning of the group C1 rfb gene cluster we localized within the cluster the rfbM and rfbK genes (Lee et al., 1992). The rfb region of group C1 showed no significant similarity on Southern hybridization to the rfb regions of groups A, B, D, C2 or E1.

* Author for correspondence. Tel. 2 692 2536; fax 2 692 4571.
† Present address: Department of Biochemistry, University of Medicine & Dentistry of New Jersey, Piscataway, New Jersey 08854-5635, USA.
‡ Present address: Department of Biochemistry, University of Queensland, Royal Brisbane Hospital, Herston, Queensland 4029, Australia.

The nucleotide sequence data reported in this paper have been submitted to GenBank and have been assigned the accession number M84642.

0001-7337 © 1992 SGM
However, one of the mannose pathway genes (rfbK) hybridized strongly to the cps region of strain LT2. To further understand the evolutionary origin of the rfb gene cluster of group C1, we have sequenced the whole region, studied its expression in E. coli K12 and compared the sequence with those of other rfb gene cluster sequences.

Methods

Media, antibiotics, DNA techniques, strains, plasmids and bacteriophages. Media, antibiotics and DNA techniques were reported previously (Lee et al., 1992), except for 2 x YT broth (1.6%, w/v, Bacto-tryptone, 1%, w/v, yeast extract, 0.5% NaCl) used for M13 phage propagation. S. enterica strain M40 was the source of all rfb region DNA used in the study. All E. coli strains used were derivatives of strain K12: NM522 (Gough & Murray, 1983) was used as host for pGEM-7zf(+) and pT7T3 19U plasmids; JM101 (Messing et al., 1981) was used for propagation of M13mp19; C600 (Maniatis et al., 1982) was used for propagation. S. enterica strain M40 was the source of all rfb region DNA. The vector used was pT7T3 19U or pGEM-7zf(+) and nested deletion sets were generated with Ex0111 and S1 nuclease (Henikoff, 1987); selected deletion derivatives were sequenced from packaged cosmid obtained using M13K07 as helper phage (Vieira & Messing, 1987). The sequencing direction was determined by the orientation of the insert in the original clone and is indicated in Fig. 1.

All the cloned fragments overlapped other fragments except at the Xhol site at position 7160; the BglII-PstI (6726-8631) fragment was cloned into M13mp19 for sequencing across this site using single stranded DNA prepared from phage.

DNA sequencing was done by the dideoxy chain termination method (Sanger et al., 1977, 1980) using dye-labelled M13 standard primer (Promega), or other oligonucleotide primers used together with dye-labelled dideoxy terminators. The methods in both cases were as described by Applied Biosystems for the 373A automatic sequencer, which was used to analyse the products. Specific primers for dye dideoxy termination sequencing were synthesized on an Applied Biosystems DNA synthesizer.

Computer analysis. Computer analysis of sequence data was done on the MBIS system at the CSIRO Biotechnology Laboratory, Sydney, Australia (Bucholtz & Reisner, 1986) or (after June 1990) on the Australian National Genomic Information Service (ANGIS) at Sydney University: both systems incorporate several sets of programs.

Sequence data was assembled using the sap program of Staden (1982, 1986). We used the program nip (Staden, 1984) to calculate G+C content, molecular weight and a hydrophobicity index. The rny preference method was used to look for open reading frames (Shepherd, 1981). The programs ALOM (GCC package; Devereux et al., 1984) based on the method of Klein et al. (1985) and PepPlot (GCC package) based on the method of Kyte & Doolittle (1982) were used to identify potential transmembrane segments, and the transmembrane segments were then further checked using the algorithm described by Eisenberg et al. (1984).

The programs BESTFIT and GAP in the GCC package were used to do sequence comparisons. The parameters for running BESTFIT were gap weight, 3.0; gap length weight, 0.1; the parameters for running GAP were gap weight, 3.0; gap length weight, 0.1. The program CLUSTAL

Fig. 1. Map of the rfb region of strain M40 and sequencing strategy. The restriction sites and the scale are as in Lee et al. (1992), but the precise map positions given here and elsewhere in this paper are based on the sequence data reported in this paper, as there were minor corrections made after sequencing. Distances from position 0 shown in kb. The ORFs inferred from the sequence are shown as boxes and the location of the gnd gene is also shown. The plasmids used for sequencing are shown with arrows showing direction of sequencing. The PstI–PstI (2.965–5.025), PstI–PstI (5.025–2.965), XhoI–XhoI (7.153–5.025), SacI–XhoI (9.850–7.153), SacI–PstI (9.90–14.26) and EcoR1–PstI (10.316–8.847) fragments were cloned into pT7T3 19U and designated pPR1251, pPR1252, pPR1254, pPR1256, pPR1257 and pPR1259, respectively. The SalI–XhoI (4.590–7.153), XhoI–EcoR1 (7.153–10.316) and PstI–SacI (12.445–9.850) fragments were cloned into pGEM-7zf(+), and designated pPR1253, pPR1255 and pPR1258, respectively. Part of pPR1257 was not sequenced and is indicated by lighter shading. B, BglII; C, ClaI; E, EcoR1; P, PstI; Pv, PvuII; Sa, SacI; S, SalI; X, XhoI.
(Higgins & Sharp, 1989) provided a good multiple alignment program for either protein or nucleic acid sequences.

Promoter search programs were generously provided by Dr M. C. O'Neill: AUTOST22 used the criteria published by O'Neill & Chiafari (1989), ALIGN2C searched for promoters on the basis of information content as described by O'Neill (1989a), and MOLBHIP used the Berg-von Hippel modified information content function described by O'Neill (1989b). The program VONHIPIC was used to calculate the index for each presumptive promoter, as described by O'Neill (1989b). A training back-propagation neural networks program (O'Neill, 1991) to define the promoters was run for us by Dr O'Neill on his computer. Promoter positions are given as the first base of the potential mRNA.

**Results**

**Sequencing the rfb region and identification of open reading frames**

The *rfb* region of strain M40 (group C1) has been located previously (Lee et al., 1992). The regions upstream and downstream of the *rfb* gene cluster have similar restriction maps to those of corresponding regions of strain LT2. The start position for our analysis of the *rfb* gene cluster of strain M40 is that used for strain LT2, as it is in the conserved region upstream of *rfb*. We sequenced from a *PstI* site at position 2965 (present in both M40 and LT2) to position 13280 (Fig. 1), which is within the *gnd* gene downstream of the *rfb* gene cluster. The sequence is presented in Fig. 2.

The sequence from bases 2965 to 4026 was very similar to the corresponding sequence of strain LT2 (Jiang et al., 1991) and the sequence from bases 13086 to 13280 was very similar to the LT2 sequence from positions 21,887 to 22,081. These regions flank the *rfb* gene cluster and had less than 2% difference in nucleotide sequence between the two groups B and C1. The 9059 bp sequence from positions 4027 to 13085 of M40 thus replaces the *rfb* cluster of strain LT2 and presumably is the *rfb* cluster of strain M40: it contains seven ORFs which were named *orf*4-11, *orf*5-19, *orf*6-17, *orf*7-17, *orf*8-36, *orf*9-79 and *orf*11-45 (Figs 1 and 2) based on their start position (kb) on the *rfb* map (e.g. *orf*4-11 starts at approximately base 4100 on the M40 *rfb* map). All ORFs read in the same direction from *orf*2-80 to *gnd* as in group B (Jiang et al., 1991).

All the ORFs started with an ATG codon and finished with a TAA codon except for *orf*7-17, which ended with a TGA codon. The sites that are the best approximations for potential Shine–Dalgarno ribosome binding sites are shown in Fig. 3. There are non-coding sequences of 384, 14 and 285 bp, respectively, before the three ORFs *orf*4-11, *orf*9-79 and *orf*11-45, whereas all other junctions have overlapping termination and initiation codons (Fig. 2 and Fig. 3). Overlaps of one base (as in *orf*4-11/*orf*5-19) and four bases (as in *orf*7-17/*orf*8-36) have already been described for polycistronic operons (Kozak, 1983). The other three ORFs show more extensive overlaps of 17 bp (*orf*5-19/*orf*6-17) and 14 bp (*orf*6-17/*orf*7-17): this kind of overlap is rather common in bacterial virus genomes (Shaw et al., 1978), but rarely found in bacterial genomes (Carlomagno et al., 1988).

**Complementation of O7 epitope expression in E. coli K12**

We previously located the *rfb* gene cluster of strain M40 by looking for clones which conferred O antigen expression on an *rfb* deletion *E. coli* strain (Lee et al., 1992).

*E. coli* does not produce O antigen as it has a mutation in *rfb*. However some *rfb* genes remain and in order to determine if the *E. coli* K12 *rfb* region could complement plasmids missing part of the M40 *rfb* region, we tested strain C600 carrying various plasmids. As can be seen (Fig. 4) the region required for group C1 O antigen expression in the presence of the K12 *rfb* region could be narrowed down to the 4.58 kb regions between the *SalI* site at base 4590 (left end of pPR1262) and the BglII site at base 9318 (right end of pPR1240). This is presumably the region required for group specificity. There are three complete ORFs (*orf*5-19, *orf*6-17 and *orf*7-17) in this region (Fig. 2), and to determine if all were required for group C1 O antigen synthesis, several subclones were constructed, transformed into *E. coli* K12 strain C600 and tested for O antigen synthesis using O7 specific antiserum (Fig. 4).

To check the requirement for *orf*5-19, we constructed pPR1250 by inserting the *SacI*-*XbaI* fragment (5498–8901) between the *SacI* and *XbaI* sites of pT7T3 19U, thereby removing 313 bp of *orf*5-19 with the remainder in frame from the polylinker. This clone did agglutinate with O7 antibody. The 14 bp *EcoRI*-*SacI* polylinker region of pPR1250 was then replaced with the 40 bp *EcoRI*-*SacI* polylinker of pYUM1118 (Futo et al., 1989), so that *orf*5-19 was read out of frame. C600 carrying the resultant plasmid pPR1289 did not agglutinate with the antibody, and as ORFs *orf*6-17 and *orf*7-17 have very good Shine–Dalgarno sequences and presumably were expressed we conclude both that *orf*5-19 cannot be complemented by *E. coli* K12 and that *orf*5-19 is functional even in the absence of the first 313 bp.

pPR1263 was constructed to remove *orf*7-17 by cloning the *SalI*-HindII fragment (4590–7720) into the *SalI*-SmaI site of pGEM-7zf (+). That this clone failed to confer O7 antigen synthesis on C600 shows that *orf*7-17 is required. We were unable to construct a derivative lacking only *orf*6-17, but clearly, when complemented by the *rfb* region of *E. coli* K12, at least two and probably three genes are required for and are sufficient for expression of the group C1 O antigen as indicated by presence of epitope O7.
Fig. 2. Sequence of part of the \textit{rfb} gene cluster of \textit{S. enterica} (M40). The \textit{PstI} site at the start of the sequence presented corresponds to the \textit{PstI} site between positions 2965 and 2970 in group B \textit{rfb} DNA (Jiang et al., 1991). Two regions are very similar to group B sequence and from 2965 to 13085. The DNA sequence, and the start of the potential transcript is indicated. The site at the start of the sequence presented corresponds to the start homology sequence, and the start of the potential transcript is indicated.
Gene        Sequence near start codon
orf4-11    ATTTGAGGCACTTAAGAGG
orf5-19    GTCAGGTTATTTACTTAAGG
orf6-17    ATTCGCTAATAAACAGTACGTAATAAAA
orf7-17    CAACTGACTTTTTATAGGAAAAAGTTA
rfbM       TCGAGCTACAGGAAAATGA
rfbK       AGGCGATACGTAATAGG
orf11-45   CTCGCGGAAGTAAAAGAGG

Fig. 3. Predicted translation initiation elements of the *rfb* genes of a group C1 *S. enterica* strain. The sequences are aligned with respect to the start ATG codons (shown in bold). Potential Shine-Dalgarno sequences are underlined and the termination codon of the preceding cistron is indicated by a bar above if it is in the sequence shown.

**Discussion**

We have sequenced the *rfb* (O antigen) gene cluster of a group C1 strain of *S. enterica* and find that there is only limited similarity to the cluster of group B (or groups A, D, E1 and C2 which are related to group B).

**GDP-Man pathway genes**

We have shown previously that genes of the mannose biosynthetic pathway lie between positions 8.86 and 11.96 kb on the *rfb* gene cluster of strain M40 (group C1) (Lee et al., 1992). *orf8-36* and *orf9-79* are at the positions where *rfbM* and *rfbK* respectively had been located. The genes were compared with those of the mannose pathway of the *rfb* regions of group B (Jiang et al., 1991) and group C2 (Brown et al., 1991, 1992), and those of the *cps* gene cluster of strain LT2 (Stevenson et al., 1991). The similarities in the sequences of the deduced polypeptides of the encoded phosphomanomutases and guanosine diphosphomannose (GDP-Man) pyrophosphorylases are shown in Table 1(a,b) and the alignments of the
mannose, N-acetylglucosamine (GlcNAc) and glucose. These three ORFs together are sufficient when complemented by the rfb region of E. coli K12 to confer O7 specificity. E. coli K12 is an rfb mutant but may retain some rfb functions. The group C1 O antigen contains mannose, N-acetylglucosamine (GlcNAc) and glucose. We know that there are mannose pathway genes functionally equivalent to rfbK and rfbM in the cps region of E. coli K12 (unpublished data) as in S. enterica LT2 (Stevenson et al., 1991), and these could substitute for rfbK and rfbM. GlcNAc is found in peptidoglycan and enterobacterial common antigen so we expect genes for synthesis of UDP-GlcNAc to be elsewhere on the chromosome, and synthesis of UDP-glucose is known to be encoded by housekeeping genes. Thus all expected biosynthetic pathway genes are accounted for in E. coli K12, and orf5·19, orf6·17 and orf7·17 presumably encode the transferases for assembly of the group C1 O antigen as these are expected to be specific for the S. enterica group C1 structure.

We can estimate how many transferases are required as the main chain of the group C1 O antigen contains two \(\alpha(1-2)\) mannosyl mannose linkages, one \(\beta(1,2)\) mannose mannose linkage, one \(\beta(1-3)\) mannosyl GlcNAc linkage and one \(\alpha(1-2)\) GlcNAc mannose linkage (Lindberg et al., 1988). One would expect a priori that there would be five transferases, one transferring the first sugar of the biological repeat unit and the others making oligomer linkages (Jann & Jann, 1984). We have assumed for the sake of economy that the transferase for the side chain glucose is not essential for O antigen polymerization and is encoded by a gene outside of rfb as this is often the case for side chain glucose (Mäkelä & Stocker, 1984).

We are left then with an expectation of five transferases but only three ORFs being sufficient to confer O antigen specificity. Unlike the situation for the other rfb gene clusters we have studied (Jiang et al., 1991; Brown et al., 1991, 1992; Wang et al., 1992), in the case of group C1 we cannot account for all expected functions with the number of ORFs present in the rfb cluster.

One possible explanation is that the polymerases are not as specific as we expect and that for example a single \(\alpha(1-2)\) mannose transferase makes both \(\alpha(1-2)\) mannosyl mannose linkages, although it would be difficult to explain how the one transferase could specify the addition of precisely two residues to give the generally accepted O antigen structure. Alternatively, perhaps the O antigen synthesized in the rfb delete strain with the three group C1 genes is incomplete but includes the O7 epitope; there are precedents for this in a mutant of the E. coli O9 rfb region which synthesizes and polymerizes a shorter repeat unit (Kido et al., 1989) and an E. coli O4 clone which expresses the O4 epitope but lacks a sugar in its O antigen (Hariguchi et al., 1991). Unfortunately, the clones we have used are unstable in Sφ874 and it was not possible to obtain sufficient O antigen for SDS-PAGE analysis to resolve this matter. Finally, it is possible that not all genes for group C1 synthesis are in the rfb cluster as for the other groups, but that some functions are provided by genes elsewhere on the chromosome, and present in both S. enterica and E. coli.

The O antigen sugar transferases are said to be membrane bound, although the sequence of the group B rfb gene cluster gave no indication of this except for the

<table>
<thead>
<tr>
<th>Table 1. Percentage amino acid and nucleotide identity among GDP-mannose pyrophosphorylases and phosphomannomutases</th>
</tr>
</thead>
<tbody>
<tr>
<td>The program BESTFIT was used to determine the alignments. We</td>
</tr>
<tr>
<td>used gap weight and gap length weight values of 30 and 0,1,</td>
</tr>
<tr>
<td>respectively. The percentage amino acid identity is shown, and</td>
</tr>
<tr>
<td>where meaningful nucleotide base identity is shown in</td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>(a) GDP-mannose pyrophosphorylases</td>
</tr>
<tr>
<td>rfbM(B)</td>
</tr>
<tr>
<td>rfbM(C1)</td>
</tr>
<tr>
<td>rfbM(B)</td>
</tr>
<tr>
<td>cpsB(B)</td>
</tr>
</tbody>
</table>

sequences are shown in Fig. 5(a,b) respectively. Dendrograms showing the relationships are given in Fig. 6. DNA sequence similarities were barely detectable except for those with amino acid identities above about 70%.

The low levels of identity show why rfbM(C1) did not hybridize with rfbM(B) or cpsB(B) by Southern hybridization (Lee et al., 1992). These data also clearly show that rfbK of group C1 is very closely related to the rfbK(B) and rfbK(C2) respectively. The percentage amino acid identity is shown, and where meaningful nucleotide base identity is shown in parentheses. We can estimate how many transferases are required as the main chain of the group C1 O antigen contains two \(\alpha(1-2)\) mannosyl mannose linkages, one \(\beta(1,2)\) mannose mannose linkage, one \(\beta(1-3)\) mannosyl GlcNAc linkage and one \(\alpha(1-2)\) GlcNAc mannose linkage (Lindberg et al., 1988). One would expect a priori that there would be five transferases, one transferring the first sugar of the biological repeat unit and the others making oligomer linkages (Jann & Jann, 1984). We have assumed for the sake of economy that the transferase for the side chain glucose is not essential for O antigen polymerization and is encoded by a gene outside of rfb as this is often the case for side chain glucose (Mäkelä & Stocker, 1984).

We are left then with an expectation of five transferases but only three ORFs being sufficient to confer O antigen specificity. Unlike the situation for the other rfb gene clusters we have studied (Jiang et al., 1991; Brown et al., 1991, 1992; Wang et al., 1992), in the case of group C1 we cannot account for all expected functions with the number of ORFs present in the rfb cluster.

One possible explanation is that the polymerases are not as specific as we expect and that for example a single \(\alpha(1-2)\) mannose transferase makes both \(\alpha(1-2)\) mannosyl mannose linkages, although it would be difficult to explain how the one transferase could specify the addition of precisely two residues to give the generally accepted O antigen structure. Alternatively, perhaps the O antigen synthesized in the rfb delete strain with the three group C1 genes is incomplete but includes the O7 epitope; there are precedents for this in a mutant of the E. coli O9 rfb region which synthesizes and polymerizes a shorter repeat unit (Kido et al., 1989) and an E. coli O4 clone which expresses the O4 epitope but lacks a sugar in its O antigen (Hariguchi et al., 1991). Unfortunately, the clones we have used are unstable in Sφ874 and it was not possible to obtain sufficient O antigen for SDS-PAGE analysis to resolve this matter. Finally, it is possible that not all genes for group C1 synthesis are in the rfb cluster as for the other groups, but that some functions are provided by genes elsewhere on the chromosome, and present in both S. enterica and E. coli.

The O antigen sugar transferases are said to be membrane bound, although the sequence of the group B rfb gene cluster gave no indication of this except for the
Fig. 5. Relationships of the amino acid sequence of protein encoded by (a) GDP-Man pyrophosphorylases and (b) phosphomannomutases. The sequences were aligned using the computer program CLUSTAL. In these comparisons, the gap penalty was set at 3. Asterisks indicate amino acids identical in all four sequences and dots those similar in all four.
galactose transferase (Jiang et al., 1991). However, in this case analysis of the three ORFs revealed that orf5'19 and orf6'17 have each one potential transmembrane segment, but orf7'17 has none. The sequences of the three ORFs were used to search for any similarities with the other sequenced genes in the serogroup. The scale indicates percentage identity of amino acids after alignment using BESTFIT.

orf11'45

The amino acid sequence of the protein encoded by orf11'45 has 12 possible transmembrane segments distributed throughout its length. We have seen such a protein in S. enterica rfb gene clusters of strains of groups B, C1 and C2 and the S. enterica cps gene cluster. The proteins are referred to by their gene symbols and serogroup. The scale indicates percentage identity of amino acids after alignment using BESTFIT.

orf4'11

The group C1 O antigen is polymerized in C600 or Sdp874 carrying pPR1042 (data not shown), which indicates that if group C1 has an O antigen polymerase (see introduction), then the rfc gene is in the rfb cluster.

The rfc genes of groups B and C2, and the putative rfc gene of group E1 have similar predicted structures with 11 transmembrane segments (Brown et al., 1992; Collins & Hacket, 1991; Wang et al., 1992). orf4'11 has a similar predicted structure and is the only ORF to do so. Thus if group C1 does have an rfc gene it is most likely orf4'11. The level of sequence similarity between presumed rfc genes is low in all cases and orf4'11 is not out of line. We constructed clones either with or without orf4'11 but instability has thus far prevented us from obtaining convincing evidence for rfc function of orf4'11.

Note that if group C1 O antigen is extended by single sugar extension, without involvement of an O antigen polymerase, this will not affect the total number of genes expected as in this case the O antigen chain will be started by addition of the first sugar to a carrier, which will require a specific enzyme as this step is different to subsequent additions of that sugar to partially completed O antigen by single sugar extension. There will thus be no effect on the argument given above on the expected number of genes.

Promoters

Our earlier studies (Jiang et al., 1991) indicated potential promoters in the intergenic regions before orf2'80 and rfbB respectively of group B, but the modified Berg–von Hippel indices (O’Neill, 1989b) were such that the promoters were not likely to function without help. This region is also present in strain M40 (Lee et al., 1992) and the sequence of the promoter before rfbB at position 3928 is the same in group C1 as in group B (Fig. 2).

Computer searches using the three programs provided by Dr. M. C. O’Neill (O’Neill & Chiafari, 1989; O’Neill, 1989a, b) and a training back-propagation neural networks program (O’Neill, 1991) revealed a large number of potential promoters but the best had a modified Berg–von Hippel index of 3.88, and again none are expected to function without help.

The genes from orf4'11 to rfbK are overlapping such that translational readthrough is possible, except for the gap of 14 bp before rfbK, which includes a good potential Shine–Dalgarno sequence (Fig. 3). It appears that these
genes comprise a single operon. There is a 285 bp gap between \( \text{rfbK} \) and \( \text{orf11-45} \) and \( \text{orf11-45} \) may or may not be in this operon. There are candidate promoters at positions 11286, 11382 and 11434 (Fig. 2) with modified Berg–von Hippel indices of 8:23, 8:34 and 5:28 respectively, but none of them would be expected to function without positive control, and identification of a promoter will require direct analysis of promoter function.

### G+C content and codon usage

The G+C content of the \( \text{rfb} \) region of strain M40 is generally low (Fig. 7 and Table 2), with discrete regions of different G+C content: \( \text{orf4-11}, \text{orf5-19}, \text{orf6-17}, \text{orf7-17} \) and \( \text{orf11-45} \) have a G+C content of about 0.30, \( \text{rfbM} \) a G+C content of 0.39 and \( \text{rfbK} \) a G+C content of 0.61 (Table 2 and Fig. 7). The regions flanking the \( \text{rfb} \) region in strain LT2 are of typical \( \text{S. enterica} \) G+C content and codon usage (Jiang et al., 1991) and these regions are similar in strain M40. As none of the \( \text{rfb} \) genes of strain M40 have a G+C content appropriate for \( \text{S. enterica} \) genes, they must all have been acquired by lateral transfer, as proposed for the \( \text{rfb} \) genes of group B strain LT2 (Jiang et al., 1991). The properties of \( \text{orf4-11}, \text{orf5-19}, \text{orf6-17}, \text{orf7-17} \) and \( \text{orf11-45} \) are typical of those of a low G+C content organism, with average P1, P2 and P3 values (Sueoka, 1988) of 0.37, 0.28 and 0.24 respectively. The low value of P3 in particular suggests that this region has not been in \( \text{S. enterica} \) for long. These properties are similar to those of the central region of the \( \text{rfb} \) gene cluster of strain LT2 (Jiang et al., 1991), and suggest a similar source for both sets of genes.

### \( \text{rfbK} \) arose by gene duplication of \( \text{rfbM} \)

\( \text{rfbM}(\text{C1}) \) has a G+C content of 0.39 and P1, P2 and P3 values of 0.51, 0.37 and 0.28 respectively, which is similar to values for \( \text{rfbM} \) and \( \text{rfbK} \) of strain LT2 with 0.4 and 0.409 G+C content and P1, P2 and P3 of 0.55 and 0.517, 0.37 and 0.431, and 0.30 and 0.295 respectively (Jiang et al., 1991). The \( \text{rfbM} \) gene of strain M40 is the gene of the \( \text{rfb} \) gene cluster most similar to a gene in the \( \text{rfb} \) region of strain LT2, but it has diverged substantially. Both \( \text{rfbM} \) genes may have been derived originally from the same low G+C content organism, and must have undergone most of their divergence in such a low G+C content species.

\( \text{rfbK}(\text{C1}) \) has a G+C content of 0.61 and P1, P2 and P3 values of 0.63, 0.42 and 0.78 respectively, and is very similar (96.4%) to the \( \text{cpsG} \) gene of strain LT2. The tree for these \( \text{rfb} \) and \( \text{cps} \) genes is shown in Fig. 6: the phosphomannomutase genes seem to evolve significantly faster than the GDP-Man pyrophosphorylase genes, based on the differences between the genes of groups C2 and B, and between that pair of strains and the \( \text{cpsG}/\text{rfbK}(\text{C1}) \) pair. However, the 3% amino acid difference between \( \text{rfbK}(\text{C1}) \) and \( \text{cpsG} \) is very anomalous compared with the 17% difference between \( \text{rfbM}(\text{C1}) \), and \( \text{cpsB}, \text{cpsG} \) and \( \text{rfbK}(\text{C1}) \) are almost identical except for the last 28 bp and presumably arose by gene duplication. \( \text{cpsG} \) is part of the \( \text{cps} \) gene cluster for the widespread M antigen and was presumably the parent gene, with \( \text{rfbK} \) of strain M40 arising via gene duplication in a relatively recent event. Strain M40 retains its \( \text{cpsG} \) gene as shown by Southern blotting but we have not checked to see if it is expressed.

The pair of \( \text{cps} \) genes must have been present over a very long period in a high G+C species as their DNA is typical of 0.61 G+C content DNA (Stevenson et al., 1991). The evolution of the group B and group C2 \( \text{rfbM} \) and \( \text{rfbK} \) genes, and the group C1 \( \text{rfbM} \) gene appears to have been in a 0.4 G+C content species, with relatively recent transfer to \( \text{S. enterica} \). It seems most likely that the group C1 gene cluster at the time of transfer to \( \text{S. enterica} \) would have included both \( \text{rfbM} \) and \( \text{rfbK} \) genes, and we can only speculate as to how the phosphomannomutase gene of the \( \text{cps} \) gene cluster came to replace the original

---

Table 2. G+C content and P1, P2 and P3 values for the genes and ORFs of the \( \text{rfb} \) gene cluster

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>G+C</th>
</tr>
</thead>
<tbody>
<tr>
<td>orf4-11</td>
<td>0.32</td>
<td>0.30</td>
<td>0.24</td>
<td>0.29</td>
</tr>
<tr>
<td>orf5-19</td>
<td>0.37</td>
<td>0.30</td>
<td>0.23</td>
<td>0.30</td>
</tr>
<tr>
<td>orf6-17</td>
<td>0.39</td>
<td>0.27</td>
<td>0.24</td>
<td>0.30</td>
</tr>
<tr>
<td>orf7-17</td>
<td>0.39</td>
<td>0.28</td>
<td>0.25</td>
<td>0.31</td>
</tr>
<tr>
<td>rfbM</td>
<td>0.51</td>
<td>0.37</td>
<td>0.28</td>
<td>0.39</td>
</tr>
<tr>
<td>rfbK</td>
<td>0.63</td>
<td>0.42</td>
<td>0.78</td>
<td>0.61</td>
</tr>
<tr>
<td>orf11-45</td>
<td>0.36</td>
<td>0.27</td>
<td>0.25</td>
<td>0.30</td>
</tr>
</tbody>
</table>

---

Fig. 7. G+C content of the \( \text{rfb} \) region of \( \text{S. enterica} \) group C1. The G+C content for the \( \text{rfb} \) genes and a few hundred bases of flanking DNA is from sequence presented in this paper (Fig. 2), and for the remainder of the flanking region is based on the sequence of a group B strain thought to have essentially the same sequence in these regions [Jiang et al. (1991) for \( \text{orf4-6} \) to \( \text{orf2-8} \) and Reeves & Stevenson (1991) for \( \text{gnd} \)].
rfbK gene, and thus become the rfbK gene for the C1 group. One possibility is that on transfer, or perhaps later, the rfbK gene suffered mutation or loss but that the gene cluster nonetheless functioned because the cpsG gene complemented the defect: gene duplication of cpsG and translocation of one copy by recombination could then have given the arrangement currently observed. If this indeed is what occurred, then recombination events to effect the replacement must have been very close to the ends of the genes. The terminal 28 bp of rfbK is of 0·3 G+C content and this may represent the end of the original low G+C content rfbK gene.

In summary, the group C1 rfb region has seven genes, arranged in an operon like manner. Of these genes, two are mannose biosynthetic pathway genes, one encodes a presumptive O antigen polymerase and four remain unidentified. One of the unidentified genes appears to encode a membrane protein and resembles genes found presumptive O antigen polymerase and four remain unidentified. The whole cluster is derived from a low G+C content organism except for rfbK which appears to be a recent copy of the cpsG gene.

We would like to thank Dr O’Neill for carrying out promoter searches for us. The work was supported by a grant from the Australian Research Council.

References


