Comparison of Lipopolysaccharide Biosynthesis Genes *rfaK*, *rfaL*, *rfaY*, and *rfaZ* of *Escherichia coli* K-12 and *Salmonella typhimurium*

JOHN D. KLENA, ELIZABETH PRADEL, AND CARL A. SCHNAITMAN*

Department of Microbiology, Arizona State University, Tempe, Arizona 85287-2701

Received 10 March 1992/Accepted 18 May 1992

Analysis of the sequence of a 4.3-kb region downstream of *rfaJ* revealed four genes. The first two of these, which encode proteins of 27,441 and 32,890 Da, were identified as *rfaY* and *rfaZ* by homology of the derived protein sequences of their products to the products of similar genes of *Salmonella typhimurium*. The amino acid sequences of proteins RfaY and RfaZ showed, respectively, 70 and 72% identity. Genes 3 and 4 were identified as *rfaK* and *rfaL* on the basis of size and position, but the derived amino acid sequences of the products of these genes showed very little similarity (about 12% identity) between *Escherichia coli* K-12 and *S. typhimurium*. The next gene in the cluster, *rfaC*, encodes a product which also shows strong protein sequence homology between *E. coli* K-12 and *S. typhimurium*, as do the *rfaF* and *rfaD* genes which lie beyond it. Thus, the *rfa* gene cluster appears to consist of two blocks of genes which are conserved flanking a central region of two genes which are not conserved between these species. Although the RfaL protein sequence is not conserved, hydropathy plots of the two RfaL species are nearly identical and indicate that this is a typical integral membrane protein with 10 or more potential transmembrane domains. We noted the similarity of the structure of the *rfa* gene cluster to that of the *rfb* gene cluster, which has now been sequenced in several *Salmonella* serovars. The *rfb* cluster also contains a gene which lies within a central nonconserved region and encodes an integral membrane protein similar to protein RfaL. We speculate that protein RfaL may interact in a strain- or species-specific way with one or more Rfb proteins in the expression of surface O antigen.

rfaYZKL is a block of genes within the *rfa* cluster (23) which is thought to be involved in completion of the outer region of the lipopolysaccharide (LPS) core and attachment of O antigen. The gene *rfaK* is defined by a set of mutants in *Salmonella typhimurium* and *S. minnesota* (14) which are resistant to bacteriophage Felix O, sensitive to one or more rough phenotype-specific phages, and lacking most but not all of O antigen. These mutants have a normal composition with respect to the core sugars glucose and galactose but lack partial substitution of the nonreducing terminal glucose residue by *N*-acetylglucosamine (14, 22). Although the properties of these mutants suggest that *rfaK* encodes *N*-acetylglucosamine transferase, this has not been proven (14). The fact that *rfaK* mutants produce small but detectable amounts of LPS molecules containing O antigen has been interpreted as indicating that a core substituted with *N*-acetylglucosamine is the preferred substrate for addition of O antigen but that this stringency is not absolute (14). The *rfaL* gene is also necessary for production of LPS containing O antigen but has not been shown to play a role in core synthesis. *rfaL* mutants of *S. typhimurium* have a distinctive phenotype with respect to other *rfa* mutants in that they lack O antigen but are sensitive to phage Felix O, as well as the rough phenotype-specific phages, and they produce an LPS which has the serological and chemical characteristics of a complete core (14). These mutants are not defective in O antigen synthesis, since they produce O antigen hapten linked to antigen carrier lipid. Other mutations that produce this phenotype map to the *rfb* gene cluster. It has been suggested that *rfaL* encodes a component of the O ligase which

translocates the polymeric O antigen from antigen carrier lipid on which it is assembled to the LPS core (14, 22).

Genes *rfaY* and *rfaZ* do not correspond to any genetic or biochemical functions which have been previously described in *Salmonella* spp. They were first identified as genes that encode open reading frames lying upstream from *rfaK* in *S. typhimurium* (13). These were recently designated as *rfa* genes (23) on the basis of their location; on the basis of the finding, described in this report, that the reading frames are strongly conserved between *Escherichia coli* K-12 and *S. typhimurium*; and on the basis of preliminary evidence from complementation experiments which indicates that the *rfaYZ* region is involved in core completion (16).

Most features of the *rfa* cluster appear to be conserved between *E. coli* K-12 and *S. typhimurium*. The physical structure of the *rfa* gene cluster is similar in both organisms (23), and the *rfaH* (*sfrB*) gene product functions interchangeably between the two organisms to regulate expression of the long *rfa* operon containing the genes for the hexose region of the core (5, 19, 21). Restriction fragments from *E. coli* K-12 carrying genes *rfaG*, *rfaP*, and *rfaB* are capable of efficiently complementing corresponding mutations in *S. typhimurium* (16, 17). The protein products of the *rfaI* and *rfaJ* genes of *E. coli* K-12 and *S. typhimurium* share substantial regions of homology, particularly surprising in view of the fact that they are thought to be sugar transferases of different specificities, and restriction fragments from *E. coli* K-12 which include both *rfaI* and *rfaJ* can efficiently complement either *rfaI* or *rfaJ* mutations of *S. typhimurium* (18).

In contrast to the results cited above, which demonstrate conservation of most of the *rfa* gene cluster across species lines, we show here that the sequences of the proteins encoded by *rfaL* and *rfaK* are poorly conserved between *E. coli* K-12 and *S. typhimurium*. In a separate report (11), it

* Corresponding author.

FIG. 1. Physical map of an *EcoRI*-*AvaI* restriction fragment from the left (*cysE*) end of the *rfa* gene cluster. The arrows at the top line indicate the locations and directions of transcription of the coding regions for the genes from *rfaD* through part of *rfaB*. The middle line shows a restriction map, and the lower portion shows all or part of the restriction fragments described in the text. The restriction fragments are designated as previously described (17, 18), by a single-letter code for restriction sites at their ends and a number indicating size in kilobases. The single-letter restriction site code is as follows: A, *AccI*; B, *BamHI*; C, *ClaI*; E, *EcoRI*; F, *HincII*; G, *BglII*; H, *HindIII*; K, *ScaI*; L, *BclI*; N, *NcoI*; Q, *BalI*; R, *EcoRV*; S, *SalI*; T, *BstXI*; V, *AvaI*.

will be shown that the *rfaK* and *rfaL* products from *E. coli* K-12 and *S. typhimurium* are functionally different and that restriction fragments from *E. coli* K-12 carrying *rfaK* and *rfaL* cannot fully complement mutations in the corresponding genes of *S. typhimurium*.

MATERIALS AND METHODS

Cloning of the left half of the *rfa* gene cluster. Figure 1 shows a restriction map of a 10.5-kb fragment extending from an *EcoRI* site between *kbl* and *rfaD* (23) to an *AvaI* site in *rfaB* (18). To provide both a useful restriction site and a selectable marker for cloning of this region, a drug resistance cassette flanked on one side by an *EcoRI* site was introduced into this *AvaI* site in the *E. coli* K-12 chromosome. This construction began with a 5.8-kb *rfa* fragment extending from the *SalI* site in *kdtA* to the *AvaI* site in *rfaB* (18) in which both the *SalI* and *AvaI* sites had been blunt ended and converted to *EcoRI* sites by ligation with *EcoRI* linkers. This fragment was inserted into the *EcoRI* site of pGEM3 (Promega, Madison, Wis.) in the orientation in which the polylinker was adjacent to *rfaB*'. The approximately 2.3-kb fragment from this plasmid which extended from the *AvaI* site in the pGEM polylinker to the *BglII* site in *rfaP* was then inserted into the internal *BglII* and *AvaI* sites of a pGEM3 plasmid containing the SG 9.0 fragment (18) which extends from the *SalI* site in *kdtA* to a hybrid *BamHI*-*BglII* site derived from the *BglII* site in *rfaY*. This resulted in the introduction of an *EcoRI* site just 5' to the *rfaB* *AvaI* site which was separated from the *AvaI* site by a short sequence consisting of the *AvaI*-*EcoRI* segment of the pGEM3 polylinker. The *AvaI* site of this construct was opened and blunt ended and converted to a *BamHI* site, and the 2.3-kb *BamHI* fragment carrying the Ω Km^r cassette (7) was ligated into this *BamHI* site. The plasmid carrying this construction was linearized and crossed onto the chromosome of a *recBC sbcB* strain as previously described (19) by selecting for kanamycin resistance. The chromosomal construction was then crossed into CS180 (16) by P1 transduction to generate strain CS1902 (*rfaB*:: Ω Km^r *EcoRI*). This construction was verified by its *rfa* phenotype and by Southern blotting (data not shown).

Chromosomal DNA from CS1902 was cleaved with *EcoRI*

and ligated into the *EcoRI* site of low-copy-number plasmid pRK415-1 (6). This DNA was introduced into NEM259 (from N. Murray) by electroporation (17). A transformant carrying the *rfaDFCLKZYJIB*' insert was obtained by selecting for kanamycin resistance, and the insert was determined to be correct by its restriction map and its ability to complement an *rfaC* mutant of *S. typhimurium*. A 5.5-kb *BglII* fragment carrying *rfaFCLKZY*' (Fig. 1, GG 5.5) was subcloned from this plasmid into the *BamHI* site of pGEM4.

DNA sequencing and analysis. DNA sequencing and analysis with PC/GENE were done as previously described (17, 18).

Nucleotide sequence accession number. The sequence reported here has been submitted to GenBank and assigned accession no. M95398.

RESULTS

Sequence of the *rfaYZKL* region. Figure 2 shows the sequence of a 4.3-kb region of the *rfa* cluster (23) which includes *rfaYZKL*. The properties of the four genes and the protein products derived from the sequence are shown in Table 1. The *rfaZ*, *rfaK*, and *rfaL* genes of *S. typhimurium* and their products (13) are included for comparison. As found for most other *rfa* gene products (13, 17, 18, 24), the proteins are basic and the RfaL protein, with isoelectric points of 9.8 in *E. coli* and 9.7 in *S. typhimurium*, and the RfaY protein, with an isoelectric point of 10.2, are the most strongly basic Rfa proteins which have been found. There is reasonable agreement between the molecular weights of the proteins of the two organisms.

As has been noted for other *rfa* genes (13, 17, 18) and for genes of the *rfb* cluster (10), these genes have an A+T content which is significantly higher than the approximately 50% A+T content which is the average of the whole genomes of *E. coli* K-12 and *S. typhimurium*. In the case of the *rfb* region, the high A+T content has been interpreted to indicate that these genes were transferred to the enteric bacteria from an ancestor with a genome which was A+T rich (3, 10). There is roughly a 10% difference in A+T content between the *rfaK* and *rfaL* genes of *E. coli* K-12 and *S. typhimurium*. This is interesting in view of the data presented below on the lack of homology between the RfaK and RfaL proteins of the two organisms. If the high A+T content is indicative of an ancestral transfer of genes, the difference in A+T content suggests that these genes were transferred at different times or from different organisms.

With the exceptions of the short region of homology between proteins RfaK and RfaG (17) and the homology between the four *E. coli* K-12 genes and their counterparts in *S. typhimurium*, the derived protein sequences of these four *E. coli* K-12 genes showed no significant homology to other proteins in the data base available on PC/GENE, release no. 6.60. Hydropathy plots of RfaY, RfaZ, and RfaK showed no significant potential transmembrane sequences, indicating that these, like most other Rfa proteins, are likely to be cytoplasmic proteins or peripheral membrane proteins on the inner face of the cytoplasmic membrane.

In contrast, a hydropathy plot of protein RfaL from both organisms (Fig. 3) is indicative of an integral membrane protein, with 10 or more potential membrane-spanning regions. Some of these definitely span the cytoplasmic membrane, since we have been able to isolate several blue (alkaline phosphatase active; 15) *TnphoA* protein fusions within the *rfaL* gene (data not shown), while fusions to *rfa* genes other than *kdtA* (4, 18) gave only white colonies. The

9830 9840 9850 9860 9870 9880 9890 9900 9910
GAAGATAAATAAAGAAATTATTGGACATCAAGAAATAGTCATTATTAATCCATAGGTGCAAAAATAATCGCGCTTACGTTTCGAG
GluAspLysIleValGluPheIleGlyAspThrArgIleValIleLeuAsnProGluGlyAlaLysIleCysArgLeuThrPheGlu

9920 9930 9940 9950 9960 9970 9980 9990 10000
CAATAATGCTTATATATCAAGAAGTCAAAACACATCTGAAATATTCAGGATATTATACCGGGTACCGCAAGCAATTATGCACATA
GlnIleLysValIleTyrGlnGluValLysThrHisPheGluAsnTyrArgIlePheThrGlyLeuProGlnAspLeuLeuThrLe

10010 10020 10030 10040 10050 10060 10070 10080 10090
CCAAATGCTCAAGATCGACACAGCTATGCATGATTTATTAACGGTTGGCTTCCCAAGTATAGCTATTTTCGATCTTCCTGGT
ProIleLeuGluLeuGlyThrLeuProPheAspGluPheIleTyrThrValAlaLeuThrLysTyrSerAspPheValIleSerValAsp

10100 10110 10120 10130 10140 10150 10160 10170 10180
ACACGCTTGTTTCATATGGCTGGCGGCTTATAAACAACCAAGCTGCATTTATCCGCTAATTCAGCTATCCGGGAATATCCCTCGCATTA
ThrAlaLeuHisMetAlaHisMetAlaLeuTyrHisLysProThrLeuLeuPheThrProAsnSerArgThrProGluTyrProIleGlyLeu

10190 10200 10210 10220 10230 10240 10250 10260 10270
ATTGGTCCCGGAATCATATAAATCATAGCAATATCCCAACATATACGGTAAGAATATGCATATGCAAACTTAACTAATCTTA
IleTyrPseProAsnHisHisLysSerIleGlnIleValSerProThrTyrThrValLysAspIleAspThrGluThrLeuThrAsnSer

10280 10290 10300 10310 10320 10330 10340 10350 10360
GTAAAGAGCTTAAGTGTATATAGAAAGAGCTGACTTTTAACCTGCTTAACCGTTATTATTCCTAATAATGTATGTTTACAGGATT
ValLysArgLeuSerCysCysIleAspLysLys---End rfaK End rfaL---AsnIleThrAsnAsnArgAsnAsn

10370 10380 10390 10400 10410 10420 10430 10440 10450
TAATGACGAGTAAAGCATATACCGGATATATGATATATTCGAACTCTGCTGCCATATGATCATACTAGTACCTCAATCCCAACA
IleValLeuLeuLeuValIleAlaSerIleIleLeuPheLeuSerArgAlaThrPheIleValAspSerLeuGlyIleGlyValIle

10460 10470 10480 10490 10500 10510 10520 10530 10540
TGGCAAGCGTTAATATCAACAACCCAAAGCTGCTTTTATATGCTGTATATAAATAGAAAATAGAGGAAAAGCTGGAAAAGCTT
GlyLeuThrLeuIleLeuLeuGlyLeuAlaArgLysLysTyrAlaIleTyrPheLeuSerPheTyrLeuThrThrThrPheIleGly

10550 10560 10570 10580 10590 10600 10610 10620 10630
CCATCAGACCTTTCAGTACGCTCTGTCATTAATCTCATATTATGATAGATAGCTAGAGAAATCCAACTGCCCTCTAGGCTATTGCTT
MetLeuGlyLysLeuSerGlyArgGluIleIleGluAsnHisLeuHisValAsnSerPheGluLeuAlaGlyArgLeuArgAsnHisGlu

10640 10650 10660 10670 10680 10690 10700 10710 10720
CTGCACATAACAATCATCTTTCAGCGCGTCACTGCTGATATAGAAAGAAAGCTGACTTTATGAATATTTAAACCAATTCTGT
AlaValLeuLeuAsnMetSerGluAlaArgSerGluAlaSerThrPheSerPheProSerLysIlePheIleAsnLeuGlyIleGlyLeu

10730 10740 10750 10760 10770 10780 10790 10800 10810
ACATTCCAGCTTTCGACATAGGGAAGTACAGCTATTAGCATGGTATACACTGTTTAAGTCAATTAACTGCTTCATTATAGCGGATCTGTA
MetAlaLeuLeuAlaCysLeuSerThrValSerAlaAsnAsnLeuTyrSerSerAsnAsnAsnLeuAlaGluAsnTyrArgAsnGlnIle

10820 10830 10840 10850 10860 10870 10880 10890 10900
TTGCTTTTAAATATATAACAATCTAGCTATATATACCAATTATAGCAATATAGCAAGGCAATATCTTGGCTGATTTATTTATTAAT
ProLysAsnPheIleIleValIleSerAlaLeuIleAlaIleLeuValIleSerAsnThrPheLysTyrProGlySerThrLeu

10910 10920 10930 10940 10950 10960 10970 10980 10990
AAGCTTATTAATCAGCACAACAATATAGGCAACAGAGTAGCGGCTGCTTGGTTTGTCTTAGCGGCAGAACATAAAGTACCGCCGCAAC
AlaIleLeuAlaAlaValLeuLysTyrProValProThrProAsnSerGlnAsnThrAsnAlaCysSerCysLeuThrGlyAlaCysSer

11000 11010 11020 11030 11040 11050 11060 11070 11080
TATTTATAAAAAATAAAGGATGATTTTCTTGATATAAAGATCCGACGCCGCACTATCCGCTATAGCACTTGTGAATCTGCTG
AsnLeuLeuPheLeuPheProHisAsnLysLysTyrThrLeuIleAlaValCysLysValIleGlyIleLeuMetThrTyrSerAlaAla

11090 11100 11110 11120 11130 11140 11150 11160 11170
CTGCTTTCGGGTTCTCAGACCAAAAGAAAGCGGGTCAATTTCATAGTCTTATTAATATACAGTTCGATCCGCAATATAGAAATGACA
GlyThrAlaThrIleGlyCysPheSerIleArgAspAsnGluHisIleSerAsnIleTyrMetAlaIleGlyAlaThrPheAsnLeu

11180 11190 11200 11210 11220 11230 11240 11250 11260
GAGATTAACAAGTCTATAATACAGCTCTTTTTTGATTTAGCTGGCTAGTACTGTCCAGAAACCATAAAGACGCAAAATATAAAT
SerTyrThrThrTyrLeuLeuPheLeuLeuLeuGlnSerThrLeuThrLeuPheLeuValIlePheSerGlyPheIlePheIle

11270 11280 11290 11300 11310 11320 11330 11340 11350
TTTTGGCAGCTATTAATACACTATGTATAGTACAGCAAAATGCGGAATATTACGTTATAAAGCGGATACCAAAATTAATACCAAGCA
LysAlaThrAsnLeuTyrSerHisTyrThrAlaArgPheProSerAsnAspValLysPheAlaSerTyrTrpIleLeuAspLeuLeuGly

11360 11370 11380 11390 11400 11410 11420 11430 11440
CTATTAAATATAGAGAGCGGAATATAGCTTTTATATATTAATTTCTGCTGGCAGCATAAATAGTACAAATAGCAATGATA
IleLeuPheIleSerLeuProLeuIleLeuAsnLysIleAsnTyrAsnGluGlnArgGlyArgLeuIleLeuSerLeuLeuValIle

11450 11460 11470 11480 11490 11500 11510 11520 11530
TAGCAGTGATATTATATTATTATGCTCAAGGATATTATGACGAAATGCAATATGATCATGAAAAAACATATAATTAAGTATTATATG
AlaThrIleAsnTyrIleIleLysSerPheThrAsnAspValPheAlaIleIleMetSerPhePheCysLeuLysTyrIleIleGlu

11540 11550 11560 11570 11580 11590 11600 11610 11620
CTAAATCTATGATGTTTCAGAGTCAAGGTTCAAGTCAAGAACTTAAGAGTATTCAGATGTTTACTGTTATATCATGATCACT
LeuIleMetSerSerLysLeuThrTyrProLysLeuSerHisLeuLysPheSerThrLeuMetC---rfaL---LeuSerSerLeuLys

11630 11640 11650 11660 11670 11680 11690 11700 11710
TTTCCAAACCTGCTGACTTTCAGGATTTAAATTTCTTGCTGGCGGCTAGATACCATCTGATTTCCGATACCGCAAAATTAAT
GluLeuValAlaGlnSerAsnLeuAsnIleLeuAsnGluProAlaArgCysValMetMetGlnAsnLysTyrGlyCylGlyIleLeuGly

11720 11730 11740 11750 11760 11770 11780
CCGAGTCGGTGTGCTAAACAGCGGATATGGCTGTATACGTCGGCGGCTTAAATGGCTTAACCGGCTATCC

ProAspThrProGlyTyrThrValIleThrAsnProAspAlaSerAlaGlnHisSerLeuS---rfaC

coli K-12 and *S. typhimurium*. Figures 4 and 5 show comparisons of the derived protein sequences of these genes. In an examination of the proteins of the *rfaYZK* region (Fig. 4), the portion of RfaY which is known for both organisms and RfaZ (excluding the N-terminal 29 amino acids) show significant

TABLE 1. Properties of *rfaY*, *rfaZ*, *rfaK*, and *rfaL* of *E. coli* K-12 and *S. typhimurium*

Organism and gene	% A+T	Protein properties	
		pI	Mol wt
<i>E. coli</i>			
<i>rfaY</i>	65	10.2	27,441
<i>rfaZ</i>	67	9.5	32,890
<i>rfaK</i>	69	8.7	41,619
<i>rfaL</i>	68	9.8	46,683
<i>S. typhimurium</i>			
<i>rfaZ</i>	62	8.7	31,003
<i>rfaK</i>	59	9.2	43,152
<i>rfaL</i>	57	9.7	46,034

regions of homology, with 70 and 72%, respectively, of the residues identical. These regions of homology terminate abruptly after residue 254 of *E. coli* K-12 RfaZ, and the last 29 residues of RfaZ and all of RfaK show few significant regions of homology, with only about 12% of the residues identical. The potentially identical residues are scattered throughout the RfaK protein. An almost identical situation is seen when the sequence extending from RfaC into RfaL is examined (Fig. 5). The RfaC proteins from *E. coli* K-12 and *S. typhimurium* show numerous regions of homology throughout their lengths (data not shown) until just before their N termini, and the short segment of RfaC from both organisms shown in Fig. 5 contains a remarkable segment in which 31 of 32 residues are identical. This region of homology also ends abruptly, 26 residues before the N terminus of *E. coli* K-12 RfaC, and the two RfaL proteins show only a barely significant patchy region of homology, with about 13% of the residues identical. We interpret the abrupt appearance of regions of low sequence identity at the ends of RfaZ and RfaC as junctions or sites of illegitimate recombination where genes or block of genes have been exchanged. The *rfa* cluster appears to consist of sequence blocks which differ significantly in A+T content (Table 1 and references 13, 17, and 18). This has also been observed for the *rfb* cluster and has been proposed to reflect an evolutionary origin of the cluster in which blocks of genes were exchanged (3, 10, 12).

It is thought that protein RfaL functions in attachment to the LPS core of O antigen which is synthesized and polymerized on the dolichol analog antigen carrier lipid. Therefore, the RfaL sequence was scanned visually for possible homology to the consensus dolichol-binding sequence found in *Saccharomyces cerevisiae* sugar transferases (1) and reported near the C terminus of enzymes involved in K1 capsular polysaccharide synthesis (26). Potential sequences near the C termini of the two RfaL proteins are indicated by underlining in Fig. 5, although it should be noted that these are not strong matches.

Comparison of sequences provided by other laboratories (20, 24, 25, 28) for the left end of the *rfa* cluster (data not shown) indicates that proteins RfaD and RfaF resemble proteins RfaC, RfaY, and RfaZ in exhibiting numerous regions of homology (more than 70% identical residues) between the two organisms.

DISCUSSION

The left (*cysE*) end of the *rfa* locus, which contains the *rfaDFC* genes involved in biosynthesis and attachment of

FIG. 3. Hydrophobicity plots of the RfaL proteins of *S. typhimurium* and *E. coli* K-12. The scale at the left indicates relative hydrophobicity, and positive values indicate increased hydrophobicity.

heptose, appears to be strongly conserved between *E. coli* K-12 and *S. typhimurium*. The derived amino acid sequences show more than 70% identity, and restriction fragments carrying these genes from *E. coli* are effective in complementing mutations in the corresponding genes in *S. typhimurium* (11). The block of genes to the right of the 3' end of *rfaZ* which are involved in synthesis and completion of the hexose region of the core also appear to be strongly conserved between these two organisms. *rfaZ* and most of *rfaY* have been sequenced in both organisms, and the derived amino acid sequences of proteins RfaY and RfaZ show identity of 70% or more. The proteins encoded by *rfaI* and *rfaJ* also show substantial sequence similarity between these organisms (18). The sequence for the *S. typhimurium* genes to the right of the *rfaIJ* region has not been reported, but the efficient complementation of *S. typhimurium* *rfaP*, *rfaG*, and *rfaB* mutations by restriction fragments carrying these genes from *E. coli* K-12 suggests that the proteins encoded by this region are also likely to be very similar. In contrast, the *rfaK* and *rfaL* genes which lie between these conserved blocks encode proteins which are poorly conserved between the two organisms and show almost no significant homology at the primary-structure level. Despite the lack of sequence identity, it appears that these represent genes of similar functions. The corresponding pairs of genes are very similar in size and arrangement within the locus, and the proteins which they encode have similar properties (Table 1). The two RfaL proteins are intrinsic membrane proteins with remarkably similar hydrophobic domain structures, even though restriction fragments carrying *E. coli* K-12 *rfaL* have almost no ability to complement an *S. typhimurium* *rfaL* mutant (11). On the basis of computer sequence analysis, the two RfaK proteins do not appear to be as structurally similar to each other as do the RfaL proteins (data not shown), but a restriction fragment carrying *E. coli* K-12 *rfaK* exhibits partial complementation of an *S. typhimurium* *rfaK* mutant (11), indicating that they are functionally somewhat more similar.

The fact that the *rfaK* genes show partial cross-species complementation while the *rfaL* genes do not may be due to

[illegible]

FIG. 5. Comparison of the amino acid sequences of the C-terminal portion of proteins RfaC and RfaL from *E. coli* K-12 (EC) and *S. typhimurium* (ST). Designations are as in Fig. 4. Underlining indicates possible antigen carrier lipid-binding sequences, as described in the text.

cosamine to the completed core. The *rfa* core genes up to and including *rfaZ* have been shown either to be interchangeable between *E. coli* K-12 and *S. typhimurium* (17, 18) or to encode proteins with similar sequences (this study). Therefore, LPS cores of *E. coli* K-12 and *S. typhimurium* should be able to function interchangeably as acceptors for *N*-acetylglucosamine added by transferases from either species. However, there is evidence that the sites of attachment of *N*-acetylglucosamine are not the same in *E. coli* K-12 and *S. typhimurium* (8, 9), and this is consistent with the difference in primary structure between the two RfaK proteins. This suggests that the incomplete complementation of *S. typhimurium* mutants by *E. coli* K-12 *rfaK* is due not to inefficient addition of *N*-acetylglucosamine but to the inability of an LPS core with *N*-acetylglucosamine attached at the wrong site to function as an efficient acceptor of O antigen.

In contrast to the RfaK proteins, the RfaL proteins show no similarity to proteins which are thought to be sugar transferases. If the proposed function of RfaL in the transfer of O-antigenic polysaccharide to the core is correct, it must function as part of a complex which interacts in a highly specific way with the acceptor LPS, with the O polysaccharide, and with one or more Rfb proteins. This would require RfaL to participate in very specific protein-protein and protein-carbohydrate interactions. This may be why RfaL proteins with different primary structures show no cross-complementation.

The sequence of the *rfaFC* region of the *E. coli* K-12 locus is now sufficiently complete to indicate that the products of these genes are similar in structure to their *S. typhimurium* counterparts (20, 24, 25, 28). Since all of the remaining *rfa* genes and the genes immediately flanking the *rfa* locus have been sequenced (this work; 17, 18, 25), it is possible to draw accurate conclusions about the overall structure of the *E. coli* K-12 *rfa* locus. If the 26-kDa gene adjacent to the *rfaD* gene at the left end and the *kdtA* and 18-kDa genes at the right end (23) are included, there are 17 genes in the cluster. The *rfa* cluster shows some remarkable similarities to the *rfb* cluster. In *Salmonella* spp., the *rfb* cluster also has about 17 genes (10), although it may vary considerably, depending upon the sugar composition of the O unit.

In comparing the sequence of the *rfb* gene clusters of group B, C2, and D serovars of *Salmonella* spp., it was found that there was a region of about two or three genes in the middle of the cluster which was not conserved, flanked on both sides by strongly conserved regions (3, 12). Interestingly, an *rfb* gene product, designated Orf-12.8, which is found within this nonconserved block is similar in size and structure to the RfaL protein (10, 12). Like protein RfaL, the Orf-12.8 protein is the only protein in its gene cluster which has a series of potential transmembrane domains throughout its length, which identifies it as a typical integral membrane protein. The RfbP protein has significant hydrophobic domains, but these domains are confined to one end of the protein. The RfbP protein is more similar to the KdtA protein encoded by the *rfa* cluster, which also has a hydrophobic domain at one end and which is the only other *rfa* protein with a significant transmembrane region (4, 10). RfbP and KdtA have other similarities; both are thought to encode sugar transferases which initiate the synthesis of complex carbohydrates, and both are the products of genes which lie at or near one end of their respective gene clusters (10, 23). This latter fact is significant if the ends of the clusters are more strongly conserved than the central regions.

It has been suggested that *rfb* product Orf-12.8 has a function which is specific for different O antigens (3), and this is consistent with its location in the variable central region of the *rfb* cluster. We envision a somewhat similar strain- or species-specific function for protein RfaL. It has been observed that O antigens are not always expressed efficiently when *rfb* gene clusters are transferred between different strains or species of enteric bacteria, and in some cases it has been possible to relate this to an incompatibility at the *rfa* locus. An example involves construction of a hybrid typhoid-cholera vaccine strain (27), in which it was found necessary to introduce the *rfa* chromosomal region from *E. coli* K-12 into *S. typhi* vaccine strain Ty21a to obtain efficient expression of cell surface *Vibrio cholerae* O antigen. It was suggested that this was due to a difference in the structure of the LPS core polysaccharide.

Because of the lack of conservation of the RfaL protein sequence which we have observed between *E. coli* K-12 and *S. typhimurium* and the analogous situation seen with *rfb*-encoded Orf-12.8 among *Salmonella* serovars, we propose alternatively that the failure to express surface O antigens efficiently from genes transferred across strain or species lines may reflect specific interactions between protein RfaL and one or more proteins encoded by the *rfb* cluster. The two alternatives of specific features of core structure and specific protein-protein interactions are not mutually exclusive, and it is possible that constraints on both determine the efficiency of expression of O antigens from *rfa*-*rfb* hybrids. As more *rfa* sequences from different organisms become

available, it will be interesting to see whether conservation of the RfaL protein sequence accompanies the ability to express heterologous O antigens.

ACKNOWLEDGMENTS

We thank A. Wright, S. Raina, and K. Sanderson for making sequence data available prior to publication.

This research was supported by NSF grant DMB89-9626 and Public Health Service grant GM-39087.

REFERENCES

- Albright, C. F., P. Orlean, and P. W. Robbins. 1989. A 13-amino acid peptide in three yeast glycosyltransferases may be involved in dolichol recognition. *Proc. Natl. Acad. Sci. USA* **86**:7366-7369.
- Bachmann, B. J. 1990. Linkage map of *Escherichia coli* K-12, edition 8. *Microbiol. Rev.* **54**:130-197.
- Brown, P. K., L. K. Romana, and P. R. Reeves. Molecular analysis of the *rfb* gene cluster of *Salmonella* serovar Muenchen (strain M67): genetic basis of the polymorphism between groups C2 and B. *Mol. Microbiol.*, in press.
- Clementz, T., and C. R. H. Raetz. 1991. A gene coding for 3-deoxy-D-manno-octulosonic-acid transferase in *Escherichia coli*. *J. Biol. Chem.* **266**:9687-9696.
- Creager, E. S., T. Schulte, and L. I. Rothfield. 1984. Regulation of membrane glycosyl transferases by the *sfrB* and *rfaH* genes of *Escherichia coli* and *Salmonella typhimurium*. *J. Biol. Chem.* **259**:3064-3069.
- Ditta, G., T. Schmidhauser, E. Yakobson, P. Lu, X.-W. Liang, D. R. Findlay, D. Guiney, and D. R. Helinski. 1985. Plasmids related to the broad host range vector pRK290, useful for gene cloning and for monitoring gene expression. *Plasmid* **13**:149-153.
- Fellay, R., J. Frey, and H. M. Kricheldorf. 1987. Interposon mutagenesis of soil and water bacteria: a family of DNA fragments designed for *in vitro* insertional mutagenesis of Gram-negative bacteria. *Gene* **52**:147-154.
- Holst, O., and H. Brade. Chemical structure of the core region of lipopolysaccharides. In D. C. Morrison and J. L. Ryan (ed.), *Bacterial endotoxin lipopolysaccharides*, vol. 1. Molecular biochemistry and cellular biology of lipopolysaccharides, in press. CRC Press, Inc., Cleveland.
- Jansson, P.-E., A. A. Lindberg, B. Lindberg, and R. Wollin. 1981. Structural studies on the hexose region of the core in lipopolysaccharides from Enterobacteriaceae. *Eur. J. Biochem.* **115**:571-577.
- Jiang, X.-M., B. Neal, F. Santiago, S. J. Lee, L. K. Romana, and P. R. Reeves. 1991. Structure and sequence of the *rfb* (O antigen) gene cluster of *Salmonella* serovar typhimurium (strain LT2). *Mol. Microbiol.* **5**:695-713.
- Klena, J. D., and C. A. Schnaitman. Unpublished data.
- Liu, D., N. K. Verma, L. K. Romana, and P. Reeves. 1991. Relationships among the *rfb* regions of *Salmonella* serovars A, B, and D. *J. Bacteriol.* **173**:4814-4819.
- MacLachlan, P. R., S. K. Kadam, and K. E. Sanderson. 1991. Cloning, characterization, and DNA sequence of the *rfaLK* region for lipopolysaccharide synthesis in *Salmonella typhimurium* LT-2. *J. Bacteriol.* **173**:7151-7163.
- Makela, P. H., and B. A. D. Stocker. 1984. Genetics of lipopolysaccharide, p. 59-137. In E. T. Rietschel (ed.), *Handbook of endotoxin*, vol. 1. Chemistry of endotoxin. Elsevier, Amsterdam.
- Manoil, C. 1990. Analysis of protein localization by use of gene fusions with complementary properties. *J. Bacteriol.* **172**:1035-1042.
- Parker, C. T., A. W. Kloser, C. A. Schnaitman, M. A. Stein, S. Gottesman, and B. W. Gibson. 1992. Role of the *rfaG* and *rfaP* genes in determination of the lipopolysaccharide core structure and cell surface properties of *Escherichia coli* K-12. *J. Bacteriol.* **174**:2525-2538.
- Parker, C. T., E. Pradel, and C. A. Schnaitman. 1992. Identification and sequence of the lipopolysaccharide core biosynthetic

- genes *rfaQ*, *rfaP*, and *rfaG* of *Escherichia coli* K-12. J. Bacteriol. 174:930-934.
18. Pradel, E., C. T. Parker, and C. A. Schnaitman. 1992. Structures of the *rfaB*, *rfaI*, *rfaJ*, and *rfaS* genes of *Escherichia coli* K-12 and their roles in the assembly of the lipopolysaccharide core. J. Bacteriol. 174:4736-4745.
19. Pradel, E., and C. A. Schnaitman. 1991. Effect of *rfaH* (*sfrB*) and temperature on expression of *rfa* genes of *Escherichia coli* K-12. J. Bacteriol. 173:6428-6431.
20. Raina, S., Personal communication.
21. Rehemtulla, A., S. K. Kadam, and K. E. Sanderson. 1986. Cloning and analysis of the *sfrB* (sex factor repression) gene of *Escherichia coli* K-12. J. Bacteriol. 166:651-657.
22. Rick, P. D. 1987. Lipopolysaccharide biosynthesis, p. 648-662. In F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger. (ed.), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, vol. 1. American Society for Microbiology, Washington, D.C.
23. Schnaitman, C. A., C. T. Parker, E. Pradel, J. D. Klena, N. Pearson, K. E. Sanderson, and P. R. MacLachlan. 1991. Physical map of the *rfa* locus of *Escherichia coli* K-12 and *Salmonella typhimurium*. J. Bacteriol. 173:7410-7411.
24. Sirisena, D. M. 1990. Molecular studies of the inner core region of the lipopolysaccharide of *Salmonella typhimurium*. Ph.D. dissertation, University of Calgary, Calgary, Alberta, Canada.
25. Sirisena, D. M., K. A. Brozek, P. R. MacLachlan, K. E. Sanderson, and C. R. H. Raetz. Cloning, expression and nucleotide sequence of the *rfaC* gene for ADP-heptose:lipopolysaccharide heptosyltransferase-I for lipopolysaccharide synthesis of *Salmonella typhimurium*. J. Biol. Chem., in press.
26. Steenbergen, S. M., T. J. Wrona, and E. R. Vimr. 1992. Functional analysis of the sialyltransferase complexes in *Escherichia coli* K1 and K92. J. Bacteriol. 174:1099-1108.
27. Tacket, C. O., B. Forrest, R. Morona, S. R. Atttridge, J. LaBrooy, B. D. Tall, M. Reymann, D. Rowley, and M. M. Levine. 1990. Safety, innunogenicity, and efficacy against cholera challenge in humans of a typhoid-cholera hybrid vaccine derived from *Salmonella typhi* Ty21a. Infect. Immun. 58:1620-1627.
28. Wright, A. Personal communication.