Structure of the *rhsA* Locus from *Escherichia coli* K-12 and Comparison of *rhsA* with Other Members of the *rhs* Multigene Family

GREGORY FEULNER, JANE ANTAL GRAY, JANE A. KIRSCHMAN,[†] ANDREAS F. LEHNER,[‡]
ALESIA B. SADOSKY, DONALD A. VLADNY, JI ZHANG,[§] SHENG ZHAO, AND CHARLES W. HILL*

Department of Biological Chemistry, The Milton S. Hershey Medical Center, The Pennsylvania State University,
Hershey, Pennsylvania 17033

Received 16 August 1989/Accepted 15 October 1989

The complete nucleotide sequence of the *rhsA* locus and selected portions of other members of the *rhs* multigene family of *Escherichia coli* K-12 have been determined. A definition of the limits of the *rhsA* and *rhsC* loci was established by comparing sequences from *E. coli* K-12 with sequences from an independent *E. coli* isolate whose DNA contains no homology to the *rhs* core. This comparison showed that *rhsA* comprises 8,249 base pairs (bp) in strain K-12 and that the *Rhs*⁰ strain, instead, contains an unrelated 32-bp sequence. Similarly, the K-12 *rhsC* locus is 9.6 kilobases in length and a 10-bp sequence resides at its location in the *Rhs*⁰ strain. The *rhsA* core, the highly conserved portion shared by all *rhs* loci, comprises a single open reading frame (ORF) 3,714 bp in length. The nucleotide sequence of the core ORF predicts an extremely hydrophilic 141-kilodalton peptide containing 28 repeats of a motif whose consensus is GxxxRYxYDxxGRL(I or T). One of the most novel aspects of the *rhs* family is the extension of the core ORF into the divergent adjacent region. Core extensions of *rhsA*, *rhsB*, *rhsC*, and *rhsD* add 139, 173, 159, and 177 codons to the carboxy termini of the respective core ORFs. For *rhsA*, the extended core protein would have a molecular mass of 156 kilodaltons. Core extensions of *rhsB* and *rhsD* are related, exhibiting 50.3% conservation of the predicted amino acid sequence. However, comparison of the core extensions of *rhsA* and *rhsC* at both the nucleotide and the predicted amino acid level reveals that each is highly divergent from the other three *rhs* loci. The highly divergent portion of the core extension is joined to the highly conserved core by a nine-codon segment of intermediate conservation. The *rhsA* and *rhsC* loci both contain partial repetitions of the core downstream from their primary cores. The question of whether the *rhs* loci should be considered accessory genetic elements is discussed but not resolved.

Families of homologous but nonidentical genes present a number of special genetic questions. These questions include the possible specific roles of individual members, the degree to which members exchange heterologies through intrachromosomal recombination, and the effects of these exchanges on function. The presence of these multigene families also has important implications for chromosome rearrangement and evolution (17). However, except for the *rrn* operons encoding ribosomal RNA, multigene families are quite rare in *Escherichia coli* (19). We have recently reported an unusual and complex family, the *rhs* family, that is comparable to the *rrn* family in number, length of shared homology, and degree of sequence similarity. The *rhs* loci were originally detected through their action as rearrangement hot spots, providing homology for *recA*-dependent intrachromosomal recombination (4, 11). Consequently, the *rhs* loci were defined to include the homologous sequences shared by two or more of the respective loci. Four *rhs* loci of *E. coli* strain K-12 have been characterized extensively, and evidence for a fifth has been noted (4, 11, 20).

A distinctive feature of the *rhs* loci is that they share a

highly conserved 3.7-kilobase (kb) core sequence. The cores are generally flanked by dissimilar sequences, and for two of the loci, *rhsA* and *rhsC*, one or more partial core repetitions are present downstream from the intact core. Sequence comparison of the first 300 nucleotides of the *rhs* cores (20) revealed that the core homology begins precisely with a start codon initiating an open reading frame (ORF) and that the *rhsA*, *rhsB*, and *rhsC* cores are closely related, showing only 1 to 2% sequence divergence. By contrast, *rhsD* is 18% divergent from the others. A total of nine mismatches distinguish *rhsA*, *rhsB*, and *rhsC* through these 300 nucleotides, but none of the nine causes an amino acid substitution in the core ORFs. However, *rhsD* differs from the others by eight amino acids. This degree of divergence through predominantly neutral mutation indicates that the cores have been evolving independently for quite some time. Application of the mutation rate estimated for enteric bacteria by Ochman and Wilson (15) suggests that the *rhsA* and *rhsC* cores diverged on the order of 10 million years ago, and the extent of sequence divergence of *rhsD* would indicate that it radiated from the others on the order of 100 million years ago. The nature of the *rhs* core is a primary consideration of this report, and we have approached this issue by sequencing *rhsA* and comparing it with selected portions of other *rhs* loci.

Originally, the *rhs* loci were defined operationally to include the homologies shared by two or more of the respective loci. However, this definition failed to describe them adequately as either functional or evolutionary units.

* Corresponding author.

[†] Present address: Advanced Research Division, Agrigenetics Corp., Madison, WI 53716.

[‡] Present address: Department of Cell and Molecular Biology, Medical College of Georgia, Augusta, GA 30912.

[§] Present address: Department of Pharmacology and Molecular Biology, The Chicago Medical School, North Chicago, IL 60085.

This inadequacy of definition became especially clear when we found that the first base of the core homology initiates a large ORF. A definition of locus based strictly on homology would therefore exclude sequences potentially responsible for the expression of the *rhs* core ORF. A second reason to find a more inclusive definition of *rhs* came as we realized that the shared homologies at individual *rhs* locations are interspersed with large nonhomologies. As will be described, the approach we took to redefine the limits of *rhsA* and *rhsC* was to compare sequences from *E. coli* K-12 with an *E. coli* strain that has no homology to the K-12 *rhs* core. The source of this *Rhs*⁰ strain was the ECOR collection. The ECOR collection is a collection of 72 *E. coli* strains which was assembled with the intent of including strains embracing the maximum genetic diversity of the species (14). The *Rhs*⁰ strain we used was one of a number of *Rhs*⁰ strains found when the ECOR collection was surveyed for the presence of *rhs* core homology.

MATERIALS AND METHODS

Bacterial strains. The *E. coli* K-12 strains used for the initial cloning of the *rhs* loci have been described previously (11, 20). Strains from the ECOR collection (14) were supplied by Robert Selander. Plasmids were maintained in either of two *recA* host strains, CH1480 (2) or DH5 α (Bethesda Research Laboratories, Inc.).

Plasmids and cloning techniques. Plasmids containing the *E. coli* K-12 *rhs* loci have been described previously (11, 20). Subclones for nucleotide sequencing were prepared by using pUC18 or pUC19 (13) as the vector. Cloning of chromosomal segments from ECOR strains involved the digestion of genomic DNA with the specified restriction endonucleases, ligation of these digests with analogous digests of pUC19, and transformation of the *E. coli* K-12 derivative DH5 α , selecting Amp^r. Groups of several hundred transformants were pooled, and their plasmid DNA was extracted and tested for homology to appropriate probes. A pool showing positive results was used in secondary transformations, and secondary pools were made of about 100 transformants. The process was repeated, reducing the pool sizes until a single transformant that gave a positive hybridization was obtained. In some cases it was expedient to gel purify pooled plasmid DNA in order to select appropriate sizes. Plasmid fusion constructs for assessing the expression of *rhs* ORFs were initially created by using the plasmid pMC1403 (5). For each fusion, an appropriate fragment from the cloned *rhs* locus was ligated in frame with the β -galactosidase gene of pMC1403. In order to eliminate the complicating transcriptional activity provided in the reverse sense by the promoter region of the Amp^r gene, the Amp^r gene of the fused constructs was replaced (in the same anti- β -galactosidase orientation) by the chloramphenicol resistance gene of pBR325. β -galactosidase activity was observed in the Lac⁻ strains M182 (6) and an F⁻ derivative of JM109 (13). Enzyme specific activity was estimated by the method of Silhavy et al. (26).

Maxicell experiments for detecting products of peptide fusion. In general, the protocol of Sancar et al. (21) was followed. Various plasmids derived from pBR325 were transformed into the *recA1 uvrA6 phr-1* mutant CSR603 (21). After irradiation of the cells for 40 s at 3 ergs mm⁻² sec⁻¹ to induce chromosomal degradation, the bacteria were grown out and then treated with cycloserine (200 μ g/ml) for 8 to 12 h. After cycloserine removal, the bacteria were labeled with [³⁵S]methionine for 15 min to selectively label plasmid

proteins. Extracted proteins were fractionated by electrophoresis on 10% polyacrylamide gels and visualized both by staining with Coomassie brilliant blue and by autoradiography. Molecular weight standards included the endogenous β and β' subunits of RNA polymerase and the drug resistance proteins encoded by pBR325.

DNA isolation, restriction enzyme analysis, and nucleotide sequencing. Procedures for the isolation of genomic and plasmid DNA, the production of overlapping deletions, DNA analysis, and DNA sequencing were as specified previously (11, 20), except that in addition to chemical sequencing, the dideoxy method (22) was used in some cases. For dideoxy sequencing, a kit purchased from United States Biochemical Corp. was used. The sequence of both strands was determined in all cases. Sequence analysis was aided by use of the MacMolly program (Soft Gene Berlin) and the BIONET Resource. The Sharp and Li codon adaptation indices (23) were calculated by using the DNA Strider program (Christian Marck, Gif-Sur-Yvette Cedex, France).

RESULTS

***rhs* elements are not present in all *E. coli* strain.** In order to determine whether *rhs* genes are universally present throughout the *E. coli* species, strains from the ECOR collection were tested for homology to the *rhs* core. Genomic DNA from each strain was digested with restriction endonucleases and subjected to Southern analysis. The papers were probed by using the 3.0-kb *EcoRI-BamHI* fragment from the *rhs* core (Fig. 1). While the majority of the ECOR strains produced multiple homologous bands, a significant number of strains did not. Details of the complete survey will be published elsewhere (M. Pencek, G. Feulner, and C. Hill, manuscript in preparation). The lack of *rhs* sequences in some *E. coli* strains indicated that these genes are not essential for general viability. Their absence also provided an opportunity to define the limits of the loci by comparing the regions containing the *rhs* loci in *E. coli* K-12 with the corresponding regions from one of the strains devoid of *rhs* core homology. Strain ECOR 55 was selected for this purpose.

A series of probes specific to portions of the *E. coli* K-12 genome, flanking both the *rhsA* and the *rhsC* cores, was used to analyze digests of ECOR 55 DNA. The probes and whether they gave a positive signal with ECOR 55 DNA are represented in Fig. 1. The probe designated A-us hybridized strongly to ECOR 55 DNA. Since A-us extended from the *SalI* site to a position about 0.2 kb in front of the *rhsA* core, the left end of *rhsA* could be placed to the right of this *SalI* site. The A-ds probe, a 382-base-pair (bp) *PvuII* fragment, also hybridized strongly with ECOR 55 DNA, while an *AccI-PvuII* fragment immediately to its left was negative. This placed the right end of *rhsA* near the border of these two probes. Furthermore, both the A-us and the A-ds probes hybridized to the same-sized fragment, indicating that a single 2.7-kb *SalI-HindIII* fragment from ECOR 55 contained the site at which *rhsA* resides in strain K-12. A similar analysis showed that the C-us and C-ds probes both hybridized to a 3.8-kb *EcoRI-SphI* fragment from ECOR 55 genomic DNA. This fragment presumably contained the site of *rhsC* inclusion. In view of these findings, *rhsA* and *rhsC* were defined as all genetic material contiguous with their respective cores which is present in *E. coli* K-12 but absent from ECOR 55. In order to further characterize *rhsA* and *rhsC*, the 2.7-kb *SalI-HindIII* fragment that hybridized to A-us and A-ds and the 3.8-kb *EcoRI-SphI* fragment that

FIG. 1. Restriction maps of the *E. coli* K-12 *rhsA* and *rhsC* loci. The filled bars indicate the *rhs* core sequence and the partial core repetitions, while the open bar indicates the extension of the core ORF at its carboxy terminus. ORFs predicted from the *rhsA* sequence are shown below the *rhsA* map. Probes specific for various segments of both *rhsA* and *rhsC* or for the flanking regions are shown by the bars above the restriction maps. Probes which gave a strong positive signal when used with ECOR 55 DNA are indicated with (+), while those that were negative are indicated with (-). Restriction site designations: Ac, *AccI*; B, *Bam*HI; Bs, *Bst*EII; C, *Clal*; Hc, *Hinc*II; K, *Kpn*I; M, *Mlu*I; P, *Pst*I; R, *Eco*RI; S, *Sal*I; V, *Pvu*II. Only selected *AccI*, *Bst*EII, *Clal*, *Mlu*I, and *Hinc*II sites are shown. The coordinates are in kilobases.

hybridized to C-us and C-ds were cloned from ECOR 55. Figure 2 shows restriction maps of these two segments, aligning them with a representation of the corresponding *E. coli* K-12 regions. The site of *rhsA* inclusion could be narrowed to a segment bounded by conserved *Bst*EII and

Clal sites, while the location for *rhsC* was in a segment bounded by an *Eco*RI site unique to ECOR 55 and a conserved *Pvu*II site. These fragments were subcloned and used for nucleotide sequencing.

Demarcation of the *rhs* boundaries. Sequence comparison of the vacant *rhs* sites in ECOR 55 with the corresponding *E. coli* K-12 sites permitted a more precise definition of the loci. This comparison is shown for *rhsA* in Fig. 3a and for *rhsC* in Fig. 3b. For *rhsA*, the homology between *E. coli* K-12 and ECOR 55 ended 191 bp upstream of the *rhsA* core and resumed 8.2 kb beyond the initial divergence. Interestingly, the points of sequence divergence and reconvergence were not contiguous in ECOR 55 but were separated by a 32-bp segment. For *rhsC*, the K-12 and ECOR 55 homologies ended 190 bp upstream of the *rhsC* core and resumed 9.6 kb downstream. A short 10-bp pyrimidine-rich sequence separating the upstream and downstream homologies in ECOR 55 was found. The limits of the two loci as defined by this sequence comparison are indicated in Fig. 1.

Sequence of *rhsA*. The preceding results established that the *rhs* loci are considerably larger than their 3.7-kb conserved cores. In order to better understand their structure, the nucleotide sequence of a 9,234-bp *Sall*-*Pvu*II segment that includes *rhsA* was determined (Fig. 4). Features of this sequence are designated in Fig. 1 and 4. The entire *rhsA* unit contained 8,249 bp.

The *rhsA* core. The boundaries of the *rhs* core homologies were defined through sequence comparisons of *rhsA*, *rhsB*, *rhsC*, and *rhsD* (20). Examination of the complete *rhsA* core sequence revealed that it extends 3,714 bp. It is situated 191 bp from the left end of the locus as defined above. Two partial repetitions of the distal portion of the core are located downstream from the intact core. The first repetition is 430 bp long, while the second is 160 bp long. Previous work showed that the core homology begins precisely at a start codon that initiates an ORF extending at least the first 300 bp (20). Inspection of the complete sequence shows that this ORF actually extends the full length of the core. Thus, the core is capable of encoding a peptide of 1,238 amino acids. The core ORF does not terminate at the end of the core, however, but extends another 417 bp or 139 codons beyond the core into a region unique to *rhsA*. A translation of this ORF is shown in Fig. 5. The molecular mass of a protein produced from the complete ORF would be 156 kilodaltons (kDa). A second ORF, containing 280 codons, is initiated 21

FIG. 2. Alignment of the restriction maps of the *rhsA* and *rhsC* regions of ECOR 55 with those of *E. coli* K-12. The position of the *rhs* insertion into K-12 is shown above the aligned maps. Restriction site designations: Ac, *AccI*; Bs, *Bst*EII; C, *Clal*; D, *Hind*III; M, *Mlu*I; R, *Eco*RI; S, *Sal*I; Sp, *Sph*I; V, *Pvu*II.

FIG. 3. Sequence comparison of *E. coli* K-12 and ECOR 55 *rhs* sites. (a) Comparison at *rhsA*; (b) comparison at *rhsC*. Matches (:) and mismatches (*) between the K-12 and the ECOR 55 sequences are indicated. The *PvuII* site present in K-12 is eliminated in ECOR 55 by a base substitution. Shown below the right flank of the *rhsA* locus is the position of a stop codon that terminates a potential open reading frame emanating from beyond a *PvuII* site located 374 bp beyond the depicted sequence in strain K-12 (see Fig. 1).

nucleotides beyond the termination of the extended core ORF. It is designated ORF-2 in Fig. 1 and 4.

Cores of *rhsB* and *rhsC*. Previous work demonstrated that *rhsA*, *rhsB*, and *rhsC* possess highly conserved cores. This conclusion was based on restriction mapping, S1 endonuclease-resistant heteroduplex formation (11), and sequence comparison of portions of the cores (20). However, the cores are not identical. As mentioned in the introduction, the degree of divergence suggests that they have been evolving independently for a considerable period. Therefore, it was of interest to determine whether the *rhsB* and *rhsC* cores have also been maintained as a single ORF, since in the absence of selection, nonsense mutations interrupting the core ORF would be expected to accumulate. Evidence that the ORF was maintained, at least for a major portion of each of the three cores, was obtained by gene fusion. The 3.0-kb *EcoRI*-*Bam*HI segment from each locus was ligated between the *EcoRI* and *Bam*HI sites of pBR325 (1). The *EcoRI* site of pBR325 is in the amino-terminal portion of the *Cam^r* gene (18), allowing the formation of an in-frame *Cam^r*-*rhs* gene fusion. The predicted 133-kDa hybrid protein would contain the first 71 codons of *Cam^r*, 994 codons of the respective *rhs* cores, and 88 out-of-frame codons from the pBR325 *Tet^r* gene. In maxicell experiments (21), the presence of each of the three constructs produced a large peptide doublet estimated to be 130 and 135 kDa in molecular mass (Fig. 6, lanes d, e, and f) which was not produced by the pBR325 control (lane a). Although we do not understand the reason for the doublet, we conclude that the core ORF of all three loci is maintained over the 3.0-kb *EcoRI*-*Bam*HI segment. Plasmids producing these hybrid proteins caused the cells to be very sick. Faster-growing derivatives could be obtained by overgrowth of liquid cultures or by inspection of cultures on agar plates for larger colonies. The plasmids of several of these fast-growing derivatives were shown by restriction analysis to have acquired an *IS1* element inserted within the *Cam^r* sequences. Maxicell experiments demonstrated that the large peptide doublet was no longer produced by any one of the three *IS1* insertion variants tested (Fig. 6, lanes c, g, and h). The elimination of the large peptide doublet by *IS1* insertion provided further evidence that the peptides were derived from the *rhs* gene fusions. The appearance of the peptide doublet was also affected by a 2.2-kb deletion from the *Cam^r*-*rhsA* fusion (Fig. 6, lane b).

Divergent extensions of the core ORFs. The *rhs* core was originally defined as the region of high sequence similarity shared by the *rhs* loci (11, 20). The finding that the *rhsA* core contained a single ORF which continued 139 codons beyond the end of the core revealed an intriguing aspect of *rhs* structure. Each *rhs* appears to have the potential to produce a very large peptide that consists of a highly conserved 141-kDa amino fragment followed by a highly divergent carboxy terminus. In Fig. 7, we compare the sequences surrounding the core ends, including the partial repetitions found in *rhsA* and *rhsC*. This comparison shows that the high conservation of nucleotide sequence ends with a proline codon (nucleotides 49 to 51 in Fig. 7), and we have chosen to define this position as the core end. An intermediate degree of conservation is observed for the 27 nucleotides following the core. The amino acids encoded by the first 9 of these 27 nucleotides vary considerably, but the amino acids encoded by the remaining 18 nucleotides are more highly conserved. This region is designated the intermediate region in Fig. 7. With one exception, the nucleotide sequences beyond the intermediate region diverge completely for all pairwise comparisons. Despite this sequence divergence, the ORFs continue for a considerable distance before being interrupted by a stop codon (Fig. 8). The number of amino acids potentially encoded by each of these extensions are: *rhsA*, 139; *rhsB*, 173; *rhsC*, 159; and *rhsD*, 177. Interestingly, *rhsB* and *rhsD* represent the exceptional pair of core extensions, sharing 50.3% amino acid similarity (underlined in Fig. 8).

Expression of the ORFs. The ATG codons at the start of both the core (nucleotide 759) and ORF-2 (nucleotide 4913) lie within sequences that suggest they would function well as initiation codons. They are preceded by AGGA and GGAGG sequences, respectively, which are highly appropriate ribosome-binding sites (24). In addition, both ATG codons are followed by sequences characteristic of functioning initiation codons in *E. coli* (16); these are AAC for the core and TCAA for ORF-2. For both the core ORF and ORF-2, in-frame fusions with the β -galactosidase gene were constructed such that the inserted *rhs* segment must provide both an active promoter and an initiation codon for β -galactosidase synthesis (see Materials and Methods). One construct, representing nucleotides 2 to 847 (Fig. 4), fused the first 29 codons of the core ORF to β -galactosidase. β -Galactosidase expression was very low from this con-

FIG. 4. Nucleotide sequence of the *rhsA* locus and flanking regions (GenBank accession no. M29716). The sequence extends from the *SalI* site to the right-hand *PvuII* site shown in Fig. 1. Features are shown above the sequence. Key start and stop codons as well as the proline codons ending the core and core repeats are underscored. Candidate promoter sequences (−35 and −10; [10]) and ribosome-binding sites (rb; [16, 24]) flanking the ATG initiation codons are overscored. A dyad symmetry preceding the core is depicted by wavy underscoring.

struct, even when it was carried on a multicopy plasmid. In fact, we estimate that no more than one enzyme monomer per cell was produced. The ORF-2 construct fused nucleotides 4311 to 5015 to β -galactosidase and thereby included the first 34 codons of ORF-2. In contrast to the core-ORF fusion, high levels of β -galactosidase were obtained, as judged by the dark blue color of colonies grown in the presence of X-gal. A parallel ORF-2 construct which included only nucleotides 4899 to 5015 gave white colonies, suggesting that the promoter used was encoded, at least in part, by nucleotides 4311 to 4898. Candidate −35 and −10 sequences for this promoter are indicated in Fig. 4. The expression of these ORFs and their control are currently being investigated.

DISCUSSION

The function(s) of the *rhs* multigene family remains to be determined. However, consideration of their distribution among *E. coli* isolates, along with aspects of their structure, provides some insight. They are clearly nonessential for cell growth, since they are not present in a number of *E. coli* strains. In addition, the cores of the *rhsA*, *rhsC*, and *rhsD* of strain K-12 can be individually disrupted without affecting cell growth. For example, the region between the *KpnI* sites of the chromosomal copy of *rhsC* (Fig. 1) can be replaced with a drug resistance determinant without affecting the routine culture of the cells (20). Similar disruptions of the *rhsD* (20) and *rhsA* cores (unpublished observations) also had no observable effects. Despite their nonessential nature,

```

TGAACAGCCTGCTGTCTGACCGGATTACACCTGGAACGACAAACGGCGAACTCATCCGATCAGCAGCCCGCCAGACCCGGAGTTACAGTACAGCAC
3400
CACC GGCGAGGTGACCGGGTTACACACCGCAGCGAATCTGGATATCCGATCCCGTATGCCACAGACCCGGCAGGTAAACCGCTGCCGACCCGGAG
3500
CTGCACCCGGACAGCACCTCAGCATGTGGCCGATAACCGTATCGCCGTGACGCGCACTATCTTTACCGGTATGACCGTCACGGCAGGTGACAGAGA
3600
AAACCGACCTCATCCCGAAGGGGTTATCCGCACGGATGATAGCGGACTCACCGGTACCATTACGACAGTCAGCACCGGCTGGTGCACTACACGGGAC
3700
ACAATATGAAGAGCCGCTGGTTCGAAAGTCGTATCTTTACGACCCGCTGGGCGCAGGGTGGCAAAACGGGTGTGGCGCGTGAAACGGGACCTGACGGGC
3800
TGGATGTCGTGTACGGAACCGCAAGTGACCTGGTACGGCTGGGACGGCGACCGGTGACCAACAATCAGAACGACAGAACCCGCATCCAGCAGATT
3900
ATCAGCCGGGAGCTTCACGCCACTCATCAGAGTTGAAACCGCCACCGGTGAGCTGGCGAAACGACGCGCGCAGCCTGGCGGATGCGGTTACGACAGTC
4000
CGCGGGCGAAGACGGTGGCAGTGGTGTTCGCCCGGTGTGGTGAGATGCTCGACCGGTGGAAGTGAATCCTGGCTGACCGGGTGAGTGAGGAA
4100
AGCCCGCGTGGCTGGCATGTCGGCCTGACCGTGAGCAGATGCAAAACAGATGGACCCGGTGTACACGCGCGCGAAAAATCCACCTGTACCACT
4200
GCGACCATCGCGGCTGCGGCTGGCCCTTATCAGCAAGGAAGGCAACAGAATGGTGGCGAGAATACGATGAATGGGGCAACCTGCTGAATGAAGAGAA
4300
CCCGCATCAGCTGACGACGCTTATCCGCTGCCGGGCGCAGTATGATGAGGAGTCCGCGCTGTATTACAACCGCCACCGCTATTATGACCCGCTGACG
4400
GGCGGTATACCTCAGGATCCGATTGGGCTGAAGGGGGTGAATTTTATCAGTATCCGTTGAATCCAGTTACGAATACAGATCCTCGGGGTTAG
4500
AAGTTTTCTAGACCATTCCTTGGCAATTCATGGCCCAAGCCCTGCACAGCAGCAAGCAGATGATAATGCTGCAAAAGCATTGACAAAATGGTG
4600
GAACGATACAGCATCAAAAGAAATTTGACTCTCTAATATTGAATAATCCGGGACTAGCATTAGATATAACAATGATAGCTTCTCGTGAATGTTGCA
4700
GACACAGGGATAACTGATCGTCAATGACATAATAATGACAGATTCTGGAGTGATGGAAAAACCCGACAGATGTGACGTACTTCAGGAATAATTG
4800
ATTGTGGTATATTAGTGTCAAGATGCAAAAGCACAGAAAGCCGGAATTGTCGTCACCTCCAGACAGTCAAACGATAAAAAAGATAGCCCTTGTG
4900
GAGGTTCTCGCAATGCAATACATACCAGAAAAGAAAGGCAAGTAAAGAGTATGGTTTATATAATCAATGAAGAACTAAATGATGATGAATATTTC
5000
GCTTACTGTATGATCACAATTCCTTGAAGAGATTTCATCTGCCAGAGTATTACAGTTAAGAGGTGGGCAAGACGCTGTAGATTGGCAATTGAGTCTG
5100
CTCTGATAAAATATATCCGTAGAGATATCGGAGCATTATCTCGGCAATAAAATTTGCAAAATGCGAAGATAATGTTTTAATATTTTGAAC
5200
AATATGGCATTGAATGATAAGAGCGCTTGGCTTCAGAGTACGGCAATCGAGTCAACGGCTCAGCGATGCAAGAAAAACCAATTTATTCACCTAAATAG
5300
TAGAACAATCTCAAATTACTGCTTTGATAAATCGACTAATGTGACAGTGTACAGCATTTCGTTATTTCTGTTATCAATGATAAAGCAACAATTCCTACT
5400
ATTGATTAACTGTTAAAGATCCAAATGGAGATGTGAGAACTGGGCGCATTGCAATAAATCAATAAATATGATAATAGTGATATTAGGGATTGT
5500
TTTGTGGAGATGCTTCAGGATAAAATGAGGAAGTCCGTATTGAAGCAATAATCGGACTTTCTCAGAAAAAGATAAAAGGGTTTTATCTGTTTATGCGG
5600
ATGAGTTAAAAAAATACTGTTTATGATGATATCATTGAAGCTGCGGTTAGGTGATAAAACGCTACTTCCTGTTTATGATACTATGTTGTACAA
5700
GTTTGATGACAATGAATATAACTTCCGCTATTGATAAGCTGAAGCGTTCAATAGCTGAGTTGCCTATGCACAGTGGGGATTCCCGCCGGACCGGTGC
5800
AGATGCTCGACCGGCTGGAAGTGAATCTGGCTGACCGGTGAGTGAGGAAGCCGCGCTGGCTGGCATCGTGGCGCTGACCGGTGAGCAGATGAA
5900
AAACAGATGGACCGGTGTACACGCGCGCGAAAAATCCACCTGTACACTGCGACCATCGCGGCTGCGGCTGCGGCTCATCAGCAGGAAGGCCA
6000
ACAGCGTGGTGGCAGAATATGATGAATGGGGCAACCTGCTGAGTGATGAGAACCCGATCATCTGCAGCAGCTCATTGCTGCCGGGTGACGAGTATG
6100
ATGAGGAGTCCGCGCTGTATTACAACCGCCACCGCTATTATGACCGCTGCTGGGCGGTATATCACTCAGGATCCGATTGGACTGAAGGGGGATGGAA
6200
TTTTCATCAGTATCCGTGAATCGGTCATAAATGATAGTCCGCAAGGTTTGGTTGATATAAATTTATACCCGAAAGTGATCTTATCATTCTGTAGCT
6300
GATGAGATTAAATCCAGCGGTTTTACAAATCGGGGCGATGTTACCCACATCTATTGAATCCGCAACGCGCAGTATCATGACAGCTAAAGATCTAG
6400
CATATCTAATTAATTTGATGGGAATTATAAAGATGGGATGACAGTTTGGTTATTTCTTGAATACAGGTAAGGACAAAAATTCATTGCTAGCTAATT
6500
AGCTAAAGAGTTACATACAAATGTAATAGGACCTGACACGCTATGGACGTGGTGGGGCGAGGAACATAATGGTAAGTTAAAAATGGATACAGTGCTAACA
6600

```

FIG. 4—Continued

the integrities of the *rhs* loci have been maintained throughout a considerable period of evolution. There are seven third-position mutations distinguishing the first 100 codons of the *rhsA* and *rhsC* cores (20). If there were no selection for the integrity of the core ORF, approximately 43 nonsense mutations would have been expected to accumulate within the 3,714-bp core during the evolutionary time required for the occurrence of these neutral mutations. Since the *rhsA*, *rhsB*, and *rhsC* core ORFs have remained intact, their maintenance implies that loss of function of any one of the three elements would be detrimental to the cell. What this nonessential function might be is unclear, but one obvious possibility is that the *rhs* elements play an important role in the natural ecology of the cell.

Examination of the amino acid sequence predicted for the *rhsA* core reveals several remarkable features (Fig. 5). It is strongly hydrophilic, being especially rich in hydroxylated amino acids; 20% of the residues are threonine, tyrosine, or

serine. The dipeptide Tyr-Asp occurs 19 times throughout the core ORF (Fig. 5), and certain positions near the Tyr-Asp dipeptides display significant nonrandomness. Closer inspection of the sequence reveals the existence of the motif, GxxxRYxYDxxGRL(I or T), which is present 28 times. Portions of the core ORF are arranged in Fig. 9 to emphasize this motif. The motif tends to occur at regular intervals, especially within amino acids 418 through 711, where the period of repetition is generally every 20 or 21 amino acids. The internal repetition, along with the hydrophilic nature and large size of the *rhs* core peptide, has precedence among certain microbial cell surface proteins such as the fibronectin-binding protein of *Staphylococcus aureus* (25) and hemolysin of *E. coli* (8). Surface proteins generally contain sequences that serve either membrane-spanning or signal functions. The putative *rhsA* core peptide does, in fact, contain a hydrophobic stretch of 28 uncharged amino acids (shown by underscoring in Fig. 5) indicative of

```

6700
GCACCAACGACCTTAATTCAAATAAGGATCTAATGGCTATAACAACAAAGACCTTGGTAATGGATAACATATGGGCCATCTGGGCCACCCATTTCTA
|ORF-3→
6800
ATATGCAAGGTACGCCAGAAAAACCCAGTGATATAAGATAGGTGTAGATGTAAGAAAGCATGCTTGTACTATTTTTTATTCTCTTTTATTGTCAA
6900
TTGCATGGTCTGTATGTGAAAAAACAACGAAAGTATGATGGGGTCTACAGCAATGACCTATGATTTAAGTGAAGAAAAGTTGATGAACTGAAGT
7000
ATAAATCACAAATGGTATTCAGAGGCATCATTTCGACTATATCAATATTACTGCTTTACTAAAAATAACATTACAAACAATTGCGATTCTGGAAAG
7100
ATCAGCATCTCAGGGGAATGTTACAGCGCAGTTTAATTATGGGGTCTTTTATCAGATACAAATCCAACTATTACAGAAATATTAAATTTGAATAGAGCT
7200
ATTTATTGGATGGAATTCGCTGTAAATAACGGTAATATTGATGCAAAAGAGCAAACTTCAAGAGCTCAAAAAGCTAAAGCGAATGGATAGAGGAAGAATA
←ORF-3| | core-repeat A'→
7300
AGGAGAATCCGTAAAGCAATAGTTCATCAGACTGCCCGCAGCAGTAGTATGATGAATCAGGGCTGTACTATACTATTATCGGTAACATTAAACCGCA
←core-repeat A' |
7400
GCAGGGGCGGTATAACACTCAGGCTCCGATTGGGGTGGAAAGTAGTGGAAATTTTATCAGTATCCGTTGAACCGGATTTCTGGAATCGATCCCTTTGGG
7500
ATTAGCCACCTGTTTATACAGCATAACTTTAGGTATGCTTTCGTGTGTTTCAGATATACCATGTGATGATAATTTTATGATGTCTTAGTATCCAGTC
7600
GCTTCGTGTAATATGCAATGTAAAAAATATCCCGTTGTACACATTATAAATAGATGGCCAACTCAAGGCCCTATGGAGTTGGAATGTTAATGA
7700
CCCGGAGCAGCAATAGAAAACCTGACGCATACGATTAGTACCGTCAGTTAATACAGGAACCTATAATCGCAATGGATTTTCAATACATAGTTGTTA
7800
AACGCTTTTGGTCCGCTTTAGGACCTCGATTCTGTTCCGAGGGTGTATAACTGGTTTATCTAATGACATGCAAAAATTAATAGATTAAATTTTCAG
7900
AGCCAGATAGCACAATAACAGTTACGGACTAAAAAATATGGCAAGGTTATTAACAATTGTATTTTTTCTATTCTGCGAGTGCAGTTACTGCAAGAC
8000
AATTTACAGAGAGTTAGATTGTCAGACAAAGCAATATTATGGGCACTAATATCAGCCAGTACTAAAGAGGGACGAAAAGCGTGTCTCACTAAGCTATTT
8100
TGCCTGTAAGGCCCGCAGAGGCTGAATCGGACTGCGCATATGGCTGCGAATGATAACAAGGAATTTCTACATCCTTATCTAATATAATGAGGTACAAA
8200
ATAGATGCTGGACTTCCGAATCCTATACATGCTATTTAATAGTAAGGAAAGATCATAAGACCATATCTGAAAAATCTAAATCCACTCCAACCTGCGC
8300
CTGATTGCATTGAACAGTAAATAAAATAAAGATAAGAAATAAAAAAATCATTGATATAAATCAGTTAATTTTGTAGTGATGATAAAAAATATAAAT
8400
TGAGAGTTAATAGTACTATTATGGCCATTGATGATTCATATAAGTGCATCGATGAATGATAACTTAATTTTATTTAAAAATATATTAATCACTCTC
8500
CAGGAGTTTTTTAACTTAGTTTTAATGAATACGATAAATAAAGTGAAGAAAGAAATAGTTGTATACATATATTCATGACCCCTACATTGTGATCTA
8600
TAGATTAAATATTTTTAAATGTACTGTCATCCTTCAGATAATAAAGCCATCATCGCAATAGATGGAAAGCACTCCTGCTCTTTTTGATTAGAGCGG
8700
TCGAGGAGGAACAACCTATATCATCATTGCGATCTCGACGATGTATAGTCTCATCTCGCGCAGAACAGAGCGGATAAAAAATCCAATGAGAGCACACT
8800
CATCCAGAACTTTAATACATAATGAATATTAAGAAAAATCATCGAAGCTAATGAAGTAATATGCAAAAATAATAGTCTTACAAGGAGCAAGTTCCA
← rhaA || ORF-ds←
8900
TAATCTCCCCCTCCCCCTCAATGATCCAAATAAGATAATGCATCCAGCTGGTCATTCTTAGCAGCACTTCCGCATCACCAGAACATGGTGAAATGG
9000
TCGGAGTAAACCCACCTGGCGCTAGATGCCGTGGGTAAAGCATCGATATCATCGTTAGGGTCCAGTTCAATGGTTCCAGCACACCGTCGTCGCCGG
9100
GCACGACCGTTAATGATTGCAATACCCCTGCGCTGATAAGAACCGCTGGCAGCAGGTAATACTAGTCAGTTTGGCGTGAACACCTGCCCGAGG
9200
TAGCGCGTTAAACACCATTCCGCTATCGCCAGGTTTCAGACGTAACAGCGAGTTTTCGCCGAAATTGGCGGACAATTTGCCGTTTTTCTCGGGGATG
AAGACCATCACCAGGACGAGCGGCAAGGCAGCTG

```

FIG. 4—Continued

a potential membrane-spanning domain (27). In addition, the first 26 residues of the amino terminus contain two positively charged amino acids, followed by a run of 15 uncharged ones, providing a potential candidate for a signal peptide (28).

Expression of the core ORF would require the presence of a promoter within the first 191 nucleotides preceding the core. One candidate sequence that compares well with the promoter consensus (10) is located between nucleotides 674 and 703 (Fig. 4). However, β -galactosidase fusions, con-

```

100
MSCKPAARQCDMTQYGGSIQGSAGVRIGAPTGVACSVCPGGVTSCHPVNPLLCANVLPGETDIALPGPLPFIILSRYSYRRTKTPAPVCSLPGWKMFA
200
DIRLQLEDNTLILSDNGGRSLYFEHLFFGEDGYSRSESLWLVRGGVAKLDEGRLAALWQALPEELRLSPHRYLATNSPQGPWLLGWCEVPFEDEVLP
300
APLPFYRVLTGLVDRFRGTQTFHREAAGEFSGEITGVTGAWNRHRLVLTQQAQRAEARQQAISGGTEFSAPFDLTGCTEYGRDNGIRLSAVWLTHDF
400
EYPENLPAAPLVRYGWTFRGELAVVDRSGKQVRSTFYDDKYRGRMVANRHTGRPEIRTYDSGRVTEQLNPAGLSYTYQYKDRITITDSDRRREVLE
500
TQCEAGLKRVRVKEHADCSVTQSQFDVAGLRAQTDAAGRITTEYSPDVVTCLITRITTFDGRASAFYNNHNLQTSATGPDGLELRREYDELGRLIQETA
600
PDGDIITRYRYDNPHSDLPATEDATGSRKMTWSRYGQLLSFTDCSGYVTRYDNDRFQMTAVHREELSLQYRAYDSRGQLIAVEDTQGHETRYEYNIAG
700
DLTAVIAPDGRNGTQYDANGKAVRTTQGLTRSMYDAAGRVIKLTSENGSHTFRYDVLDRLIQETGFGRTQRYHNDLTGKLIRSEDEGLVTHWYD
800
EADRLTERTVKETAERNQYDERGWLTDISHISEGHRVAVHYRYDEKRLTGERQTVEHPQTRALLWQHETREHAYNAQGLANRCIPDLSFAVWLTYGSG
900
YLACMKLGDTPLEVYTRDLRHRETLRSFGRYELTAYTPAGQLSQHLNLSLSDRDYTWNDNGELIRISSFRQTRSYSTTGRLTGVHTAANLDTIRIP
1000
YATDPAGNRLPDPFELHPSSTLSMFPDNRIARDANLYRYDRHGRLEKTDLIPEGVIRTDDETRHYRYDSQRLVHYTRTQYEEPLVESRYLYDPLGRR
1100
VAKRVWRNRDLTGWMSLRKFPQVTWYQNDGDRLLTITQNDRTRIQTIYQPGSFTPLIRVETATGELAKTQRRSLADALQSGGEGCGSVFFPVVLVQMLD
1200
RLESEILADRVSESRRLASCGLTVEQMGNQMDPVVTPARKIHLNCHDRGLPLALISKEGTTEWCAEYDEWGNLLNEENPQLQLLIRLPGQQYDEES
← core | core-ext
1300
GLYNNRRRYDPLQCRYITQDPICLKGWNFYQYPLNPTNTDPLGLEVFPFPFPLPFWPKSFAQQQADDNAAKALTKNWNTASQRIPLSLNNEFGL
ALDITMIASRGHVDATGIDRVNDIINDRFWSGDKKPRDCLVQLDLCGDISAKDAESTQKAWNCHRSRQSNDEKKR

```

FIG. 5. Predicted amino acid sequence of the *rhaA* core ORF and its extension. A hydrophobic region containing 28 uncharged amino acids is underlined, as are Tyr-Asp dipeptides.

FIG. 6. Production of a fused *rhs* core peptide in maxicells. Cells were analyzed as described in Materials and Methods. Lane a, pBR325; lanes b through h, pBR325 derivatives carrying an *EcoRI*-*Bam*HI *rhs* core fragment ligated between analogous pBR325 sites: lane b, pJG1537 deleted between the *rhsA* *Hpa*I site (nucleotide 2179 of Fig. 4) and the vector *Ava*I site; lane c, pJG1609 Cam^r-*rhsA*::IS1; lane d, pJG1537 Cam^r-*rhsA*; lane e, pJG1575 Cam^r-*rhsC*; lane f, pJG1590 Cam^r-*rhsB*; lane g, pJG1598 Cam^r-*rhsC*::IS1; lane h, pJG1600 Cam^r-*rhsB*::IS1. The IS1 insertions were all in or near the respective Cam^r promoters. The positions of the 37-kDa Cam^r and the 26-kDa Tet^r proteins of pBR325 are indicated, as are the 130- and 135-kDa doublets (fs) that were specific to plasmids carrying the *rhs* gene fusions.

structed to obtain evidence for promoter activity, did not result in significant enzyme activity. This result might indicate the lack of expression or it might be that the fused protein was made but was simply not active. Preliminary results (D. Vlazny and C. Hill) indicate that the lack of enzyme activity was, in fact, due to low expression. Upstream of and overlapping the putative core promoter sequence is a large dyad symmetry (nucleotides 620 to 688). The arms of this dyad match at 22 of 31 positions. Experimentation is in progress to determine whether this sequence is involved in the control of *rhsA* core expression.

One of the most novel aspects of the *rhs* family is the extension of each of the core reading frames into adjacent divergent regions. These extensions would lengthen the putative protein products by 139 to 177 amino acids. If such core polypeptides serve a function at the cell surface, their extensions would provide a unique quality to each core protein. The first nine codons of the core extensions exhibit an intermediate level of conservation (Fig. 7). The predicted

amino acids for the fourth through the ninth of these codons are relatively well conserved, while those for the first three are more variable. In fact, no two of the seven sequences compared are identical at all of these first three codons. Beyond this intermediate region, only *rhsB* and *rhsD* show homology. If two single amino acid gaps are introduced into the *rhsB* sequence, the *rhsB* and *rhsD* amino acid sequences agree at 89 of 177 positions (Fig. 8). This is surprising, since the core of *rhsD* is the most divergent among the loci (20). The linkage of the conserved *rhs* core to several dissimilar sequences poses interesting evolutionary questions about the origin of these structures. The implication of a site-specific recombination mechanism capable of joining the end of the core to a variety of other sequences would seem strong.

The transition from core to core extension is accompanied by a dramatic change in both DNA base composition and codon usage. The G+C-rich core contains 61.7% G · C base pairs, while the core extension of *rhsA* is 40.5% G+C. Both of these values differ significantly from the 50% G+C generally found for *E. coli* (12). Although infrequently used codons appear in both segments of the ORF, the abnormal usage is more extreme in the core extension. For example, the rarely used arginine codons AGA and AGG appear 13 times in the core ORF (out of a total of 122 arginine codons). However, in the core extension, six out of nine arginine codons are either AGA or AGG. The Sharp and Li codon adaptation index (23) is 0.332 for the core ORF. This value is about average for low-to-moderate-abundance proteins. The index is only 0.157 for the *rhsA* core extension. This value is lower than that for any of 165 genes analyzed by Sharp and Li. Core-ORF extensions of *rhsB*, *rhsC*, and *rhsD* were remarkably similar to that of *rhsA* in base composition, having percentage GC values of 37.4, 40.7, and 36.7, respectively. Their codon adaptation indices were also low: *rhsB*, 0.189; *rhsC*, 0.185; and *rhsD*, 0.173. Whether these low-index values and low G+C contents reflect recent alien origins of the core extensions or whether they have some other significance is an open question.

A second 280-codon ORF, designated ORF-2 in Fig. 4, appears immediately after the extended core ORF of *rhsA*. The sequences surrounding the ATG codon of ORF-2 (Fig. 4) are predictive of an effective start codon. Fusions with the β -galactosidase gene indicated the presence of an active promoter preceding ORF-2, somewhere between nucleo-

FIG. 7. Comparison of the nucleotide sequence surrounding the *rhs* core ends. *rhsA'*, *rhsA''*, and *rhsC'* refer to partial core repetitions found in *rhsA* and *rhsC*. The first nucleotide of each sequence is within the *Bam*HI site (or its homolog) that occurs near the end of each core or core repeat (Fig. 1). For the *rhsA*, *rhsA'*, and *rhsA''* sequences, this nucleotide is 4422, 6175, and 7326, respectively, in Fig. 4. Consensus nucleotide and deduced amino acid sequences are shown. An x in the consensus sequences indicates that no nucleotide or amino acid appears at that position in four or more of the individual sequences. A period indicates that the nucleotide or amino acid at that position is identical to the one in the consensus.

FIG. 8. Nucleotide and amino acid comparisons of the core extensions for *rbsA*, *rbsB*, *rbsC*, and *rbsD* (GenBank accession no. M29717, M29718, and M29719 for *rbsB*, *rbsC*, and *rbsD*, respectively). The first nucleotide of each sequence is the one immediately following the proline codon terminating the core; for *rbsA*, this is nucleotide 4473 in Fig. 4. Amino acids which are common to both the *rbsB* and *rbsD* sequences are underscored.

tides 4311 and 4898. The segment covering nucleotides 4641 to 4672 contains sequences predicted to have strong promoter activity (10). It is interesting that this putative promoter is part of the sequence encoding the extension of the core ORF. The region containing ORF-2 is 35.8% G+C. The codon adaptation index for ORF-2 is only 0.196. The remaining 3.0 kb of *rbsA* beyond ORF-2 contains the two partial core repetitions. These 430- and 160-bp core fragments may be relics of ancestral versions that were successively replaced by more advantageous types, possibly because the newer versions had more suitable core extensions. Between the partial core repeats is an additional ORF, designated ORF-3 in Fig. 4. The last 1.4-kb segment of the *rbsA* locus is very A+T rich, having only 34.3% G+C. It has no features suggesting function.

The limits of *rbsA* and *rbsC* were established by comparing flanking homologies of *E. coli* K-12 with the *Rhs*⁰ strain ECOR 55 (Fig. 3). The nearly perfect sequence similarity surrounding *rbsA* ended abruptly between strains K-12 and ECOR 55, allowing a rather clear demarcation of the element. The 8.2-kb *rbsA* locus in strain K-12 and the unrelated 32-bp sequence in ECOR 55 are located between the same

chromosomal base pairs. The situation is somewhat less clear for *rbsC*. Similar to *rbsA*, the upstream and downstream homologies flanking *rbsC* could be readily located. If the boundaries of *rbsC* are taken as the positions where the nearly perfect sequence match ends, we conclude that the 9.6-kb *rbsC* locus of strain K-12 and a 10-bp sequence in ECOR 55 (as shown in Fig. 3b) are alternate residents of the same chromosomal location. However, this 10-bp sequence from ECOR 55 (CTTTTTTCGG) matches at 6 out of 10 positions with the extreme right end of *rbsC* (CATTTCTCAA). Thus, it is somewhat arbitrary whether we assign these 10 bp to the element or to the host. If we were to assign them to the host, we would conclude that the insertion or excision of *rbsC* had been between adjacent nucleotides.

The strain K-12 *rbsA* locus is abutted by two ORFs, both of which are oriented in the direction opposite to the *rbsA* core ORF (Fig. 4). ORF-us begins 28 nucleotides to the left of *rbsA* and extends leftward at least 531 nucleotides to the upstream *SalI* site. ORF-ds presumably initiates somewhere downstream from *rbsA* and terminates two nucleotides to the right of *rbsA*.

The fact that the *rbs* elements are not present in all *E. coli*

FIG. 9. Internal repetitions predicted for the *rhsA* core peptide. Sequence blocks in panels a, b, and c are each continuous. The first and last amino acids of each block are numbered according to the conventions used in Fig. 5. Line breaks are introduced to optimize the alignment, and residues that match the motif GxxxRYxYDxxGRL(I or T) are boxed.

raises the question of whether they were present at the point of divergence during the evolution of strains K-12 and ECOR 55. In other words, is the difference due to loss or acquisition? Unfortunately, this important question cannot be answered at this time. It has been shown that *Salmonella typhimurium* DNA does not have homology to the *rhs* core (11). The *rhs* sequences might comprise a family of accessory genetic elements such as transposable elements or prophages. If this were true, then examination of their terminal sequences could give a clue as to their nature, since accessory elements (3) typically exhibit small direct or inverted repetitions that serve as either substrates or products of site-specific recombination (7). Examination of the termini of *rhsA* and *rhsC* (Fig. 3) gave no indication of the presence of short repetitive sequences flanking the *rhs* elements. Thus, they fail this criterion for transposable elements and prophages. It also seems noteworthy that *rhsA* and *rhsC* share no homology at their termini. This contrasts with the 98% sequence similarity shared by their cores. The *rhs* termini, nevertheless, do exhibit some interesting sequence features. The left end of *rhsA* is very purine rich, 21 of the first 26 bases being either A or G. The right end is very pyrimidine rich, 19 of the last 22 bases being either C or T. These particular features do not appear in *rhsC*. However, a run of 12 consecutive pyrimidines is present in the region of the ECOR 55 chromosome corresponding to the position of *rhsC* insertion. Another line of experimentation also failed to provide evidence that the *rhs* loci are transposable elements. A property of transposable elements is their ability to cause cointegrant formation. Application of the system described by Grindley and Joyce (9) for the detection of cointegrate formation between plasmids bearing transposons and a modified F plasmid gave negative results for *rhsA* (C. Hill,

unpublished observations). We anticipate that the patterns of *rhs* occurrence among natural *E. coli* isolates will provide clues as to their origin.

ACKNOWLEDGMENTS

This work was supported by Public Health Service grant GM16329 from the National Institutes of Health. Computer resources were provided by the BIONET National Computer Resource for Molecular Biology, which is funded by the Biomedical Research Technology Program, Division of Research Resources, National Institutes of Health, grant P41 RR01685.

LITERATURE CITED

1. Bolivar, F. 1978. Construction and characterization of new cloning vehicles. III. Derivatives of plasmid pBR322 carrying unique *EcoRI* sites for selection of *EcoRI* generated recombinant DNA molecules. *Gene* 4:121-136.
2. Brody, H., A. Greener, and C. W. Hill. 1985. Excision and reintegration of the *Escherichia coli* K-12 chromosomal element e14. *J. Bacteriol.* 161:1112-1117.
3. Campbell, A. 1981. Evolutionary significance of accessory DNA elements in bacteria. *Annu. Rev. Microbiol.* 35:55-83.
4. Capage, M., and C. W. Hill. 1979. Preferential unequal recombination in the *glyS* region of the *Escherichia coli* chromosome. *J. Mol. Biol.* 127:73-87.
5. Casadaban, M. J., J. Chou, and S. N. Cohen. 1980. In vitro gene fusions that join an enzymatically active β -galactosidase segment to amino-terminal fragments of exogenous proteins: *Escherichia coli* plasmid vectors for the detection and cloning of translational initiation signals. *J. Bacteriol.* 143:971-980.
6. Casadaban, M. J., A. Martinez-Arias, S. K. Shapira, and J. Chou. 1983. β -Galactosidase gene fusion for analyzing gene expression in *Escherichia coli* and yeast. *Methods Enzymol.* 100:293-308.
7. Craig, N. L., and N. Kleckner. 1987. Transposition and site-

- specific recombination, p. 1054–1070. 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. 2. American Society for Microbiology, Washington, D.C.
8. **Felmlee, T., S. Pellett, and R. A. Welch.** 1985. Nucleotide sequence of an *Escherichia coli* chromosomal hemolysin. *J. Bacteriol.* **163**:94–105.
 9. **Grindley, N. D. F., and C. M. Joyce.** 1980. Genetic and DNA sequence analysis of the kanamycin resistance transposon Tn903. *Proc. Natl. Acad. Sci. USA* **77**:7176–7180.
 10. **Hawley, D. K., and W. R. McClure.** 1983. Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucleic Acids Res.* **11**:2237–2255.
 11. **Lin, R.-J., M. Capage, and C. W. Hill.** 1984. A repetitive DNA sequence, *rhs*, responsible for duplications within the *Escherichia coli* K-12 chromosome. *J. Mol. Biol.* **177**:1–18.
 12. **Normore, W. M.** 1975. Guanine-plus-cytosine (GC) composition of the DNA of bacteria, fungi, algae and protozoa, p. 65–235. In G. D. Fasman (ed.), *Handbook of biochemistry and molecular biology*, 3rd ed., vol. 2. CRC Press, Inc., Boca Raton, Fla.
 13. **Norlander, J., T. Kempe, and J. Messing.** 1983. Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. *Gene* **26**:101–106.
 14. **Ochman, H., and R. K. Selander.** 1984. Standard reference strains of *Escherichia coli* from natural populations. *J. Bacteriol.* **157**:690–693.
 15. **Ochman, H., and A. C. Wilson.** 1987. Evolutionary history of enteric bacteria, p. 1649–1654. 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. 2. American Society for Microbiology, Washington, D.C.
 16. **Petersen, G. B., P. A. Stockwell, and D. F. Hill.** 1988. Messenger RNA recognition in *Escherichia coli*: a possible second site of interaction with 16S ribosomal RNA. *EMBO J.* **7**:3957–3962.
 17. **Petes, T. D., and C. W. Hill.** 1988. Recombination between repeated genes in microorganisms. *Annu. Rev. Genet.* **22**: 147–168.
 18. **Prentki, P., F. Karch, S. Iida, and J. Meyer.** 1981. The plasmid cloning vector pBR325 contains a 482 base-pair-long inverted duplication. *Gene* **14**:289–299.
 19. **Riley, M., and S. Krawiec.** 1987. Genome organization, p. 967–981. 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. 2. American Society for Microbiology, Washington, D.C.
 20. **Sadosky, A. B., A. Davidson, R.-J. Lin, and C. W. Hill.** 1989. *rhs* gene family of *Escherichia coli* K-12. *J. Bacteriol.* **171**:636–642.
 21. **Sancar, A., A. M. Hack, and W. D. Rupp.** 1979. Simple method for identification of plasmid-coded proteins. *J. Bacteriol.* **137**: 692–693.
 22. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**:5463–5467.
 23. **Sharp, P. M., and W.-H. Li.** 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**:1281–1295.
 24. **Shine, J., and L. Dalgarno.** 1974. The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to non-sense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci. USA* **71**:1342–1346.
 25. **Signas, C., G. Raucchi, K. Jonsson, P.-E. Lindgren, G. M. Anantharamaiah, M. Hook, and M. Lindberg.** 1989. Nucleotide sequence of the gene for a fibronectin-binding protein from *Staphylococcus aureus*: use of this peptide sequence in the synthesis of biologically active peptides. *Proc. Natl. Acad. Sci. USA* **86**:699–703.
 26. **Silhavy, T. J., M. L. Berman, and L. W. Enquist.** 1984. Experiments with gene fusions. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
 27. **von Heijne, G.** 1981. Membrane proteins: the amino acid composition of membrane-penetrating segments. *Eur. J. Biochem.* **120**:275–278.
 28. **von Heijne, G.** 1985. Signal sequences: the limits of variation. *J. Mol. Biol.* **184**:99–105.