Nucleic Acids Research, Vol. 19, No. 25 7177–7183

The *RhsD-E* subfamily of *Escherichia coli* K-12

Alesia B.Sadosky+, Jane Antal Gray and Charles W.Hill*
Department of Biological Chemistry, The Milton S.Hershey Medical Center, The Pennsylvania State University, Hershey, PA 17033, USA

EMBL accession nos X60997 – X60999 (incl.)

ABSTRACT

The *Escherichia coli* K-12 chromosome contains a family of five large, unlinked sequences known as the *Rhs* elements. They share several complex homologies, the most prominent being a 3.7 kb *Rhs* core. The elements are divided into two subfamilies, *RhsA-B-C* and *RhsD-E*, according to the sequence similarities of the cores. The *RhsD* core is 3747 bp long compared to 3714 bp for *RhsA*. Despite a 22% sequence divergence, the *RhsD* core conserves features previously noted for *RhsA*. Similar to *RhsA*, the *RhsD* core maintains a single ORF, the start codon coinciding with the first nucleotide of the homology. The *RhsD* core-ORF continues 177 codons beyond the homology, resulting in a carboxy terminal extension unrelated to that of *RhsA*. The *RhsD* core retains all 28 copies of the repeated motif GxxxRYxYDxxGRL(I/T) seen in *RhsA*. The other member of the *RhsD-E* subfamily, *RhsE*, has been mapped to minute 32 of the *E. coli* map. It appears defective in that it contains only the last 1550 bp of the 3.7 kb core. Its sequence is more closely related to that of *RhsD* than *RhsA*. In addition, *RhsE* and *RhsB* share a 1.3 kb homology, known as the H-repeat. The H-repeats from *RhsE* and *RhsB* are more closely related than their cores, showing only 1% nucleotide divergence.

INTRODUCTION

The *Rhs* elements comprise a family of large, complex genetic homologies of *Escherichia coli*. Their existence was first suggested by the observation of Folk and Berg (1) that the chromosomal segment containing the *glyS* locus was subject to frequent amplification in *E. coli* K-12. The novel feature of this region responsible for this phenomenon is the presence of two *Rhs* elements, *RhsA* and *RhsB*. Recombination between their homologous cores causes amplification of the intervening 140 kb region including the *glyS* locus (2,3). *E. coli* K-12 contains five *Rhs* elements, and collectively they account for nearly 1% of the genome. While they are widespread, not all wild *E. coli* contain *Rhs* elements, an indication that they probably do not encode essential genetic information (4). Their function is yet to be determined.

Each *Rhs* element shares one or more homologies with the others, but each also contains sequences unique to the individual element. The largest homology is the 3.7 kb *Rhs* core. The *RhsA* and *RhsB* cores are sufficiently similar to form a heteroduplex resistant to limited S1-nuclease digestion (3). The *Rhs* core homology begins with an ATG start codon, initiating an ORF that extends through and beyond the entire length of the core into divergent downstream sequences. In the case of *RhsA*, the core consists of 1238 codons, with the unique core extension contributing an additional 139 codons to the ORF (4). The predicted protein product is remarkable in a number of ways, including its large size, extreme hydrophilicity and internal repetition. The structure of the *RhsA* element is shown schematically in Fig. 1.

Four of the elements, *RhsA*, *RhsB*, *RhsC* and *RhsD*, have been cloned and located on the *E. coli* genetic map (5). The existence of the fifth element, *RhsE*, was inferred by Southern analysis of genomic DNA by virtue of the fact that an additional hybridizing fragment could not be assigned to the other four elements (3). The fragments associated with *RhsD* and *RhsE* gave much weaker hybridization signals compared to the other three elements when a *RhsA* core probe was used. These weak signals were specifically reduced at higher stringencies (3), suggesting that the *Rhs* elements could be divided into two subfamilies based on core sequence divergence. In this paper, we report the cloning and mapping of *RhsE* and describe various features of the *RhsD-E* subfamily.

MATERIALS AND METHODS

Bacterial and phage strains

ECOR #39 (6) was supplied by Robert Selander. Tn*10* strains used for mapping *RhsE* were supplied by Peter Keumpel. Lambda 274 from the Kohara miniset (7) was supplied by Ken Rudd. Other *E. coli* strains have been described previously (5). Procedures for bacterial growth, P1 transduction and transformation were as specified previously (8).

* To whom correspondence should be addressed

+ Present address: Department of Microbiology, College of Physicians and Surgeons, Columbia University, New York, NY 10032, USA

Plasmid construction

The vector pUC19 (9) was used for the construction of most *Rhs* clones. Techniques for screening plasmid pools prepared from digests of wild *E. coli* DNA have been described (3). Preparation of plasmids containing *RhsD* and flanking regions from *E. coli* K-12 has also been described (5). *RhsE* was isolated from the K-12 chromosome in a manner similar to *RhsD*. Initially, the 1 kb *Sal* I-*Hin*dIII fragment from pRL390 (3) was transferred into pUC19, creating pAS3154. The *Sal* I site from pAS3154 was then destroyed by blunt ending using Klenow polymerase, resulting in pAS3156. Finally, the 333 bp *Acc* I fragment internal to the insert of pAS3156 was replaced with the *Sal* I fragment containing the Kan^r determinant from pUC4K (10), creating pAS3158. pAS3158 was introduced into the *polA1* strain, CH1330. The resulting strain, CH3159, had the plasmid integrated into the host chromosome at the site of insert homology (Fig. 2a). Recombinant clones carrying sequences upstream (pAS3161) and downstream (pAS3165) from the *RhsE* element were then cloned by selective digestion and ligation of genomic DNA flanking the integrated plasmid. pAS3161 and pAS3165 were used to generate the restriction map of *RhsE*.

DNA sequencing

Dideoxy sequencing of double stranded plasmid DNA was performed as specified previously (4). Oligonucleotide primers necessary for sequencing were purchased from the Hershey Medical Center Macromolecular Core Facility. DNA sequences have been submitted to the EMBL Data Library with the following accession numbers: K-12 *RhsD*, X60997; K-12 *RhsE*, X60998; ECOR #39 alternative to *RhsD*, X60999.

RESULTS

Cloning and mapping *RhsE*

The cloning, mapping and partial sequence analysis of *RhsD* have been reported previously (5), but details of *RhsE* have not been described. Better understanding of the *RhsD-E* subfamily required cloning *RhsE*. *RhsE* was originally defined on the basis of an extra *Sal* I-*Hin*dIII fragment that showed homology to an *Rhs* core-specific probe. Since the other *Rhs* elements maintained a conserved 3.7 kb core, it seemed probable that this 1 kb *Sal* I-*Hin*dIII fragment represented an incomplete version of the *RhsE* element. In order to isolate genomic DNA flanking the 1 kb fragment, we used an approach that had proven successful for isolating *RhsC* and *RhsD*. The basic cloning strategy was as follows. Initially, a fragment of interest is cloned from the bacterial chromosome. The recombinant plasmid is then forced into the chromosome by recombination between the insert and its chromosomal homolog. The drug resistance of the inserted plasmid aids in both the genetic mapping and cloning of regions adjacent to the starting fragment (5,8). Cloning of the original 1 kb *Sal* I-*Hin*dIII fragment from *RhsE* was previously reported (3). A more useful form, pAS3158, was prepared by inserting the 1 kb fragment into pUC19 and then placing a Kan^r determinant within the cloned insert (Materials and Methods). pAS3158 was used to transform a *polA1* recipient, selecting for both Amp^r and Kan^r. Since this vector requires the PolA function for plasmid replication, only those transformants which have the recombinant plasmid integrated into the host chromosome survive. The result of pAS3158 integrated at the *RhsE* locus was strain CH3159 whose structure is shown in Fig. 2.

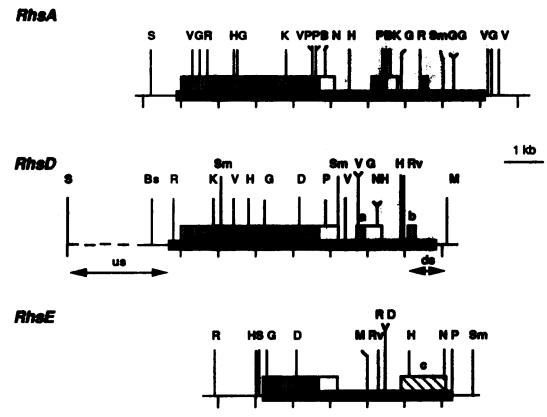


Figure 1. Schematic representation of *Rhs* elements of *E. coli* K-12. The complete sequence of *RhsA* (4) and a preliminary description of *RhsD* (5) have been published. Details of *RhsD* and *RhsE* are provided in the text. The solid bar indicates the extent of the individual elements. The stippled bars indicate whole or partial core homologies. The open bars indicate the various unique extensions of the core-ORFs. The hatched bar indicate the H-repeat. Homology blocks discussed in the text are designated a, b and c. Fragments used as probes in the cloning of *RhsD*^{0} sites from ECOR strains are designated *us* and *ds*. Restriction site designations are: B, *Bam*HI; Bs, *Bst*EII; D, *Hin*dIII; G, *Bgl* I; H, *Hin*cII; K, *Kpn* I; M, *Mlu* I; N, *Nco* I; P, *Pst* I; R, *Eco*RI; Rv, *Eco* Rv; S, *Sal* I; Sm, *Sma* I; V, *Pvu* II. Only selected *Bst*EII and *Eco*Rv sites are shown.

DNA flanking the integrated plasmid was cloned from CH3159 by digesting genomic DNA with either *Bam*HI or *Sal* I and religating. This resulted in two new, enlarged plasmids which could be recovered by transforming a *polA*$^+$ strain, selecting Kan^r. pAS3161 contained genomic DNA to the left of the original fragment, and pAS3165 contained genomic DNA to the right. Analysis of these clones resulted in a restriction map covering a 16.2 kb region of the genome (Fig. 2b). This map was compared to the physical map of the chromosome (7), and a very similar restriction pattern was observed for the region between coordinates 1540–1550. This assignment was confirmed by the observation that the recombinant phage from the Kohara miniset, lambda 274, contained sequences identical to those contained in pAS3161 and pAS3165. We also verified the position of *RhsE* through genetic means, using phage P1 cotransduction. These experiments used a series of mutants with Tn*10* insertions (11) linked to the region identified by the physical mapping. The Tn*10* mutants served as donors, and the elimination of the Amp^r or Kan^r character of CH3159 among Tet^r transductants was scored. The cotransduction frequencies, summarized in Fig. 3, show that *RhsE* is located near min. 32 of the *E. coli* map.

Sequence of the *RhsD* and *RhsE* cores

In previous work, we presented evidence that *RhsD* contains a core analogous to the ones found for the *RhsA-B-C* family. We identified sequences in *RhsD* that were homologous to both ends of the *RhsA* core, and observed that these sequences were separated by approximately 3.7 kb, the length of the *RhsA* core. We wished to determine whether some of the unusual features observed for the *RhsA* core-ORF were also present in *RhsD* despite the apparent sequence divergence. Therefore, we sequenced the entire *RhsD* core and compared it to that of *RhsA*. The 4846 bp sequence extending from 268 bp before the *Eco*RI site to the first *Pvu* II site (Fig. 1) has been submitted to the

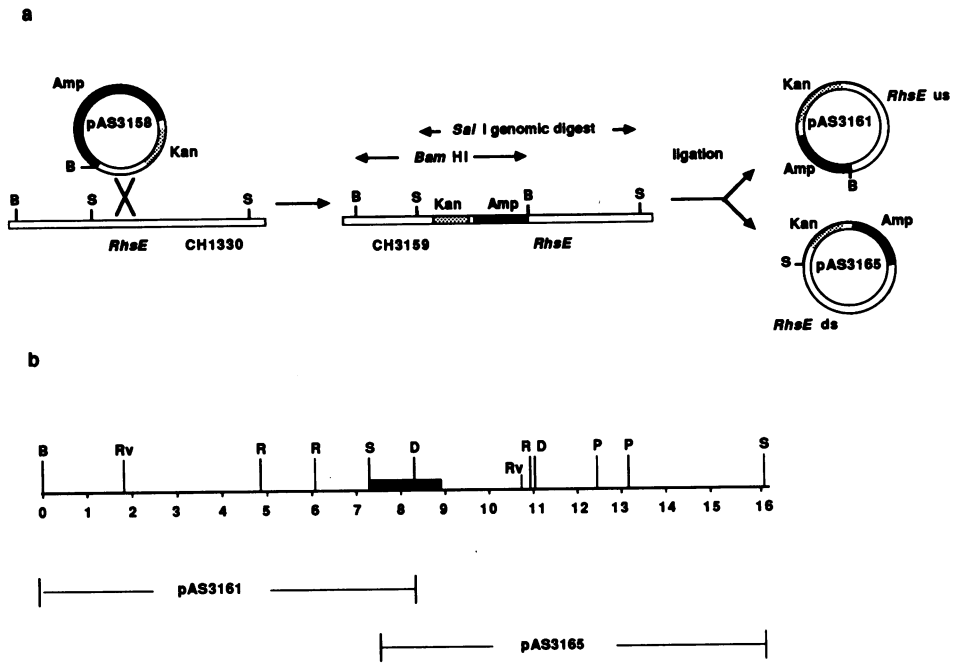


Figure 2. *RhsE* cloning. a) Illustration of the procedure used to clone *RhsE* upstream (us) and downstream (ds) sequences. Preparation of plasmid pAS3158 is described in Materials and Methods. Its insert is the 1 kb *Sal* I-*Hin*dIII fragment from *RhsE* into which was placed Kan^r. pAS3158 became integrated at the *RhsE* chromosomal locus upon transformation of the *polA1* strain CH1330, generating CH3159. Adjacent *E. coli* sequences were isolated by digestion of CH3159 genomic DNA with either *Bam*HI or *Sal* I, religation to produce plasmids containing the vector and adjacent host DNA, and transformation of a recipient *polA*$^+$ strain, selecting Amp^r Kan^r. Solid black lines denote vector sequences; stippled regions, Kan^r; and open regions, chromosomal DNA. b) Restriction map of the *E. coli* K-12 *RhsE* element. Two overlapping clones carrying sequences upstream (pAS3161) and downstream (pAS3165) from *RhsE* were used to generate the 16.2 kb restriction map. Numbering refers to length in kb. Restriction sites are designated as in Fig. 1.

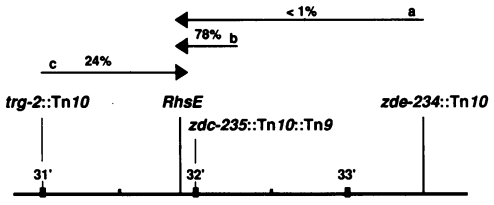


Figure 3. *RhsE* mapping by P1 transduction. P1 lysates of the Tn*10* strains were used to infect CH3159, selecting Tet^r and scoring for loss of the drug resistance provided by the integrated plasmid. Donor stains used were: cross a, PK1110 (*zde-234*::Tn*10*); cross b, PK1269 (*zdc-235*::Tn*10*); cross c, PK1148 (*trg-2*::Tn*10*).

EMBL data library (accession no. X60997). We found the *RhsD* core to be slightly larger than that of *RhsA*, 3747 bp compared to 3714 bp. Like *RhsA*, the entire *RhsD* core comprised a single ORF. The predicted amino acid sequences of the *RhsD* and *RhsA* cores were very similar throughout (Fig. 4). The length differences were primarily due to two non-homologies. In one case, a fifteen codon sequence in *RhsD* (residues #261 – #275) replaced an unrelated 10 codon sequence in *RhsA*. In the second case, *RhsD* contained a block of seven additional codons not present in *RhsA* (residues #835 – #841).

We knew from hybridization studies that at least part of the *RhsE* core homology was contained in the 1 kb *Sal* I-*Hin*dIII fragment (Fig. 2). To clarify the arrangement, this fragment and adjacent regions were sequenced. The 2440 bp sequence beginning at the *Sal* I site has been submitted to the EMBL data library (accession no. X60998). The results showed that *RhsE* contained only part of the core, corresponding to the last 1550 bp of the *RhsD* core. This core homology began 14 bp downstream from the *Sal* I site (Fig. 1). To the left of the *Sal* I site, the sequence was not similar to any known *Rhs* sequence. As expected from the hybridization studies, nucleotide sequencing revealed that the *RhsE* core was more closely related to *RhsD* than to *RhsA*. The translation of the *RhsE* partial core is compared to *RhsD* and *RhsA* in Fig. 4. *RhsE* retained the seven codon block (residues #835 – #841) that was present in *RhsD* but not in *RhsA*. In fact, the predicted amino acid sequence differed from *RhsD* at only 21 of the remaining 509 positions, while it differed from *RhsA* at 105 positions. The ORF of the partial *RhsE* core extended 158 codons beyond the 3′ end of the core into adjacent DNA, as compared to 177 codons for the *RhsD* core-extension and 139 codons for *RhsA* (Fig. 4). Except for the intermediate level of conservation observed within the first nine codons immediately following the core (4), these three extensions showed no sequence similarity at either the nucleotide or amino acid level. We previously reported that the extensions of *RhsD* and *RhsB* do have about 50% amino acid similarity (4).

Boundaries of the *RhsD* element

Some wild *E. coli* do not exhibit homology with *Rhs* core-specific probes. It is not yet known whether the ancestral lineage of these strains never possessed such elements or whether they were once present but have been lost through deletion. Nevertheless, we have taken advantage of this situation to define an *Rhs* element as all of the DNA in an *Rhs*$^+$ strain that is contiguous with the

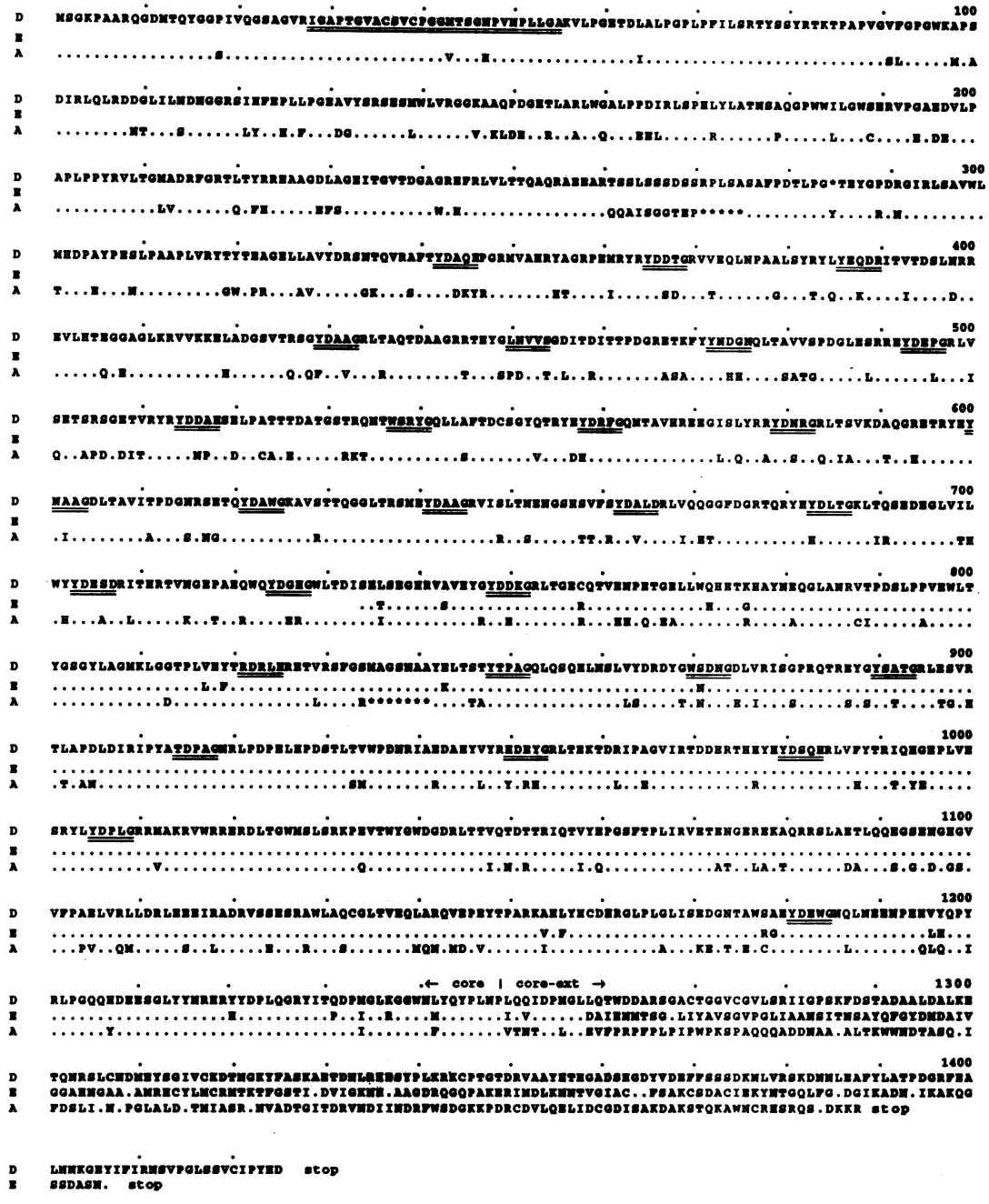


Figure 4. Alignment of the proteins predicted from the cores of *RhsD*, *RhsE* and *RhsA*. The *RhsA* translation is derived from the published nucleotide sequence (4). The sequences of the *RhsD* and the *RhsE* cores were determined for this work and have been submitted to the databases (Materials and Methods). The *Rhs* core has been defined to extend through the proline residue at position #1249 (4). The extension of the core-ORF begins with the next codon and proceeds into a region that is unique to each element. The potential membrane spanning region between residues #28 – #55 is underlined. Also underlined are portions of each repeat unit. The five residues underlined correspond to the sequence YDxxG from the consensus GxxxRYxYDxxGRL(I/T). A dot indicates that the amino acid is identical to the one above in *RhsD*. An * is inserted into a sequence as a filler to facilitate alignment.

core homology, yet absent from the *Rhs*^{o} strain. Using this criterion, we previously found that *E. coli* K-12 *RhsA* consists of 8.2 kb of DNA replacing 32 bp of DNA in an *Rhs*^{o} strain, while *RhsC* consists of 9.6 kb of DNA replacing a 10 bp unrelated segment. The same approach has now been used to define the limits of *RhsD*.

The first step was to identify restriction fragments containing sequences that flank *E. coli* K-12 *RhsD*. By our definition, such fragments should show homology with genomic DNA from *Rhs*^{o} strains. We found that the 2.85 kb *Sal* I-*Eco*RI fragment upstream from the *RhsD* core and the 1.07 kb *Eco*RV-*Mlu* I fragment downstream both hybridized with genomic DNA from the *Rhs*^{o} strain ECOR #39. These probes are identified as us and ds in Fig. 1. Using these probes, we screened recombinant plasmids from a library of ECOR #39 and found homology with an 11 kb *Sal* I-*Bgl* II insert (pAS3135). The limits of *RhsD* were

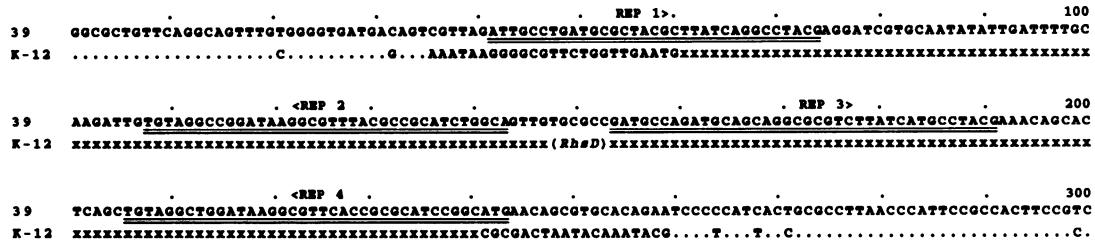


Figure 5. Limits of *RhsD*. Sequences flanking *RhsD* of *E. coli* K-12 (lower) are aligned with the homologous sequences from ECOR #39 (upper). The points of divergence and convergence are separated by 224 bp in ECOR #39 and 7.3 kb in *E. coli* K-12. A dot indicates that the base in *E. coli* K-12 is identical to the one above in ECOR #39, while xxxx indicates the position of the 7.3 kb *RhsD* element in K-12. The positions of sequences matching the REP consensus A(T/A)TGCC(G/T)GATG.CG(G/A)CG(C/T).....(G/A)CG(C/T)CTTATC(C/A)GGCCTAC(A/G) (15,16) are underlined in the ECOR #39 insert.

narrowed by the finding that a 2.0 kb *Bst*EII-*Mlu* I subclone (pAS3141) retained homology to both upstream and downstream probes. Preliminary sequence analysis showed that the *Bst*EII site of pAS3141 was analogous to the *Bst*EII site that occurs approximately 780 bp upstream of the *RhsD* core in *E. coli* K-12 (Fig. 1). The limits of the non-homologies that distinguish *E. coli* K-12 from ECOR #39 were narrowed to the regions 300−400 bp to the right of the *Bst*EII site and 200−300 bp to the left of the *E. coli* K-12 *Mlu* I site (Fig. 1) Synthetic oligonucleotides based on known sequences from *E. coli* K-12 were then used to sequence portions of the DNA cloned from ECOR #39. Sequences nearly identical to *E. coli* K-12 were found, and the site where the sequence diverged from *E. coli* K-12 was identified. In ECOR #39, the left and right divergences were separated by a stretch of 224 bp not found in *E. coli* K-12 (Fig. 5). Instead, these 224 bp were replaced by 7.3 kb of DNA which we consider to constitute the *RhsD* element.

The limits of *RhsE* have not yet been defined at the sequence level. However, the 0.5 kb *Pst* I-*Sma* I fragment downstream from the core (Fig. 1) hybridizes well with *Rhs*^{o} genomic DNA, placing the right hand limit of *RhsE* near the *Pst* I site.

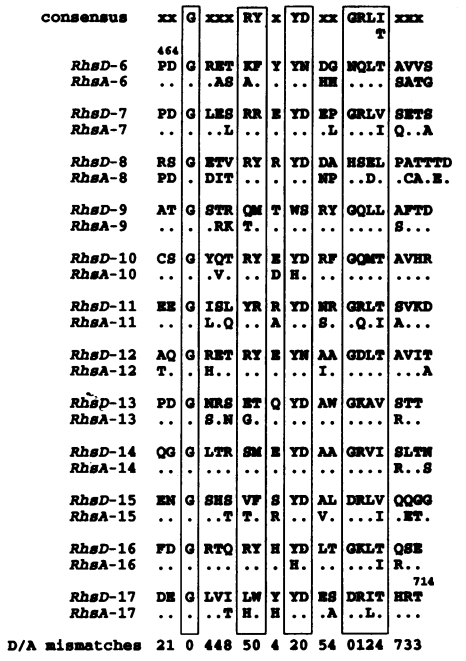
DISCUSSION

The protein potentially encoded by the *Rhs* core is a unusual one (4). The peptide product of *RhsA*, including the core-ORF extension, would have a molecular mass of 156 kilodaltons, making it one of the largest proteins of *E. coli*. The *RhsA* core protein would be extremely hydrophilic, due in part to the fact that the three hydroxylated amino acids, serine, threonine and tyrosine, constitute 20.1% of the residues. The most striking feature of the core is the presence of a motif whose consensus is GxxxRYxYDxxGRL(I/T). This motif appears in the *RhsA* core 28 times. Finally, a potential membrane spanning domain has been noted near the amino terminus of the protein. Collectively these features have precedents among certain microbial cell surface or binding proteins. Whether the *Rhs* core product has such a function has not been determined. A major objective of the work presented here was to determine the relationship of the *RhsD-E* subfamily to *RhsA*, and to establish whether the *RhsD* core maintains the features observed for *RhsA* despite the apparent sequence divergence.

Nucleotide sequence comparison showed that the *RhsD* core is slightly longer than *RhsA*, 3747 bp vs. 3714 bp. The length differences are accounted for by differences at three positions (Fig. 4). These differences are a 15 codon sequence in *RhsD* that replaces an unrelated 10 codon sequence in *RhsA* (residues #261 − #275), a one codon deletion that occurs in *RhsD* (position #285), and a seven codon deletion (residues #835 − #841) that occurs in *RhsA*. The remaining 3681 bp of the *RhsD* and *RhsA* cores can be aligned precisely, differing by 22.1% or 813 of the 3681 bp. This degree of divergence is comparable to the more extreme divergences observed for homologous loci in *E. coli* and *Salmonella typhimurium* (12,13). It is important to note that the 813 nucleotide changes cause only 258 changes in the predicted *Rhs* core amino acid sequence. Since theoretically 73% of all possible nucleotide substitutions result in non-synonymous codons, the accumulation of only 258 amino acid changes indicates a strong sequence constraint on the evolution of the *Rhs* core protein. The conservation of the amino acid sequence near the ends of the cores appears greater than in the interior. Only 4 amino acid substitutions are produced by 46 nucleotide changes in the first 90 codons, and only 2 amino acid substitutions are produced by 19 nucleotide substitutions in the last 42 codons.

The general features noted previously for the *RhsA* core-ORF are conserved for *RhsD* as well. Although two amino acid substitutions occur in the postulated membrane spanning region (residues #28 − #55), the amino acid sequence in *RhsD* is still compatible with a membrane spanning function. The first 20 amino acids of this sequence in *RhsA* and *RhsD* get scores of 23.3 and 22.3 kcal/mol for their transfer free energy, where a score above 20 is predictive of a membrane spanning helix (14). The core of *RhsD* is even richer in hydroxylated amino acids than *RhsA*, 22.5% compared to 20.1%. This increase is due largely to the eight serine residues present in the fifteen amino acid segment (residues #261 − #275) unique to *RhsD*.

The conservation of the repeated motif, GxxxRYx-YDxxGRL(I/T), is of special interest. We were interested in knowing how the two cores compared as to the number of repetitions and the amino acid sequence of individual repetitions. The total number of repetitions was identical for both cores; analogs of all 28 repeats in *RhsA* were found in *RhsD*. Their positions in *RhsD* have been marked in Fig. 4 where the residues corresponding to YDxxG in the consensus are underlined. From previous work on *RhsA* (4), it was observed that a block of 12 such repetitions between residues #464 and #714 were particularly regular in both their agreement with the consensus motif and in their spacing. With a single exception, the motif was repeated at 20 or 21 residue intervals. These 12 repetitions from each of the cores are aligned in Fig. 6. The number of mismatches between *RhsD* and *RhsA* is tabulated for each of 20 positions, beginning with the two positions preceding the first



consensus	xx	G	xxx	RY	x	YD	xx	GRLI T	xxx	
	464									
RhsD-6	PD	G	RET	KF	Y	YN	DG	NQLT	AVVS	
RhsA-6	..	.	.AS	A.	.	..	HH	....	SATG	
RhsD-7	PD	G	LES	RR	E	YD	EP	GRLV	SETS	
RhsA-7	..	.	..L	..	.	..	.L	...I	Q..A	
RhsD-8	RS	G	ETV	RY	R	YD	DA	HSEL	PATTTD	
RhsA-8	PD	.	DIT	..	.	..	NP	..D.	.CA.E.	
RhsD-9	AT	G	STR	QM	T	WS	RY	GQLL	AFTD	
RhsA-9	..	.	.RK	T.	.	..	..	....	S...	
RhsD-10	CS	G	YQT	RY	E	YD	RF	GQMT	AVHR	
RhsA-10	..	.	.V.	..	D	H.	..	....	....	
RhsD-11	EE	G	ISL	YR	R	YD	HR	GRLT	SVKD	
RhsA-11	..	.	L.Q	..	A	..	S.	.Q.I	A...	
RhsD-12	AQ	G	RET	RY	E	YN	AA	GDLT	AVIT	
RhsA-12	T.	.	H..	..	.	..	I.	....	...A	
RhsD-13	PD	G	NRS	ET	Q	YD	AW	GKAV	STT	
RhsA-13	..	.	S.N	G.	.	..	..	....	R..	
RhsD-14	QG	G	LTR	SM	E	YD	AA	GRVI	SLTN	
RhsA-14	..	.	...	..	.	..	..	....	R..S	
RhsD-15	EN	G	SHS	VF	S	YD	AL	DRLV	QQQG	
RhsA-15	..	.	..T	T.	R	..	V.	...I	.ET.	
RhsD-16	FD	G	RTQ	RY	H	YD	LT	GKLT	QSE	
RhsA-16	..	.	...	..	.	H.	..	...I	R..	
									714	
RhsD-17	DE	G	LVI	LW	Y	YD	ES	DRIT	HRT	
RhsA-17	..	.	..T	H.	H	..	.A	..L.	...	
D/A mismatches	21	0	448	50	4	20	54	0124	733	

Figure 6. Alignment of 12 of the 28 repeat units from a select region of the *RhsD* and *RhsA* cores. The consensus sequence derived previously from the *RhsA* sequence (4), is shown at the top. Residues #464 – #714 from *RhsD* and *RhsA* (Fig. 4) are arranged to match each successive repeat with both the consensus and with each other. Note that the consensus was derived by considering all 28 repetitions in *RhsA*. Positions specifically identified in the consensus are enclosed in boxes. The number of times that *RhsD* differs from *RhsA* at each position within the repeated unit is tabulated at the bottom.

glycine of the consensus. With one exception, *RhsD* and *RhsA* match well at positions which are specifically identified in the consensus. In the case of the first glycine in the consensus (position 3 in the repeat unit as depicted), all 12 repeats have glycine in both of the sequences. The conservation between *RhsD* and *RhsA* holds even if the residue in question does not agree with the consensus. For example, in the case of the second glycine (position 14 in the repeat unit), four of the repeats do not have the consensus glycine. Nevertheless, *RhsD* and *RhsA* have the same alternative amino acid in each repeat. The least conserved of the specified positions is the first arginine (position 7 in the repeat unit) where *RhsD* and *RhsA* differ five times. Interestingly, two positions not specified by the consensus are nevertheless highly conserved between the two cores. These are the two amino acids immediately preceding the first glycine. Of the 12 comparisons, *RhsD* and *RhsA* match ten times at the first of these positions and eleven times at the second.

Taken together, these sequence comparisons suggest that the *RhsA* and *RhsD* cores diverged a very long time ago, roughly at the time of the divergence of *E. coli* and *S. typhimurium* as separate species. The internal repetitions were fully established at the time these core subfamilies diverged, and the maintenance of this repetitive pattern through so much evolutionary time clearly indicates a strong functional constraint on the pattern. We noted previously that the *RhsA* core and its core-ORF extension are anomalous among *E. coli* sequences in their GC content, the core being GC rich and the extension being GC poor. These anomalies hold for *RhsD*. Its core sequence is 63.5% GC, while its core-extension is only 36.7%. This departure from the 50% GC content observed for most *E. coli* sequences suggests that it originated outside of the *E. coli* species and entered *E. coli* relatively recently.

A combination of physical and genetic mapping has placed *RhsE* at minute 32 on the *E. coli* map. This is in the middle of the large phenotypically silent region containing the termination of replication. The *RhsE* core is incomplete in that it contains only the distal 1550 bp of the 3.7 kb core homology. It is probable that the proximal portion of the core was lost by deletion in a recent ancestor of *E. coli* K-12, since we have found that the *RhsE* element of ECOR strains closely related to *E. coli* K-12 retain proximal core homology (unpublished). We also observe that the *RhsE* core is more closely related to *RhsD* than to *RhsA*. At the nucleotide level, *RhsD* and *RhsE* differ by 4.1% or 64 substitutions within the 1550 bases of the residual core. This compares to the 22.1% divergence between *RhsD* and *RhsA*. On the other hand, the 4.1% divergence between *RhsD* and *RhsE* is considerably greater than the roughly 1% divergence observed between members of the *RhsA-B-C* family (5). Furthermore, the mismatches between *RhsD* and *RhsE* are quite unevenly distributed. There is not a single mismatch in the 846 bp region encoding amino acids 872 – 1153 (Fig. 4), while there are 36 mismatches in the preceding 415 bp and 28 mismatches in the last 289 bp of the core homology. This situation might have arisen if the *RhsE* and *RhsD* cores originally diverged long enough ago to differ at about 9% of their base pairs (the average divergence of the segments flanking the identical 846 bp). Much more recently, the ancestral *RhsD* and *RhsE* cores must have recombined in a way that caused them to be identical in this 846 bp region.

The comparison of sequences flanking *RhsD* with analogous sequences from the wild *Rhs°* strain ECOR #39 revealed the presence of a 7.3 kb insert in *E. coli* K-12 which we define as *RhsD* (Fig. 1). The core homology begins 425 bp from the left end of the element, which is somewhat greater than the 191 bp found for *RhsA*. Promoter sequences required for *RhsD* core expression presumably lie within this 425 bp segment. Interestingly, the analogous promoter regions for *RhsA*, *RhsB* and *RhsD* share no sequence similarity. Instead, their homology begins precisely at the ATG start codon of the core (5). The conditions required for expression from these putative promoters are being investigated. At the right hand boundary, the *RhsD* element ends 2.6 kb beyond the extended core-ORF. Two additional homologies have been revealed by our preliminary sequencing of this region. These are two short repetitions of portions of the *RhsD* core. One segment is similar to the last 213 bp of the core, while the other is similar to the first 94 bp. These are identified in Fig. 1 as block a and block b, respectively. In block b, the core ATG start codon has been altered to ATT. Much larger partial repetitions involving the distal core ends occur in both *RhsA* and *RhsC* (4). The functional and evolutionary significance of these partial repetitions are obscure.

In place of the 7.3 kb *RhsD* element of *E. coli* K-12, ECOR #39 carries an unrelated 224 bp sequence (Fig. 5). This is somewhat reminiscent of the situation for *RhsA* and *RhsC*, where it appears that these elements have replaced unrelated DNA segments in the *Rhs°* strains. However, the sequences replaced were much smaller, 32 bp and 10 bp respectively (4). Examination of ECOR #39 revealed that it contains four regions that are excellent matches to the REP consensus (15,16). REP sequences are short sequences generally appearing between cistrons and at the end of operons. The positions of the four REP

elements are marked in Fig. 5. Interestingly, a search of the GenBank data base for homologies revealed a significant match of the entire region bracketed by REP 1 and REP 2 with sequences from the 3′ non-coding regions of several *E. coli* genes such as *gyrB* and *pfkA*. As an example, we found that the ECOR #39 sequence matches *gyrB* at 90 of 100 positions.

The precise limits of *RhsE* have not been defined by a similar *Rhs*^{o} comparison. As discussed above, the proximal portion of the *RhsE* core appears deleted in *E. coli* K-12, and it is likely that the deletion included the left hand boundary as well. Hybridization experiments indicate that the right boundary is close to the *Pst* I site. Preliminary sequencing has shown that the *RhsE* element contains another large homology that has been associated with other *Rhs* elements. This homology, known as the H-repeat, constitutes a 1.3 kb sequence that is also present in *RhsB* and *RhsC*. Its position in *RhsE* is depicted as block c in Fig. 1. The H-repeat has been best characterized in *RhsB*, where it has been shown to contain a 1137 bp ORF (unpublished). Preliminary sequencing indicates that the *RhsE* and *RhsB* H-repeats are only 1% divergent in nucleotide sequence. Since the cores of the *RhsD-E* and *RhsA-B-C* subfamilies are 18% divergent, the 1% divergence of the H-repeats of *RhsE* and *RhsB* indicates that these composite elements have been assembled from components with very different evolutionary histories.

ACKNOWLEDGEMENTS

It is with the most sincere feelings of gratitude and affection that CWH dedicates this paper to Paul Berg on his 65th birthday. Paul has the wonderful ability to give differently to each associate, matching his gift to their natural bent. To me, this special gift was a trust in my own instincts in science. We are indebted to Greg Feulner for contributions to the *RhsE* sequence and to Robert Selander, Peter Keumpel and Ken Rudd for bacterial strains and phage. This work was supported by Public Health Service grant GM16329 from the National Institutes of Health.

REFERENCES

1. Folk, W.R. and Berg, P. (1971) *J.Mol.Biol.*, **58**, 595–610.
2. Capage, M. and Hill, C.W. (1979) *J.Mol.Biol.*, **127**, 73–87.
3. Lin, R.-J., Capage, M. and Hill, C.W. (1984) *J.Mol.Biol.*, **177**, 1–18.
4. Feulner, G., Gray, J.A., Kirschman, J.A., Lehner, A.F., Sadosky, A.B., Vlazny, D.A., Zhang, J., Zhao, S. and Hill, C.W. (1990) *J.Bacteriol.*, **172**, 446–456.
5. Sadosky, A.B., Davidson, A., Lin, R.-J. and Hill, C.W. (1989) *J.Bacteriol.*, **171**, 636–642.
6. Ochman, H. and Selander, R.K. (1984) *J.Bacteriol.*, **157**, 690–693.
7. Kohara, Y., Akiyama, K. and Isono, K. (1987) *Cell*, **50**, 495–508.
8. Greener, A. and Hill, C.W. (1980) *J.Bacteriol.*, **144**, 312–321.
9. Norrander, J., Kempe, T. and Messing, J. (1983) *Gene*, **26**, 101–106.
10. Taylor, L.A. and Rose, R.E. (1988) *Nucleic Acids Res.*, **16**, 358.
11. Singer, M., Baker, T.A., Schnitzler, G., Deischel, S.M., Goel, M., Dove, W., Jaacks, K.J., Grossman, A.D., Erickson, J.W. and Gross, C.A. (1989) *Microbiol.Rev.*, **53**, 1–24.
12. Nichols, B.P. and Yanofsky, C. (1979) *Proc.Natl.Acad.Sci.USA*, **76**, 5244–5248.
13. Ochman, H. and Wilson, A.C. (1987) *J.Mol.Evol.*, **26**, 74–86.
14. Engelman, D.M., Steitz, T.A. and Goldman, A. (1986) *Annu.Rev.Biophys.Biophys.Chem.*, **15**, 321–353.
15. Stern, M.J., Ames, G.F.-L., Smith, N.H., Robinson, E.C. and Higgins, C.F. (1984) *Cell*, **37**, 1015–1026.
16. Yang, Y. and Ames, G.F.-L. (1990) In Drlica, K. and Riley, M., (eds.) The Bacterial Chromosome. American Society for Microbiology, Washington D.C., pp. 211–226.