

## Conservation of Organization in the Specificity Polypeptides of Two Families of Type I Restriction Enzymes

P. Kannan, Gill M. Cowan†, Anne S. Daniel, Alexander A. F. Gann‡  
and Noreen E. Murray

Department of Molecular Biology, University of Edinburgh  
King's Buildings, Mayfield Road  
Edinburgh EH9 3JR, U.K.

(Received 7 March 1989)

We have identified the recognition sequence for the *Citrobacter freundii* restriction endonuclease *CfrA*, a member of the A-family of type I R-M enzymes. This bipartite target sequence differs in both its components from those of other type I enzymes. We determined the nucleotide sequence of its specificity gene (*hsdS*) and a comparison of this with its relative *EcoA* identifies two extensive variable regions, an organization analogous to that found in the K-family of type I R-M enzymes. The specificity polypeptides of the A-family, unlike those of K, have an N-terminal conserved region, and this includes a sequence repeated within the central conserved region. A second repeat sequence, identified at the amino acid level, coincides with the only sequence similarity common to all type I S polypeptides. Sequences immediately downstream from the *hsdS* genes of *EcoA*, *CfrA*, *EcoK*, B and D are almost identical, consistent with an allelic chromosomal location.

### 1. Introduction

Restriction and modification (R-M) systems have been subdivided into a number of types each with quite distinct characteristics (for a recent review, see Bickle, 1987). Those classified as type I are the most complex; they are made up of three subunits, have three cofactor requirements and at least three enzymic activities. Methylation occurs within the asymmetric, bipartite target sequence, but DNA is cut at unspecified sequences remote from the unmodified target. The three genes encoding a type I restriction enzyme are designated *hsdR*, *M* and *S*, but only the polypeptides encoded by *hsdM* and *hsdS* are essential for modification. Despite their common characteristics, type I R-M systems show considerable diversity. Currently, three distinct groups, or families, of enzymes have been recognized in *Escherichia coli* and its close relatives.

*EcoK* and *EcoB* were the first type I R-M systems to be identified, and complementation tests depending upon the interchange of subunits between these enzymes established their relatedness (Boyer & Roulland-Dussoix, 1969; Glover & Colson,

1969). This was reinforced by molecular evidence, including cross-hybridization between their genes and cross-reactivity of antibodies raised against *EcoK* with *EcoB*. These same tests showed *EcoA* to be unrelated to *EcoK* (Murray *et al.*, 1982) and led to the subdivision of type I restriction enzymes into families (Fuller-Pace *et al.*, 1985; Suri & Bickle, 1985). The K-family now includes five members, the A-family three, and a third family includes the plasmid-encoded system *EcoR124* and *EcoDXXI*. To date, all well-characterized type I enzymes fit discretely into one of these three families. Other putative type I enzymes remain to be classified (e.g. see Ryu *et al.*, 1988); these may identify new families, but some could bridge the divisions between families.

The specificity polypeptide, S, confers sequence specificity to both the restriction endonuclease and its methyltransferase. A comparison of the nucleotide sequences of the *hsdS* gene of *EcoK* with those of its relatives identified two extensive variable regions (Gough & Murray, 1983), each of which correlates with the recognition of one of the two parts of the target sequence (Fuller-Pace *et al.*, 1984; Nagaraja *et al.*, 1985; Fuller-Pace & Murray, 1986; Gann *et al.*, 1987). Since all type I R-M systems, irrespective of family, recognize bipartite target sequences (*EcoK*, for example, recognizes AAC(N<sub>6</sub>)GTGC) a common mechanism of target

† Present address: Nuffield Department of Clinical Medicine, Oxford OX3 9DU, U.K.

‡ Present address: Department of Biochemistry and Molecular Biology, Harvard University, Cambridge MA, U.S.A.

recognition would predict that the S polypeptides of other families also include two recognition domains.

To test this prediction for a second family of type I enzymes, we now compare the nucleotide sequences of the specificity genes of two members of the A-family whose target sequences are dissimilar in both components. As expected, these differ in two regions. We compare the predicted amino acid sequences of the specificity genes of members of the A-family with those of other families and identify a limited similarity common to the three families. Finally, although the K and A *hsd* genes appear unrelated as judged by the criteria used to separate the families, their location on the *E. coli* chromosome is consistent with allelism (see Daniel *et al.*, 1988). Further evidence for identity of location emerges from the nucleotide sequences of the region downstream from *hsdS*.

## 2. Materials and Methods

### (a) Phage and bacterial strains

The  $\lambda$ *hsd CfrA* phages have been described by Daniel *et al.* (1988). Fragments of their *hsd* DNA were subcloned in M13mp18 and mp19 (Yanisch-Perron *et al.*, 1985) for nucleotide sequencing. Other derivatives of mp18 were used to deduce the recognition sequence for *CfrA* (Fig. 2). These included some clones derived from other sequencing projects (see the legend to Fig. 2), and derivatives made by cloning restriction fragments of either phage  $\lambda$  DNA or pBR322. An M13 sensitive strain of *E. coli* restricting with the specificity of *CfrA* (NM653, see Table 1) was made by introducing *F'kan* into a strain lysogenic for  $\lambda$ *hsd CfrA* (a derivative of phage number 6 of Daniel *et al.*, 1988). All bacterial strains used are listed in Table 1.

### (b) Enzymes and chemicals

DNA polymerase (Klenow fragment), phage T4 DNA ligase, restriction enzymes, deoxynucleoside triphosphates (dNTPs) and dideoxynucleoside triphosphates (ddNTPs) were purchased from Boehringer Mannheim; deoxyadenosine 5'-[ $\alpha$ -(<sup>35</sup>S)]triphosphate (15.2 TBq nmol<sup>-1</sup>) was from Amersham International. Oligonucleotide primers for sequencing were purchased either from New England Biolabs or from Oswel DNA Service, University of Edinburgh.

### (c) Media and microbial methods

Media, general methods (Murray *et al.*, 1977) and tests for estimating restriction and modification have been described (Fuller-Pace *et al.*, 1985).

### (d) Preparation, manipulation and recovery of DNA

The methods were those described by Midgley & Murray (1985).

### (e) Mutagenesis of *CfrA* target sites

Lysates of M13 recombinants  $\lambda$ 32 and T8 were grown first in NM648, a *mulD* strain which stimulates the frequency of point mutations, and then on the *CfrA* restricting strain NM653 to enrich for phage that had lost

Table 1  
Bacterial strains

Strain	Relevant genotype	Reference
NM52	<i>hsd</i> Δ F'	Gough & Murray (1983)
RP526	<i>mutD5</i>	Cesareni (1981)
EH55	<i>asn</i> F' <i>kan</i>	Hansen <i>et al.</i> (1984)
NM648	<i>mulD5</i> F' <i>kan</i>	This paper
NM649	<i>hsd</i> Δ ( $\lambda$ <i>hsd CfrA imm</i> <sup>21</sup> ; <i>limm</i> <sup>434</sup> )	Daniel <i>et al.</i> (1988)
NM653	F' <i>kan</i> derivative of NM649	This paper

a restriction target site. The cycle was repeated and single plaques were isolated on NM653. These clones were grown in NM522 so that the resulting phage were unmodified. Mutants plating with an efficiency of 1 on NM522 *versus* NM653 were sequenced. Some were deletions, but two, one of  $\lambda$ 32 and one of T8, had a single base-pair difference from the wild-type sequence.

### (f) DNA sequencing

The region containing the entire *hsdS* gene of *CfrA* was sequenced on both strands using overlapping restriction fragments (see Fig. 1). Synthetic oligonucleotides were used as primers when needed. Single-stranded template DNA was prepared and sequenced by the dideoxy chain termination method using [ $\alpha$ -<sup>35</sup>S]thio-dATP (Sanger *et al.*, 1980) and analysed on buffer gradient gels (Biggin *et al.*, 1983). Compressed sequences were resolved by the use of dITP in place of dGTP. DNA sequences were assembled using the programs of Staden (1982) and analysed using UWGCG software (Devereux *et al.*, 1984).

## 3. Results

### (a) Recognition sequence of *CfrA*

A type I restriction system identified in *Citrobacter freundii* (*CfrA*) has been shown to be a member of the A-family (Daniel *et al.*, 1988). We have now determined the recognition sequence of *CfrA* using an extension of a biological approach previously described by Gann *et al.* (1987) and Cowan *et al.* (1989). Phage containing an unmodified target for a restriction system plate with a reduced efficiency on a strain encoding that system (Arber & Kühnlein, 1967): in our experience a single target in an M13 phage generally confers an efficiency of plating (e.o.p.†) of 10<sup>-1</sup> and two targets an e.o.p. of 10<sup>-2</sup> (Gann *et al.*, 1987). Identifying targets in this way, rather than by *in vitro* methylation (e.g. see Suri *et al.*, 1984), obviates the need for enzyme purification.

Phage M13 plates with equal efficiency on a *CfrA* restricting strain (NM653) and on NM522, which carries no *hsd* system, indicating that M13 itself lacks a target for the *CfrA* enzyme. We screened M13 recombinants which contained fragments of known sequence until we found two that showed a reduced e.o.p. (~10<sup>-1</sup>) on the *CfrA* restricting

† Abbreviations used: e.o.p., efficiency of plating; bp, base-pair(s); kb, 10<sup>3</sup> base-pairs.

Figure 1.  
sequenced.  
arrowheads

strain. an  
each. One  
516 bp P  
(T8) con  
pBR322.  
introduc  
nitio seq  
were selec  
on the C  
sequencin  
change (s  
type I en  
trinucleot  
pentanuel  
seven or  
the entire  
within 1  
mutation  
recognitic  
Only on  
GCA(N<sub>8</sub>)  
sequence  
has a del  
strands;  
methylat  
Some t  
sequence  
or altern  
that this  
this typ  
showed 1  
other M  
found in  
sons of t  
alternat  
recogniti  
spacer  
(N)GCA  
by the C

(b) Nu  
We h  
gene o  
sequenc  
bp 234  
tide of  
upstrea  
sequenc



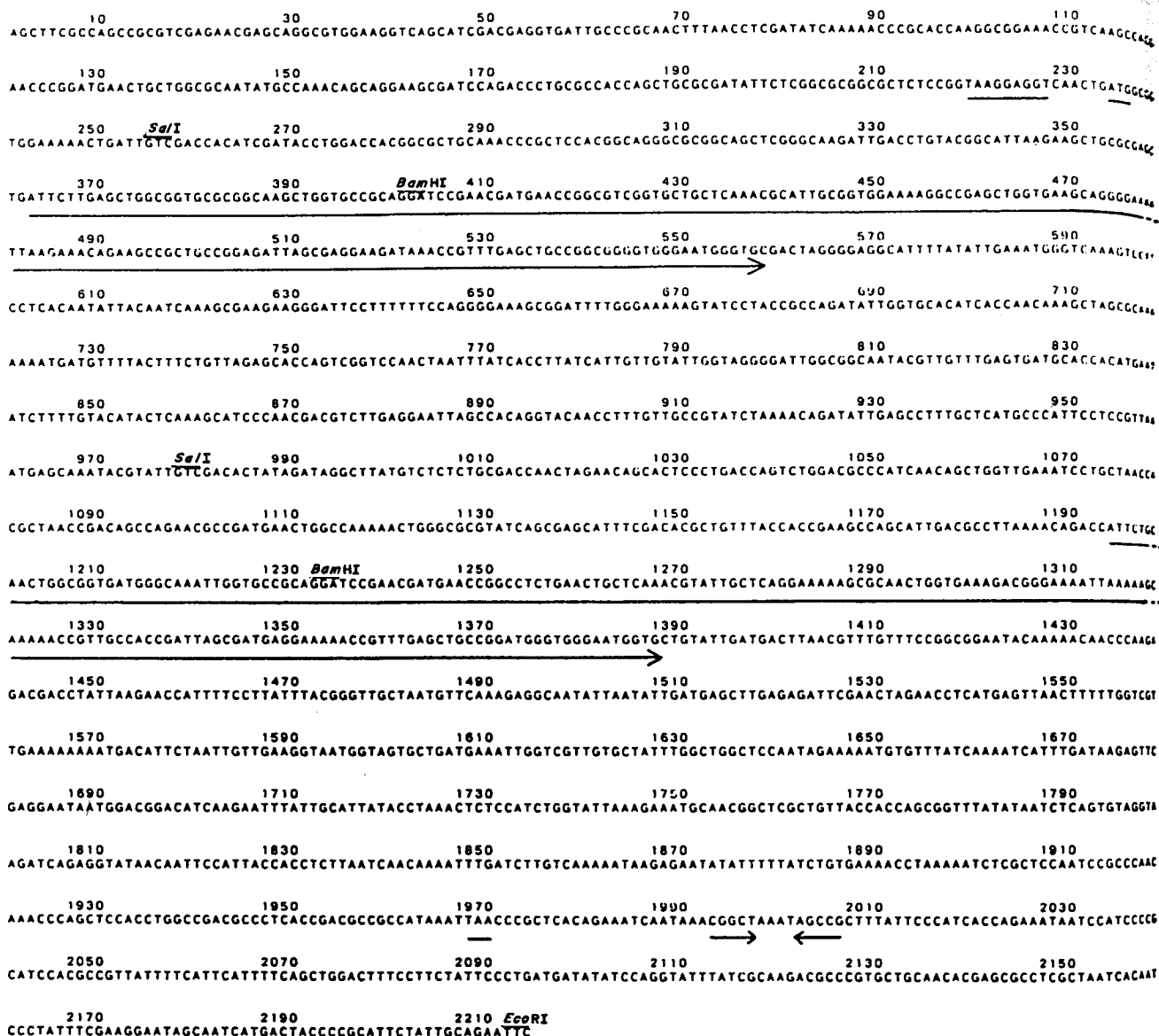


Figure 3. The nucleotide sequence of the *hsdS* gene of *CfrA*. The Shine-Dalgarno sequence, initiation and termination codons are underlined. The direct repeat of 195 bp and the inverted repeat constituting the presumed rho-independent terminator are marked by arrows.

### (c) Organization of the A-family S polypeptides

The predicted polypeptide sequences of the *CfrA* and *EcoA* *hsdS* genes were aligned (Fig. 5). The two variable regions, referred to as amino and carboxyl, are approximately 150 and 180 amino acid residues in length, respectively, as are those of the K-family (Gough & Murray, 1983). The three regions seen to be almost identical in both S polypeptides constitute the N-terminal 107 residues, the C-terminal 16 residues and a region of 131 residues in the centre separating the variable regions (see Fig. 5). Comparisons of the amino acid, rather than nucleotide, sequences revealed an additional feature: a short sequence repeated within each polypeptide of the A-family (data not shown). The level of similarity

seen when the repeats of one polypeptide are compared is also maintained when related polypeptides are compared. Consequently, in the dot matrix of *CfrA* and *EcoA* polypeptides (Fig. 6), these repeat sequences spanning 24 residues are identified by the weaker lines above and below the diagonal, but because of their conservation throughout the family they also contribute to the main diagonal. Since these repeat sequences are at the carboxy end of each variable region, their effect on the diagonal is to extend the apparent lengths of the central and distal conserved regions. While the similarity seen when the repeats are compared is obvious, it does not approach the near identity characteristic of the conserved regions (see Fig. 5). A schematic diagram of an S polypeptide is shown in Figure 7.

Figure  
represent  
indicate

(d) C

A re  
sequen  
nevert  
S poly  
Argos  
sequen  
These  
mainly  
polype  
repeat  
K-fam  
exceed  
polype

CAAGCCACC  
 TGATGGCGG  
 TGCGGAGC  
 AGGGGAAGA  
 AAAGTCTT  
 TAGCCAAA  
 CACATGAA  
 CTCGGTTA  
 TGCTAACCA  
 ACCATTCTGC  
 ATTAAAAAGC  
 CAACCCAAGA  
 TTTTGGTCGT  
 ATAAGAGTTC  
 AGTGTAGGTA  
 TCCGCCCAAC  
 TCCATCCCCG  
 TAATCACAAT

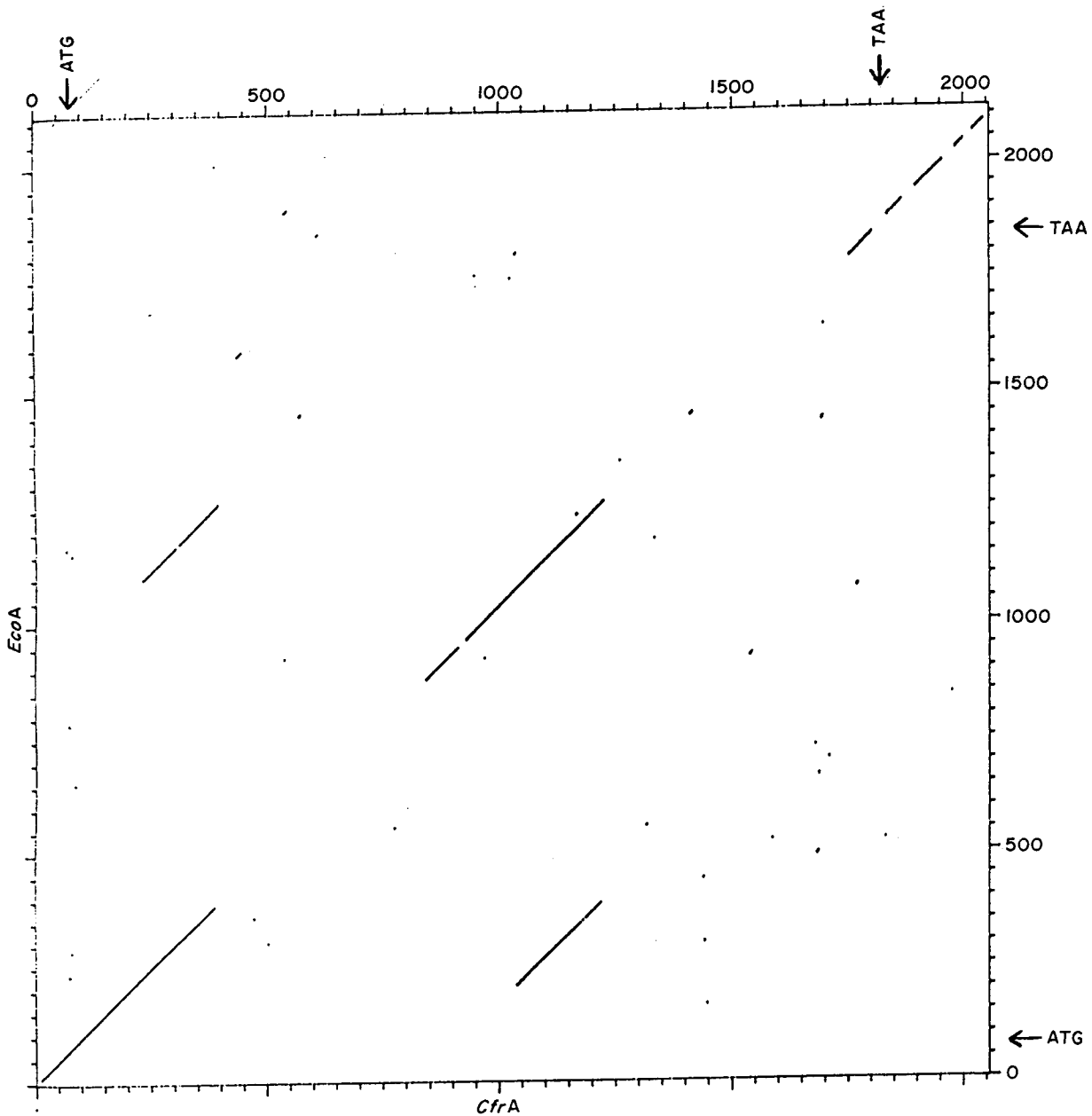


Figure 4. Dot matrix comparisons of the nucleotide sequences of the *hsdS* genes of *CfrA* and *EcoA*. Each dot represents the conservation of at least 15 bases within a region of 21 bases. The 2 lines parallel to the main diagonal indicate the repeated sequence in both *CfrA* and *EcoA*.

(d) Comparison of S polypeptides between families

A repeated sequence detected when amino acid sequences but not nucleotides are compared, but nevertheless conserved throughout the A-family of S polypeptides, is reminiscent of the finding of Argos (1985), who demonstrated a repeated sequence within each S polypeptide of the K-family. These repeats of about 60 residues are, however, mainly confined to the conserved regions of the polypeptides (Fig. 7). Inevitably, therefore, each repeat is strongly conserved throughout the K-family; the similarity between members greatly exceeding that found between the repeats of one polypeptide. We shall refer to the respective

repeated sequences as the central repeat and the carboxyl repeat. Each of these sequences in the K-family is made up of two components A and B; these, but not the intervening sequence, are repeated (see Fig. 7).

The central and carboxy repeats within the S polypeptides of the A-family are equivalent to the A components of the Argos repeats in the K-family. The nucleotide sequence of the *hsd* genes encoding the R124 system has been determined (Price *et al.*, 1989), and a repeat is also present in the S polypeptide of this third family of enzymes. An alignment of the repeat sequences from the central and carboxyl regions of *EcoK*, A and R124, reveals that all six sequences are similar (Fig. 8). Within these repeats

	1				50
<i>CfrA</i>	MaVEKLIVDH	iDTWTtALQT	RSTAGRGSSG	KIDLyGIKkL	RELILELAVR
<i>EcoA</i>	MsVEKLIVDH	mETWTsALQT	RSTAGRGSSG	KIDLyGIKkL	RELILELAVR
	51				100
<i>CfrA</i>	GKLVPQDPND	EPASvLLKRI	AvEKAELVKQ	GKIKKQKPLP	EISEEDKPF
<i>EcoA</i>	GKLVPQDPND	EPASeLLKRI	AaEKAELVKQ	GKIKKQKPLP	EISEEEKPF
	101				150
<i>CfrA</i>	LPaGWEWvrL	geafyIemgq	spSsdyyngS	eegiPffqgK	adfgkkYpta
<i>EcoA</i>	LPdGWEWttL	triaeInpki	dvSddeqeiS	fipmPlistK	fdgsheFeik
	151				200
<i>CfrA</i>	Ry.....	.wctsptkLA	qkndvllsvR	ApVgptnlsp	yhccigr gla
<i>EcoA</i>	Kwkdvkkgyt	hfangdiaIA	kitpcfensK	AaIfsglkng	igvgttelhv
	201				250
<i>CfrA</i>	Airclsdaph	eYLLyilKas	...qrrleeL	atgttfvaVs	KtdiEpllmP
<i>EcoA</i>	Arpfsdiinr	kYLLlnfKsp	nflksgesqM	tgSagqkrVp	RfffEnnpip
	251				300
<i>CfrA</i>	iPPLnEQiRI	VdtidrLMSL	CDQLEQhSLT	SLDAHQQQlVE	iLLtTLTDSQ
<i>EcoA</i>	fPPLqEQeRI	IirftqLMSL	CDQLEQqSLT	SLDAHQQQlVE	tLLgTLTDSQ
	301				350
<i>CfrA</i>	NaDELAKnWA	RISEHFDTLF	TTEASIDALK	QTILQLAVMG	KLVPQDPNDE
<i>EcoA</i>	NvEELAeNWA	RISEHFDTLF	TTEASVDALK	QTILQLAVMG	KLVPQDPNDE
	351				400
<i>CfrA</i>	PASELLKRIA	QEKAQLVKDG	KIKKQKPLPP	ISDEEKPFEL	PDGWEWCCId
<i>EcoA</i>	PASELLKRIA	QEKAQLVKEG	KIKKQKPLPP	ISDEEKPFEL	PEGWEWCrlg
	401				450
<i>CfrA</i>	dLtfvsgGiq	kqpkrrpikn	hfpYLRvANV	qrGniNidEl	erfeLE.phE
<i>EcoA</i>	sIynflnGya	fksewftsvg	.lrlLRnANI	ahGvtNwkDv	vhipnDmisD
	451				500
<i>CfrA</i>	ltfwsLkkND	IlIvegnsga	deigRcAIwl	apieKcVyyqn	hliRvRgimd
<i>EcoA</i>	fenyiLseND	IvIslDrpii	ntglKyAIis	ksdlpCl11q	rvaKfKnyan
	501				550
<i>CfrA</i>	g.hqeFIaLy	LnSpSgIkem	qrlavttsGl	ynLSvgkIrg	itiPLpPlnq
<i>EcoA</i>	tvsnSFLtIw	LqSyffInsi	d..pgrsnGv	phIStkqLem	tlfPLlPqse
	551				592
<i>CfrA</i>	QnlILSKirE	yIfiCenLKl	sLqsAqQTQL	HLADALTDAA	IN
<i>EcoA</i>	QdrIISKmdE	lIqtCnkLKy	iIktAkQTQL	HLADALTDAA	IN

Figure 5. An alignment of the predicted amino acid sequence of *CfrA* and *EcoA*. Gaps were allowed to optimize the alignment. Upper case letters represent the conserved residues. The conservative substitutions allowed were I/L, I/V, L/M, E/D, F/Y, and K/R as recommended by Collins & Coulson (1987). The repeat sequence that shows amino acid conservation when compared to the A component of the Argos repeat (see Figs 7 and 8) is underlined.

there is the same degree of similarity between the families as that detected within a single polypeptide, assessed by the fact that identities are just as likely to occur across the families as between repeats within a given polypeptide (see Fig. 8).

#### (e) Comparison of downstream DNA sequences

The sequences immediately downstream from the termination codons of *CfrA* and *EcoA* (Fig. 9) are similar. Following the potential transcriptional

Figure  
indicates  
diagonal  
to the A

termina  
nation  
also str  
sequen  
for *Eco*  
K-fami  
Murray  
*hsd* gen  
Daniel  
stream  
occupy  
chromo  
A-fami  
Obtaine  
similar  
downst

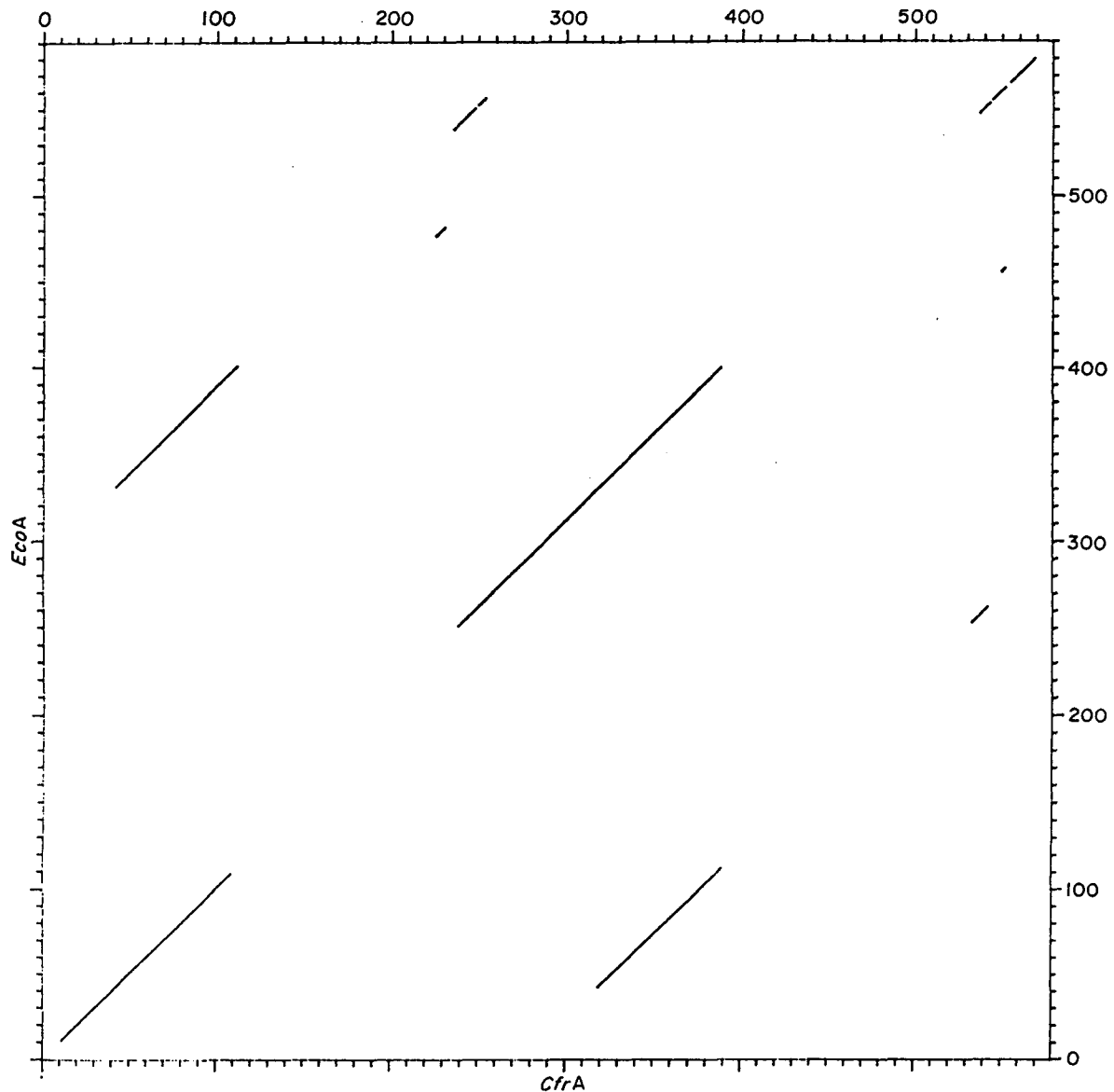


Figure 6. Dot matrix comparisons between the amino acid sequence of the S polypeptides of *CfrA* and *EcoA*. Each dot indicates the conservation of at least 14 amino acid residues within a region of 20 residues. The lines parallel to the main diagonal indicate the repeats in both *CfrA* and *EcoA*. The shorter repeat shows amino acid conservation when compared to the A component of the Argos repeat (see Figs 7 and 8).

terminators, ~100 bp downstream from the termination codons, the sequences of *CfrA* and *EcoA* are also strongly similar to those of *EcoK* (Fig. 9). The sequence downstream from the *S* genes is conserved for *EcoB* and *EcoD*, the two other members of the K-family for which data are available (Gough & Murray, 1983). Previous evidence indicates that the *hsd* genes for all these systems behave as alleles (see Daniel *et al.*, 1988). The sequence homology downstream from their *S* genes indicates that they may occupy identical locations in their respective chromosomes. *EcoE*, another member of the A-family for which sequence information has been obtained (Cowan, 1988), is anomalous and shows no similarity to the other systems in the 250 bp of downstream sequence available.

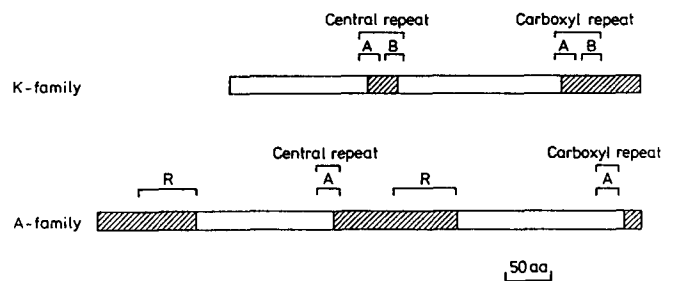


Figure 7. Schematic diagrams of specificity polypeptides. The regions conserved in each family are hatched, the variable regions are open. R represents the strong repeat seen only in the A-family. The central and carboxyl repeats in the K-family, the Argos repeats, are made up of 2 components; only 1 of these is present in the A-family. aa, amino acid residues.

<i>EcoA</i>	Central	I P F P P L Q E Q E	R I I I R F T Q L M	S L C D
<i>EcoA</i>	Carboxyl	F P L L P Q S E Q D	R I I S K M D E L I	Q T C N
<i>EcoK</i>	Central	I P I P P L A E Q K	I I A E K L D T L L	A Q V D
<i>EcoK</i>	Carboxyl	V L L P P V K E Q A	E I V R R V E Q L F	A Y A D
<i>EcoRI24</i>	Central	N P E K S L A I Q S	E I V R I L D K F T	A L T A
<i>EcoRI24</i>	Carboxyl	I P V P N I N E Q Q	R I V E I L D K F D	T L T N
Consensus		I P * p p l * E Q -	r i v * * l d * l -	a l - d

**Figure 8.** An alignment of the amino acid sequences from the repeats within the specificity polypeptides of the 3 families of enzymes. Upper case letters in the consensus sequence indicate positions where at least 5 residues are identical, lower case letters where at least 3 residues are identical. Asterisks indicate other positions where identical amino acids are present in different families. Any position in the consensus sequence indicated by either a letter or an asterisk is one where there is interfamily conservation. Of the 3 positions where this is not the case (10, 20 and 23), only one (23) shows conservations between the repeats within a given S polypeptide.

#### 4. Discussion

The specificity polypeptides of *EcoA* and *CfrA*, members of the A-family of type I R-M enzymes, differ in two extensive regions referred to as the variable regions; a situation analogous to that found in the K-family. Type I R-M enzymes recognize target sequences comprising two defined regions separated by a non-specific spacer sequence. In the K-family the two variable regions behave as independent recognition domains (Gann *et al.*, 1987).

The N-terminal domain of 150 amino acid residues is proven to specify the trinucleotide component of the target sequence (Cowan *et al.*, 1989), and extensive circumstantial evidence confines the recognition domain for the tetra-, or pentanucleotide, target sequence to the carboxy-variable region (Fuller-Pace & Murray, 1986). The following observations support the same correlations for the specificity polypeptides of the A-family.

The variable domains, like those of the K-family, are around 150 and 180 residues long, and both are very different when both components of the specified target sequences are different. In contrast, *EcoA* and its relative *EcoE* have 5'GAG as the trimeric component of their target sequences and they have almost identical amino variable domains (Cowan *et al.*, 1989). Similarly, *EcoK* and *StySP* of the K-family both recognize 5'AAC and have nearly identical amino variable domains (Fuller-Pace & Murray, 1986). Particularly relevant is a 44% identity between the amino variable regions of either *EcoA* and *StySB* or *EcoE* and *StySB*. This is the only extensive similarity found between specificity polypeptides from different families and correlates with these being the only known unrelated S polypeptides to share a common component in their target sequences: the trinucleotide 5'GAG (Cowan *et al.*, 1989). We therefore propose that the two variable regions of the S polypeptides from the A-family correlate with two DNA recognition domains. As for the K-family, the recognition domains appear to be extensive though we have no direct evidence to implicate residues throughout the length of a vari-

	1				50
<i>CfrA</i>	.....	<u>Taa</u> CcCGCTC	ACAgAAatcA	atAAaCGGCT	AAATaGCCGC
<i>EcoA</i>	.....	<u>taat</u> T TcTCgCcCTC	AtAaAACCaA	TaAAgCGGCT	AAATgGCCGC
<i>EcoK</i>		<u>tgaacatta</u> T TtTCtgGCgC	ACctttCCgg	TgcgcttttT	AttatttCaC
	51				100
<i>CfrA</i>		TTTATTCcCA TCACCAgAA.	..ATAATcCA	tcCccgCAtc	cAgccgTTA
<i>EcoA</i>		TTTATTCACA TCACCAAAAA	tTATATTTCA	cgC...tAAt	tttcatCTTA
<i>EcoK</i>		gccAaTCaTA aCcCacAtAA	aTATATTTaA	atCattCcAg	aAattgCccA
	101				150
<i>CfrA</i>		TTTTcATTcA TTTTcAGCTG	GACTTTCcTt	cTATTccCTG	ATGATATATC
<i>EcoA</i>		TTTTcATTaA TTTTtAGCTG	GACTTTCcCT	gTATTTACTG	ATGATATATC
<i>EcoK</i>		TTTTatTctA TTTTtAGCTG	GACTTTCcCc	aTATTTACTG	ATGATATATA
	151				200
<i>CfrA</i>		CAGGTATTTA tCGCaagaCG	cccGTGctgC	AACACgagCG	CcTCGCTAAT
<i>EcoA</i>		CAGGTATTTA GCGCGGTGCG	GATGTGCGCC	AACACACCCG	CAcCGCTAAT
<i>EcoK</i>		CAGGTATTTA GCGCGGTGCG	GATGTGCGCC	AACACACCCG	CATCGCTAAT
	201				250
<i>CfrA</i>		CACAATCcCT ATTTcGaaagG	AATAGCAaTc	ATGACTacCc	CGCATTCTAT
<i>EcoA</i>		CACAATCgCT gTTatcGGAG	AATAGCAGTT	ATGACTGACA	CGCATTCTAT
<i>EcoK</i>		CACAATCaCT ATTTcCtGGAG	AATAGCAGTT	ATGACTGACA	CGCATTCTAT

**Figure 9.** An alignment of the downstream nucleotide sequence of *CfrA*, *EcoA* and *EcoK*. Gaps were allowed to optimize the alignment. Upper case letters represent conserved residues, termination codons are underlined.

able region in the definition of specificity. Nevertheless, the similarity in size of the variable regions in the two families of enzymes adds support to the concept that extensive regions of polypeptide are required for sequence specificity in these complex systems. Long variable regions that correlate with DNA recognition domains have also been identified in the methyltransferases of *Bacillus subtilis* phages (Balganesh *et al.*, 1987; Behrens *et al.*, 1987).

The only general sequence similarity between the A- and K-family S polypeptides coincides with the region previously identified as being repeated within each S polypeptide of the K-family (Argos, 1985). We have also identified an equivalent repeat in the S polypeptides of *EcoR124*, a member of the third family. This sequence is conserved between the families at the same level as it is between repeats within a given S polypeptide. If, therefore, this repeat has functional significance, it could identify regions involved in an activity common to all type I enzymes.

The finding that the repeat sequences are common to all three families of polypeptides may add weight to their relevance, but any interpretation of this is complicated by the fact that in the K-family the repeats are mainly within the conserved regions while in the A-family they are mainly within the variable regions (Fig. 7). Consequently the central repeats from different K-family members, and similarly the carboxyl repeats, are far more alike than the central and carboxyl repeats within a single polypeptide. This, of course, is not the case in the A-family. The high level of conservation seen in either the central or carboxyl repeats of the K-family may not be essential for the functional integrity of the polypeptide. Indeed, one member of the family, *EcoD*, shows greater variability in its central conserved region than its relatives (Gough & Murray, 1983), indicating that complete conservation is not required for normal function. Random mutagenesis experiments should indicate the degree of conservation necessary for enzyme activity. Near identity in the conserved region of the *hsdS* genes of *EcoK* and *EcoB* could reflect a recombination event relatively recently on an evolutionary time-scale. Recombination between the central conserved regions of related *hsdS* genes has been observed, and while this results in new sequence specificities (Bullas *et al.*, 1976; Gann *et al.*, 1987) it should reduce diversification in conserved regions.

It has been suggested that the repeats within the S polypeptides of the K-family are the only visible remnants of a gene duplication (Argos, 1985; Gann *et al.*, 1987). An ancestral gene encoding a polypeptide with a single recognition domain could have duplicated, resulting in the present organization. If the A- and K-families are descended from a common ancestral system, it is simplest to think of an evolutionary pathway starting with a common ancestral gene, duplication generating the two recognition domains and familial divergence followed by generation of new specificities within the families.

The A-family S polypeptides, however, contain a

second repeat absent from those of both the K- (see Fig. 6) and R124-families. This strong repeat, detected even at the nucleotide level, is found in the amino and central conserved regions and could be an additional remnant of gene duplication. Both of these strong repeats are located in regions of the A-family specificity polypeptides that have no equivalents in those of the K-family (see Fig. 7). The absence of both of these regions from the K-polypeptides could be explained by a deletion, or insertion, prior to gene duplication, and would require separate duplication events during the evolution of the A- and K-families of specificity genes.

The genes encoding the K- and A-systems appear to be allelic. This was first suggested by genetic linkage to *serB* (Arber & Wauters-Willems, 1970) and is supported by molecular studies (Daniel *et al.*, 1988). The similarity of nucleotide sequence immediately downstream from the S genes of *EcoA*, *CfrA* and members of the K-family supports this claim and encourages a belief in their sharing a common ancestor.

We thank Julia Kelleher for constructive criticism of the manuscript, A. F. W. Coulson and J. F. Collins for advice on computing programs, Professors T. A. Bickle and S. W. Glover for communication of sequence data prior to publication, Fiona Govan for preparation of the manuscript and Annie Wilson for Figures. The work was supported by the Medical Research Council and by Science and Engineering Research Council studentships to G.M.C. and A.A.F.G.

## References

- Arber, W. & Kühnlein, U. (1967). *Pathol. Microbiol.* **30**, 946-952.
- Arber, W. & Wauters-Willems, D. (1970). *Mol. Gen. Genet.* **180**, 203-217.
- Argos, P. (1985). *EMBO J.* **4**, 1351-1355.
- Balganesh, T. S., Reiners, L., Lauster, R., Noyer-Weidner, M., Wilke, K. & Trautner, T. A. (1987). *EMBO J.* **6**, 3543-3549.
- Behrens, B., Noyer-Weidner, M., Pawlek, B., Lauster, R., Balganesh, T. S. & Trautner, T. A. (1987). *EMBO J.* **6**, 1137-1142.
- Bickle, T. A. (1987). In *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology* (Neidhardt, F. C., Ingraham, J. L., Low, K. B., Magasanik, B., Schaechter, M. & Umberger, H. E., eds), pp. 692-696. American Society for Microbiology, Washington, D.C.
- Biggin, M. D., Gibson, T. J. & Hong, C. F. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 3963-3965.
- Boyer, H. W. & Roulland-Dussoix, D. (1969). *J. Mol. Biol.* **41**, 459-472.
- Bullas, L. R., Colson, C. & van Pel, A. (1976). *J. Gen. Microbiol.* **95**, 166-172.
- Cesareni, G. (1981). *Mol. Gen. Genet.* **184**, 40-45.
- Collins, J. F. & Coulson, A. F. W. (1987). In *Nucleic Acids and Protein Sequence Analysis: A Practical Approach* (Bishop, M. J. & Rawlings, C. F., eds), pp. 323-358. IRL Press, Oxford.
- Cowan, G. M. (1988). Ph.D. thesis, University of Edinburgh.

- Cowan, G. M., Gann, A. A. F. & Murray, N. E. (1989). *Cell*, **56**, 103-109.
- Daniel, A. S., Fuller-Pace, F. V., Legge, D. M. & Murray, N. E. (1988). *J. Bacteriol.* **170**, 1775-1782.
- Devereux, J., Haerberli, P. & Smithies, O. (1984). *Nucl. Acids Res.* **12**, 387-395.
- Fuller-Pace, F. V. & Murray, N. E. (1986). *Proc. Nat. Acad. Sci., U.S.A.* **83**, 9368-9372.
- Fuller-Pace, F. V., Bullas, L. R., Delius, H. & Murray, N. E. (1984). *Proc. Nat. Acad. Sci., U.S.A.* **81**, 6095-6099.
- Fuller-Pace, F. V., Cowan, G. M. & Murray, N. E. (1985). *J. Mol. Biol.* **185**, 65-75.
- Gann, A. A. F., Campbell, A. J. B., Collins, J. F., Coulson, A. F. W. & Murray, N. E. (1987). *Mol. Microbiol.* **1**, 13-22.
- Glover, S. W. & Colson, C. (1969). *Genet. Res. Camb.* **13**, 227-240.
- Gough, J. A. & Murray, N. E. (1983). *J. Mol. Biol.* **166**, 1-19.
- Hansen, E. B., Atlung, T., Hansen, F. G., Skovgaard, O. & Van Meyenberg, K. (1984). *Mol. Gen. Genet.* **196**, 387-396.
- Loenen, W. A. M., Daniel, A. S., Braymer, H. D. & Murray, N. E. (1987). *J. Mol. Biol.* **198**, 159-170.
- Midgley, C. A. & Murray, N. E. (1985). *EMBO J.* **4**, 2695-2703.
- Murray, N. E., Brammar, W. J. & Murray, K. (1977). *Mol. Gen. Genet.* **150**, 53-61.
- Murray, N. E., Gough, J. A., Suri, B. & Bickle, T. A. (1982). *EMBO J.* **1**, 535-539.
- Nagaraja, V., Shepherd, J. C. W. & Bickle, T. A. (1985). *Nature (London)*, **316**, 371-372.
- Price, C., Lingner, J. & Bickle, T. A. (1989). *J. Mol. Biol.* **205**, 115-120.
- Ryu, J.-I., Rajadas, P. T. & Bullas, L. R. (1988). *J. Bacteriol.* **170**, 5785-5788.
- Sanger, F., Coulson, A. R., Barell, G. B., Smith, A. J. & Roe, B. A. (1980). *J. Mol. Biol.* **43**, 161-178.
- Shine, J. & Dalgarno, L. (1974). *Proc. Nat. Acad. Sci., U.S.A.* **71**, 1342-1346.
- Staden, R. (1982). *Nucl. Acids Res.* **10**, 4731-4751.
- Suri, B. & Bickle, T. A. (1985). *J. Mol. Biol.* **186**, 77-85.
- Suri, B., Shepherd, J. C. W. & Bickle, T. A. (1984). *EMBO J.* **3**, 575-579.
- Sutcliffe, J. G. (1979). *Cold Spring Harbor Symp. Quant. Biol.* **43**, 77-90.
- van Wezenbeck, P. M. G. F., Hulsebos, T. J. M. & Schoenmakers, J. G. G. (1980). *Gene*, **11**, 129-148.
- Yanisch-Perron, C., Vieira, J. & Messing, J. (1985). *Gene*, **33**, 103-119.

Edited by N. L. Sternberg