

Two DNA recognition domains of the specificity polypeptides of a family of type I restriction enzymes

(protein-DNA interaction/protein evolution)

FRANCES V. FULLER-PACE* AND NOREEN E. MURRAY†

Department of Molecular Biology, University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh EH9 3JR, United Kingdom

Communicated by F. Sanger, August 25, 1986

ABSTRACT The *hsd* genes of *Salmonella typhimurium* and *Salmonella potsdam* encode related type I restriction and modification systems designated SB and SP, respectively; the polypeptide encoded by the *hsdS* gene dictates the DNA sequence recognized. The *hsdS* genes of the SB and SP systems have a conserved sequence of around 100 base pairs flanked by two nonhomologous (variable) regions of around 500 base pairs. Recombination between the *hsdS* genes of SB and SP generated a system (SQ) with a different recognition specificity. We have localized the position of the crossover in the central conserved region by analysis of nucleotide sequences. Concomitant with the generation of a new combination of flanking variable regions is the recombination of minor differences in the central conserved region. A polypeptide domain encoded on the 5' side of the crossover dictates recognition of the trinucleotide component of the target sequence, and a second domain, encoded on the 3' side of the crossover, similarly governs recognition of the tetra- or penta-nucleotide component. Our analysis implicates at least parts of the variable regions in the determination of the specificity of interaction between protein and DNA. Furthermore, the trinucleotide components of the recognition sequences of *S. typhimurium* and *Escherichia coli* K-12 are identical, and the 5' segments of their *hsdS* genes are strikingly homologous rather than variable.

Some strains of *Escherichia coli* and *Salmonella* spp have type I restriction systems that are related to that of *E. coli* K-12 (K system). The concept of a family of enzymes, members of which carry out essentially the same reactions following the recognition of different DNA sequences, poses questions concerning both the specific interaction of the enzyme with its target sequence and the evolution of different specificities of interaction. Type I restriction endonucleases are encoded by three chromosomal genes: *hsdR*, *hsdM*, and *hsdS*. Genetic experiments show that the corresponding polypeptides (R, M, and S) can be interchanged between related systems and identify the S subunit as the determinant of the specificity of recognition (1-3). The various S polypeptides interact with related polypeptides to produce modification enzymes and with M and R subunits in the formation of restriction endonucleases. Related S subunits must retain a basic similarity for their interactions with other M and R polypeptides but have diverged to enable recognition of different nucleotide sequences.

The *hsdS* genes of related *E. coli* systems (K, B, and D) have only localized DNA sequence homology (4). The genes vary in length from 1335 to 1425 base pairs (bp), and the regions of homology are \approx 100 bp in the middle and 250 bp at the 3' end. Although these two regions are highly conserved, the remainder of each *hsdS* gene shares little or no homology with either of the other two.

The K family includes the SB system found in *Salmonella typhimurium* and the SP system of *Salmonella potsdam*. When the SP *hsd* genes were transferred by phage P1 transduction to an SB recipient, one transductant was found to have a new specificity, designated SQ (5). Heteroduplex analyses of the *hsd* genes of the SB, SP, and SQ systems (6) showed that their organization parallels that of *E. coli* K-12 and B (4), with two nonhomologous (variable) regions of about 500 bp flanking a conserved core of around 100 bp. Furthermore, the *hsdS* gene of SQ contains one variable region from SP and the other from SB, confirming that the SQ specificity was generated by recombination between the two parental specificity genes (6).

The DNA sequences recognized by type I restriction enzymes comprise specific trinucleotide and tetranucleotide (or pentanucleotide) segments separated by a spacer of defined length but of nonspecific sequence (7-9). One interpretation of the heteroduplex analyses correlates the two variable regions of the *hsdS* gene, and hence of the S polypeptide, with the two specific segments of the DNA target sequence (4). Alternatively, reassortment of minor differences within the conserved regions might suffice to generate a new specificity (4). Models in which recombination results in the reassortment of two domains of the S polypeptide predict the recognition of a recombinant target sequence. The recognition sequence of SQ does indeed comprise the trinucleotide element of SP and the pentanucleotide component of SB (10).

In this paper we define the site of the crossover that segregated the parts of the gene encoding the polypeptide domains recognizing the trinucleotide and tetranucleotide components of the target sequences and, thereby, generated the SQ specificity. We report the nucleotide sequence of the specificity gene of the SP system, of particular interest because one of the two target sequences of SP is identical to that of K (9). The K, SP, and SQ specificity systems all recognize 5' AAC; hence, an amino acid sequence common to their S polypeptides should identify the relevant recognition domain.

MATERIALS AND METHODS

Strains, Vectors, and Media. The λ *hsd* phages (Fig. 1) were propagated in the *hsd*-deletion strain NM477 (4). The respective *hsd* regions were subcloned in pBR322, propagated in the *recA hsdS* host HB101 (1), and used as probes and as sources of small fragments of DNA for cloning in M13. Vectors Mp10, Mp11, Mp18, and Mp19 (11) were used, and recombinants were grown in the *hsd*-deletion host NM522 (4).

Enzymes and Chemicals. DNA polymerase (Klenow fragment) was purchased from Boehringer Mannheim; DNA

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: bp, base pair(s).

*Present address: Research Institute of Scripps Clinic, La Jolla, CA 92037.

†To whom reprint requests should be addressed.

polymerase I was from Northumbria Biologicals Limited Enzymes (Cramlington, United Kingdom); T4 DNA ligase, restriction enzymes, phage M13 17-mer primer, and hybridization probe primer were from the New England Biolabs; deoxynucleoside triphosphates (dNTPs) and dideoxynucleoside triphosphates (ddNTPs) were from P-L Biochemicals; and deoxycytidine 5'-[α -³²P]triphosphate (110 TBq/mmol) and deoxyadenosine 5'-[α -(³⁵S)thio]triphosphate (15.2 TBq/mmol) were from Amersham.

DNA Preparation, Analysis, and Cloning. Fragments of DNA separated in 0.7% agarose gels in TBE buffer, pH 8.2 (12), were isolated by electroelution into a well lined with dialysis tubing. Small DNA fragments generated by digestion with *Alu* I or *Sau*3a or by sonication (13) were cloned in phage M13 vectors for sequencing. DNA from plaques was transferred to nitrocellulose for hybridization (14); single-stranded phage M13 DNA was added from cleared lysates (15).

DNA Sequencing. Template DNA was sequenced by the dideoxy chain-termination method (16) with deoxyadenosine 5'-[α -(³⁵S)thio]triphosphate and was analyzed on buffer-gradient gels (12). The sequences of DNA fragments were compiled with computer programs (17, 18). Sections of sequence, initially obtained on only one strand, were selected from libraries of recombinants by using strand-specific probes (15). A few areas were completed by using synthetic oligonucleotide primers.

RESULTS AND DISCUSSION

The Crossover That Generated the SQ Specificity. The *hsdS* gene of the SQ system derives one variable region from SP and the other from the SB gene (6). Hybridization of DNA from λ *hsd* SB, SP, and SQ phages with a probe covering the central conserved region of the *hsdS* gene of *E. coli* K-12 showed that, in these *hsd* phages, the central conserved region is to the right of the *Eco*RI and *Hind*III targets identified by an asterisk in Fig. 1 (6). Hybridization to an *hsdM*-specific probe from *E. coli* K-12 (data not shown) located the *hsdM* gene to the left of these same targets. This establishes the orientation of *hsdS* with respect to *hsdM*, and,

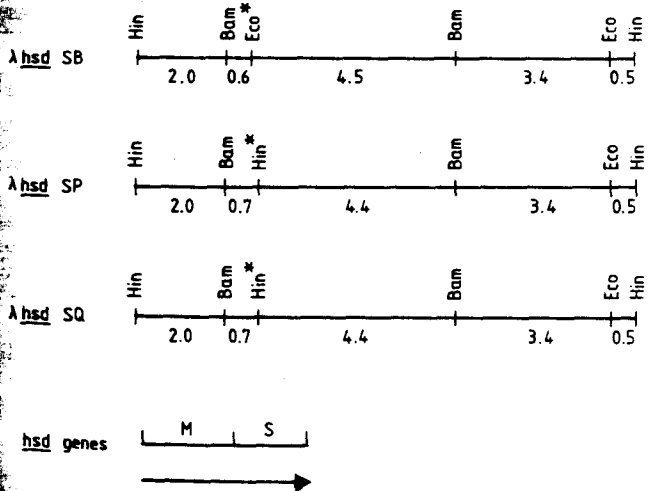


FIG. 1. The λ *hsd* SB, SP, and SQ specificity phages. The *Eco*RI (*Eco*), *Hind*III (*Hin*), and *Bam*HI (*Bam*) targets within and flanking the 11-kilobase insert of each λ *hsd* phage are indicated. A probe for the *hsdM* gene hybridizes to DNA to the left of the targets marked with an asterisk, while a 400-bp probe spanning the central conserved region of the *hsdS* gene of *E. coli* K-12 hybridizes to DNA to the right of the same targets. The deduced orientation of *hsdM* and *S*, the reverse of that suggested previously (6), is confirmed by the nucleotide sequence (Fig. 4). The arrow identifies the direction of transcription.

since transcription is initiated upstream of *hsdM* (19), the *hsdS* gene of SQ specificity derives its proximal variable region from SP and its distal variable region from SB.

The nucleotide sequences of the central conserved regions of the three *hsdS* genes show that the crossover between the SP and SB specificity genes occurred within the longest section of perfectly conserved sequence (Fig. 2). The distal conserved regions of the *hsdS* genes of SB and SQ systems have identical sequences but differ from SP (data not shown, but for amino acid sequences, see Fig. 3C). This is consistent with the origin of SQ by a single crossover.

The SQ and SP systems recognize the same trinucleotide component of the target sequence (5' AAC). Therefore, the domain of the S polypeptide conferring recognition of AAC is encoded by the segment of the *hsdS* gene located proximal to the region within which the crossover event occurred. This includes the proximal variable region and the first 33 nucleotides of the central region. The domain of the SB (or SQ) specificity polypeptide conferring recognition of RTAYG (where R = unspecified purine nucleoside and Y = unspecified pyrimidine nucleoside) must be encoded by the DNA distal to the crossover (see Fig. 3A). These conclusions refer to polypeptide domains that discriminate between specific nucleotides. Adenine residues are common to the recognition sequences, and other domains could contribute to their recognition; for example, an active site for methylation encoded by the *hsdM* gene may also dictate specificity for adenine residues as opposed to cytidine residues.

Amino Acid Sequences of the Conserved Regions of the Salmonella Specificity Polypeptides. There is only one possible open reading frame in each central conserved region, translation of which yields amino acid sequences that are conserved with respect to each other and the S polypeptides of *E. coli* (4). These sequences identify minor differences between each system and show where SQ differs from its parents (Fig. 3B). The crossover that generated SQ separates information that specifies recognition of each of the two components of the target sequence. Assuming that the organization of the domains of the *E. coli* polypeptides parallels that in *Salmonella*, we conclude that the information for recognition of the complete target sequence is not entirely within the central conserved region. While the variability within the first 11 amino acids of this region could correlate with differences in the trimeric component of the target sequence, there is insufficient variability distal to the crossover. In particular, for *E. coli* B and *Salmonella* SP, whose recognition sequences are quite different, there is complete identity after the first five amino acids (Fig. 3B).

In an alternative hypothesis (20), the two components of the target sequence are recognized by two conserved domains of the S polypeptide. Argos (20) found that the S polypeptides of *E. coli* contain repeated domains, and these repeated domains overlap the central and distal conserved

```

SP ATACCAATCCCGTCACTTGCTGAACAAAAATCATCGCCGAAAAACTCGATACGCTGCTGCCGAGGTAG
SB GTTCTCTGCCACCTCTTGCCGAACAAAAAGTCATCGCCGAAAAACTCGATACGCTGCTGCCGAGGTAG
SQ ATACCAATCCCGTCACTTGCTGAACAAAAATCATCGCCGAAAAACTCGATACGCTGCTGCCGAGGTAG

SP ACAGCACCAAAGCAGCTCTTGAGCAAATCCGCAAAATCTGAAACGTTTTTCGTCAGGCCGCTGTTA
SB ACAGCACCAAAGCAGCTCTTGAGCAAATCCCAAAATCTGAAACGTTTTTCGCAATCAGTGATA
SQ ACAGCACCAAAGCAGCTCTTGAGCAAATCCCAAAATCTGAAACGTTTTTCGCAATCAGTGATA
    
```

FIG. 2. Localization of the crossover that generated SQ. The nucleotide sequences of the central conserved regions of the SP, SB, and SQ *hsdS* genes are given. This includes base pairs 469-603 of the SP specificity gene (base pairs 695-829 in Fig. 4). The 70-bp region underlined is common to all three sequences and identifies the region in which the crossover occurred.

the beginning of each repeat could be interaction with the M polypeptides, for type I restriction enzymes are believed to include one S and two M subunits (see ref. 7).

The Nucleotide Sequence of the *hsdS* Gene of SP. A contiguous sequence of 1731 bp has been determined (Fig. 4), extending from the *Bam*HI site in *hsdM* (see Fig. 1) through the *Hind*III site to an *Alu* I target \approx 100 bp downstream of the *hsdS* gene of SP: An open reading frame beginning at bp 226 and terminating at bp 1617 encodes a polypeptide of 463 amino acids, 1 amino acid smaller than the S polypeptide of the K system (4). Comparison of this nucleotide sequence with that for *E. coli* K-12 shows good conservation in the regions flanking the *hsdS* gene, including the 3' end of *hsdM*. In addition to the homology already discussed (Fig. 3 B and C), there is strong homology within the proximal regions of the *hsdS* genes where 421 of the first 474 nucleotides are conserved. This contrasts with the complete lack of homology between the corresponding regions of the *hsdS* genes of the K, B, and D systems. The predicted sequences of the S polypeptides (Fig. 5) show that 143 of the 158 residues of the proximal variable regions of SP and K are identical, with 5 of the 15 changes being conservative. This proportion of changed amino acids parallels that found in the distal conserved regions (6 of 88 amino acids). No homology was detected when the DNA sequences of the distal conserved regions were compared, but similarity of amino acid sequences has been found (A. F. W. Coulson, personal communication).

The *hsdS* genes for four natural representatives of the K family have now been sequenced, but only when SP and K are compared is there extensive homology. In this case, the 5' segments of the genes, referred to as proximal variable regions for both the *E. coli* and *Salmonella* systems, are well conserved. The SP and K systems are also distinguished from the remainder by their very similar target sequences; while K

recognizes 5' AAC(N)₆GTGC, the SP enzyme recognizes the degenerate version, 5' AAC(N)₆GTRC. The present analysis of the SB, SP, and SQ systems taken together with their recognition sequences (10) localizes the coding sequence for the domain of the SP specificity polypeptide conferring recognition of 5' AAC as proximal to the region in which crossing-over occurred to generate SQ. The domains of the K specificity polypeptide have not been separated, but the simplest prediction is one in which the organization parallels that deduced for SP—namely, that recognition of the 5' trimeric sequence is dictated by a domain encoded by the amino-terminal part of the polypeptide and by the longer component by a domain within the carboxyl-terminal segment of the polypeptide. The conservation of nucleotide sequence in the 5' regions of the *hsdS* genes of K and SP then correlates with the identity of one component (AAC) of their target sequence.

CONCLUSIONS

Helix-turn-helix motifs, characteristic of many proteins that bind DNA (see ref. 21), have not been detected in the specificity polypeptides of type I restriction systems (4, 20). Recombination between different specificity genes can re-sort two domains of the S polypeptide, each of which specifies recognition of one component of the target nucleotide sequence (6, 10). A comparison of the predicted amino acid sequences places constraints on the localization of these two domains. Sequences in each of the two conserved domains could be involved but appear to show too little diversification for the recognition of different target sequences. Argos (20) proposed that the two recognition domains are within repeated sequences that overlap the central and distal conserved regions, respectively (see Fig. 3A). This allows further diversity because it includes some of



The nucleotide sequence of the *hsdS* gene of SP. The sequence (upper line) is aligned with that of *E. coli* K-12. The *hsdS* gene of at bp 226 and ends at bp 1617; the initiation and termination codons are underlined.

```

1 MNRKLPEDGATAPVSTVTTLIRGVTYRKEQALNYLDDYLPPIRANNIO 50
1 NSAGKLPEDGAVIAPVSTVTTLIRGVTYRKEQALNYLDDYLPPIRANNIO 50
51 NGRFDTTDLVFPVFNKLVKESQKISPEDIVIAMSSGSKSVVGSQAHLRPF 100
51 NGRFDTTDLVFPVFNKLVKESQKISPEDIVIAMSSGSKSVVGSQAHLRPF 100
101 ECSPGAFPCGALRPERFISPNYIAHFTKSFYRNKISSLSAGANINNIKPA 150
101 ECSPGAFPCGVLRPEKLI FSGFIAHFTKSSLYRNKISSLSAGANINNIKPA 150
151 SFDLINIPISPLAEOKIIAEKLDTLAQQVSTKARLEQIPIQLKRFQAV 200
151 SFDLINIPISPLAEOKIIAEKLDTLAQQVSTKARLEQIPIQLKRFQAV 200
201 LAAAVSGTLTALRN.SHSLIGWHTNLGALIVDSNGLAKROGLNGEI 249
201 LGGAVNGKLTKEWRNFEPQHSVFKLNFESILTELRNGLSSKPNESGVGH 250
250 TILRLADFKDAQRIIGNERRIKLDSKEENKYSLENDLIVRVNGSADLA 299
251 PILRISVVLRAGHVQNDIRPLECSSELNRHKLQGDLLFTRYNGSLEFV 300
300 GRFIETKSENGDIEGCFDHFIRLRLDNSKIMSRLFTYIANEGEGRFYLRNS 349
301 GVCGLLKKLOHONLLYPDKLIRARLTGDALPEYIEIFFSSPARNAMMC 350
350 LSTSAGQMTINOTSIKGLSFLPLKEQAEIVRRVEQLFAYADTIEKQVN 399
351 VKTTSQKGIKSGDKIKSQVLLFPVKEQAEIVRRVEQLFAYADTIEKQVN 400
400 NALTRVMSLTOSILAKAFRGELTAQWRAENPDLSGKNSAAALEKIKAE 449
401 NALARVMSLTOSILAKAFRGELTAQWRAENPDLSGKNSAAALEKIKAE 450
450 RAVSGGKETSRRKA 463
451 RAASGGKASRRKS 464

```

FIG. 5. The predicted amino acid sequence (in single-letter code) of the specificity polypeptide of SP. The sequence (upper line) is aligned with that of *E. coli* K-12 (lower line). The proximal variable region, as defined by Gough and Murray (4), ends at amino acid 158; the central conserved region, at amino acid 201; and the distal variable region, at amino acid 377 (see Fig. 3A for a diagrammatic representation).

the adjacent variable regions. Although these sequences are not predicted to be helical, they are adjacent to prominent α -helical domains, and they are within the most conserved parts of the repeats identified by Argos (20).

No experimental evidence argues against a correlation of the variable regions with the recognition domains—a concept that now receives circumstantial support from the near identity of the amino acid sequences of the amino-terminal segments of the K and SP polypeptides. It is tempting to assume that the variable regions contribute to the specificity most simply by including the recognition domains. Alternatively, they, or more particularly the parts adjacent to the

α -helical regions, could alter the presentation of the actual DNA binding domains within the relatively conserved regions. If single-base changes alone are sufficient to dictate a different specificity of recognition, then it should be possible to isolate mutants with novel specificities. Such mutants have not been reported, and we have failed to select mutations that relax the specificity of K to that of SP. We suggest that the variable regions provide diversity of recognition, and natural recombination can add to this diversity by reassortment of existing domains.

We thank A. Campbell and A. Daniel for the nucleotide sequence of the SB and SQ specificity genes; E. Kawashima (Biogen, Geneva) for synthetic oligonucleotides; T. Bickle, J. Collins, A. Coulson, A. Gann, D. Meek, and K. Murray for constructive criticism of the manuscript; K. Harris and A. Wilson for help in the preparation of the manuscript; and the Medical Research Council for support.

- Boyer, H. W. & Roulland-Dussoix, D. (1969) *J. Mol. Biol.* **41**, 459–472.
- Hubacek, J. & Glover, S. W. (1970) *J. Mol. Biol.* **50**, 111–127.
- Van Pel, A. & Colson, C. (1974) *Mol. Gen. Genet.* **135**, 51–60.
- Gough, J. A. & Murray, N. E. (1983) *J. Mol. Biol.* **163**, 1–19.
- Bullas, L. R., Colson, C. & Van Pel, A. (1976) *J. Gen. Microbiol.* **95**, 166–172.
- Fuller-Pace, F. V., Bullas, L. R., Delius, H. & Murray, N. E. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 6095–6099.
- Bickle, T. A. (1982) in *Nucleases*, eds. Linn, S. M. & Roberts, R. J. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 85–108.
- Nagaraja, V., Steiger, M., Nager, C., Hadi, S. M. & Bickle, T. A. (1985) *Nucleic Acids Res.* **13**, 389–399.
- Nagaraja, V., Shepherd, J. C. W., Prippl, T. & Bickle, T. A. (1985) *J. Mol. Biol.* **182**, 579–587.
- Nagaraja, V., Shepherd, J. C. W. & Bickle, T. A. (1985) *Nature (London)* **316**, 371–372.
- Messing, J. (1983) *Methods Enzymol.* **101**, 20–78.
- Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3963–3965.
- Deininger, P. L. (1983) *Anal. Biochem.* **129**, 216–223.
- Benton, W. D. & Davies, R. W. (1977) *Science* **196**, 180–182.
- Hu, N. & Messing, J. (1982) *Gene* **17**, 271–277.
- Sanger, F., Coulson, A. R., Barell, B. G., Smith, A. J. H. & Roe, B. A. (1980) *J. Mol. Biol.* **43**, 161–178.
- Devereux, J., Haerberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12**, 387–395.
- Staden, R. (1982) *Nucleic Acids Res.* **10**, 4731–4751.
- Sain, B. & Murray, N. E. (1980) *Mol. Gen. Genet.* **180**, 35–46.
- Argos, P. (1985) *EMBO J.* **4**, 1351–1355.
- Pabo, C. O. & Sauer, R. T. (1984) *Annu. Rev. Biochem.* **53**, 293–321.