Molecular Microbiology (1993) 9(1), 133–143

Conservation of motifs within the unusually variable polypeptide sequences of type I restriction and modification enzymes

Noreen E. Murray,[1]* Anne S. Daniel,[2] Gill M. Cowan[3] and Paul M. Sharp[4]
[1]*Institute of Cell and Molecular Biology, University of Edinburgh, Edinburgh EH9 3JR, UK.*
[2]*The Rowett Research Institute, Aberdeen AB2 9SB, UK.*
[3]*Institute of Molecular Medicine, Oxford OX3 9DU, UK.*
[4]*Department of Genetics, Trinity College, Dublin 2, Ireland.*

Summary

Type I restriction enzymes comprise three subunits encoded by genes designated *hsdR*, *hsdM*, and *hsdS*; S confers sequence specificity. Three families of enzymes are known and within families, but not between, *hsdM* and *hsdR* are conserved. Consequently, interfamily comparisons of M and R sequences focus on regions of putative functional significance, while both inter- and intrafamily comparisons address the origin, nature and role of diversity of type I restriction systems. We have determined the sequence of the *hsdR* gene for *Eco*A, thus making available sequences of all three *hsd* genes of one representative from each family. The predicted R polypeptide sequences share conserved regions with one superfamily of putative helicases, so-called 'DEAD box' proteins; these conserved sequences may be associated with the ATP-dependent translocation of DNA that precedes restriction. We also present *hsdM* and *hsdR* sequences for *Eco*E, a member of the same family as *Eco*A. The sequences of the *M* and *R* genes of *Eco*A and *Eco*E are at least as divergent as typical genes from *Escherichia coli* and *Salmonella*, perhaps as the result of selection favouring diversity of restriction specificities combined with lateral transfer among different species.

Introduction

It is usually assumed that restriction and modification (R-M) systems serve as barriers to the uptake and propagation of foreign DNA. Their ability to protect bacteria against viral infection is amply demonstrated in the laboratory, but the importance of such protection in nature is difficult to evaluate. Nevertheless, it is readily arguable that a restriction system of novel specificity would be advantageous to a bacterial strain establishing itself in a new ecological niche (Levin, 1986). An alternative role has been suggested in which the ends of DNA molecules that result from restriction may be substrates for recombination and hence restriction could facilitate the stable acquisition of foreign DNA (Lederberg, 1973; Price and Bickle, 1986). These proposed roles share the need to distinguish self from foreign DNA and the effectiveness of either will depend on a member of a bacterial species having a restriction system of different sequence specificity.

All type I R-M systems comprise three different subunits, R, M and S, and their alternative activities of restriction and modification are influenced by the methylation state of the target sequence. Different specificity (S) polypeptides confer alternative sequence specificities to the complex. An S polypeptide has two recognition domains, each specifying one component of the bipartite target sequence (Gough and Murray, 1983; Fuller-Pace *et al.*, 1984; Nagaraja *et al.*, 1985). One consequence of having two discrete recognition domains is that either their reassortment by genetic recombination between different *hsdS* genes (Gann *et al.*, 1987; Gubler *et al.*, 1992), or a change in the spacing between them (Price *et al.*, 1989; Gubler and Bickle, 1991), can create a new specificity. A single subunit that concomitantly confers sequence specificity to both restriction and modification facilitates the acquisition of new specificities. Indeed, changes in specificity were first observed by chance following transfer of chromosomal genes by transduction (Bullas *et al.*, 1976) and the *Eco*R124 plasmid by transformation (Hughes, 1977). The mechanisms outlined above explain the origins of new specificities in the laboratory, and are probable sources of natural variation. *Eco*K and *Sty*SP, both of which are natural members of the same family of enzymes, have in common the recognition domain that specifies the 5' component of the target sequence, and it seems likely that during evolution this domain has, as a consequence of recombination, become associated with different recognition domains for the 3' component of the target sequence (Gann *et al.*,

Received 28 January, 1993; revised and accepted 19 March, 1993. *For correspondence. Tel. (031) 650 5374; Fax (031) 668 3870.

1987). No information is available about the evolution of new recognition domains, and experiments designed to select for a relaxation in specificity from 5'-AAC(N_6)GTGC to 5'-AAC(N_6)GTG/AC have been unsuccessful (N. E. Murray, unpublished). It thus appears that even a minor change of sequence specificity may require more than one amino acid change.

The diversification of type I *hsd* genes within the *Escherichia coli* species has not been confined to the gene specifying the S polypeptide. Some strains have alternative R-M systems encoded by *hsd* genes that are apparently allelic, but have very dissimilar nucleotide sequences (Fuller-Pace *et al.*, 1985; Suri and Bickle, 1985). For example, when the M polypeptide sequences of *Eco*K and *Eco*A, the prototypes of these two alternative families of enzymes (now called Ia and Ib, respectively), were compared the amino acid identity was only 32% (Sharp *et al.*, 1992); this level of identity indicates a common ancestry, but is so low that it implies an evolutionary separation greatly in excess of that between *E. coli* and *Salmonella*. A third family of type I R-M systems (Ic) identified in *E. coli* includes the plasmid-borne system *Eco*R124 (Firman *et al.*, 1985), which is even less similar to either *Eco*K or *Eco*A than these two are to each other (Price *et al.*, 1989; Sharp *et al.*, 1992).

When the inferred M polypeptide sequences of representative members of the three different type I systems are compared, the similarities are limited and point to a few regions presumed to be of functional relevance to the methylation of adenine residues (Sharp *et al.*, 1992). The sequences of the *hsdR* genes have been reported for only two enzymes: a type Ia (*Eco*K; Loenen *et al.*, 1987), and a type Ic (*Eco*R124; Price *et al.*, 1989). Both genes encode polypeptides containing motifs characteristic of ATP-binding sites (Walker *et al.*, 1982), and both include additional sequences common to a variety of ATP-dependent helicases (Gorbalenya and Koonin, 1991). Nevertheless, in the published alignment of these two R polypeptide sequences (Price *et al.*, 1989) none of the several components that they share in common with one subgroup of ATP-dependent helicases, the so-called 'DEAD box' proteins, is coincident. In this paper, we report the sequence for the *hsdR* gene of *Eco*A, the archetypal member of the third family (Ib) of type I R-M genes. The predicted sequence of the R polypeptide appears to include all of the seven motifs of the 'DEAD box' proteins that were identified by Gorbalenya and Koonin (1991) and is readily aligned with that of *Eco*K.

We also present sequences of both the *M* and *R* genes for *Eco*E, a second member of the Ib family. These sequences are compared with those encoding *Eco*A and for both pairs of genes our analyses indicate levels of variation characteristic of comparisons between rather than within species of bacteria. This high level of divergence is evident both at silent nucleotide sites and for nucleotide changes that confer amino acid replacements. Recombination (i.e., horizontal transfer), combined with selection for maintaining diversity, are the proposed explanations for this intraspecific diversity.

Results and Discussion

The R subunit of EcoA*: conservation of motifs among ATP-dependent helicases*

In the 3.5 kb DNA fragment known to include the *hsdR* gene of *Eco*A (Fuller-Pace *et al.*, 1985) the nucleotide sequence (data not shown; EMBL/GenBank/DDBJ Nucleotide Sequence Data Libraries accession number L18758) identifies one long open reading frame that specifies a polypeptide with a molecular weight calculated to be 92 000, in good agreement with the value of 95 000 estimated from the mobility of the R polypeptide in SDS–polyacrylamide gels. Immediately upstream from the first methionine codon is a purine-rich region that could serve as a ribosome-binding site (Shine and Dalgarno, 1974) and downstream a non-coding sequence of ~100 bp separates *hsdR* from *hsdM*.

The inferred polypeptide sequence was aligned with those reported for the R polypeptides of *Eco*K (type Ia) and *Eco*R124 (type Ic). Pairwise alignment of the K and A polypeptides was straightforward and indicated 26% amino acid identity. This value is somewhat lower than that found for the M polypeptides (32%; Sharp *et al.*, 1992), but nevertheless consistent with a common ancestry for the *Eco*K and *Eco*A *hsdR* genes. The inclusion of the R polypeptide of *Eco*R124 made alignment more difficult, and quite sensitive to the choice of gap penalties. The level of identity between *Eco*R124 and either *Eco*A or *Eco*K is sufficiently low that these comparisons fall into the 'twilight zone' (Doolittle *et al.*, 1986) where it can be difficult to distinguish between extreme divergence and convergence. Nevertheless, the PILEUP Program using the recommended gap penalties (see Fig. 1), achieved an alignment in which all but one of the various motifs shared with one subgroup of helicases (see below) were found to align among the three R polypeptides. In this alignment, the *Eco*R124 polypeptide has 17–18% amino acid sequence identity to those from *Eco*K and *Eco*A.

The R polypeptides of type I R-M systems are essential for restriction, but not modification. Restriction by type I R-M systems, but not modification, is ATP-dependent; ATP is required as both cofactor and substrate (Bickle *et al.*, 1978). Thus, the presence of an ATP-binding motif in the R polypeptide sequences of *Eco*K and *Eco*R124 (Loenen *et al.*, 1987; Price *et al.*, 1989) correlates with the ATP-dependence of restriction, including the associated ATPase activity. The R polypeptide of *Eco*A, as anticipated, also includes the motifs characteristic of an ATP-binding protein.

Many enzymes containing the purine NTP-binding motif have been identified as members of a superfamily of known, or putative, helicases sharing a number of motifs including GXGKS/T, an abridged version of the 'A' component of the ATP-binding motif, instead of the classical sequence GA XXX GKS/T (Walker *et al.*, 1982). In one of these, the 'DEAD box' family, so called because prototypic members shared the motif 'Asp–Glu–Ala–Asp', there are subgroups which include either the motif 'DEAD' or 'DEXH' (where X is commonly A or C), and several other conserved regions (Linder *et al.*, 1989). Some of these proteins are known to be helicases (Schmid and Linder, 1992). Gorbalenya and Koonin (1991) searched the R polypeptide sequences of *Eco*K and *Eco*R124 for the presence of each of seven motifs they identified in one 'superfamily' of 'DEAD box' proteins. They readily identified five of these (I, Ia, II, V and VI) in both R polypeptide sequences. The counterpart to motif III was only identified with any confidence in the R polypeptide of *Eco*R124 and that of motif IV only in the R polypeptide of *Eco*K. These seven regions are indicated in our alignment of the three R polypeptides (Fig. 1). In this alignment the five conserved sequences that were readily identified by Gorbalenya and Koonin (1991) in both the K and R124 polypeptides are coincident with each other, and with similar sequences in the A polypeptide. In addition, region III of *Eco*R124 is almost coincident with the putative region III of *Eco*K, while region IV in *Eco*K aligns with a similar sequence in *Eco*A. The alignment, in the R polypeptides, of the regions common to one subgroup of 'DEAD box' proteins adds support to their consideration as regions of functional significance to restriction. This 'DEXH' subgroup also includes the Rad3 protein of yeast, a protein shown to have DNA-dependent ATPase activity and a helicase activity dependent on ATP hydrolysis (Sung *et al.*, 1988; Harosh *et al.*, 1989).

Cleavage of DNA by type I R-M enzymes occurs at non-specific sequences remote from the recognition target. This follows the formation of looped intermediates in which the enzyme is simultaneously bound to both the recognition sequence and the cleavage site. It has been proposed that DNA is 'pumped' past the bound enzyme using the energy released by the hydrolysis of ATP (Yuan *et al.*, 1980) and it is tempting to speculate that helicase activity may correlate with the ATP-dependent translocation of DNA. This suggestion appears to be contradicted by the observation (Gorbalenya and Koonin, 1991) that the R subunit of *Eco*P1, a type III R-M system, contains the same six motifs as *Eco*R124; although the restriction activity of type III R-M systems is stimulated by ATP, these enzymes cut DNA close to what classically have been termed their recognition sequences and DNA translocation has not been implicated. It may be relevant, however, that the restriction target for a type III R-M enzyme is now known to comprise two so-called recognition sequences, each of the same six nucleotides separated by an unspecified distance, but in inverse orientation (Meisel *et al.*, 1992). Neither the mechanism of cutting nor that of signalling between the two components of this bipartite target sequence is known. It therefore seems appropriate to consider, for type III R-M systems, a model of the sort suggested for *Eco*K (type Ia) (Studier and Bandyopadhyay, 1988), in which DNA cleavage is triggered when translocating enzymes meet. On this model, type I enzymes remain bound to the recognition site, translocate DNA past themselves from both sides, and cut the DNA near the point at which neighbouring enzymes meet. Perhaps type III enzymes behave in a similar way but translocate DNA from one side only and cut a fixed distance (25–27 nucleotides) from the bound recognition site when translocating enzymes meet. The patterns of DNA fragments produced by digesting a variety of plasmid substrates with a type III restriction enzyme in the presence of ATP (Meisel *et al.*, 1992) are readily interpreted in this way (F. W. Studier, personal communication). Although ATP hydrolysis is not essential for restriction by type III enzymes, a model in which ATP-stimulated helicase activity is associated with DNA translocation would provide a common explanation for strand separation associated with both type I and type III restriction enzymes. Alternative hypotheses are that the helicase activity is required for local unwinding of DNA in the region of the cleavage sites (Gorbalenya and Koonin, 1991) or in the region of the recognition sequence itself.

Divergence among the R subunits of type Ia, Ib and Ic R-M systems

We have argued that the presence in *E. coli* of alternative families of highly divergent, 'pseudoallelic' *hsd* genes (the Ia and Ib families) is most readily explained by horizontal transfer among different species which themselves are much more remote in evolutionary time than *E. coli* and *Salmonella* (Sharp *et al.*, 1992). Comparison of silent sites in the *hsdM* genes of *E. coli* K-12 and *Salmonella enterica* serotype typhimurium LT2 (formerly *Salmonella typhimurium* LT2) showed the expected level of divergence for these strains, based on comparisons of many other genes (Sharp 1991), and indicated that the M polypeptides are relatively highly conserved. Thus it seems unlikely that the extremely high levels of divergence seen between members of different families can be explained by continuous, within-species divergence, uninterrupted by lateral transfer. The R polypeptides of *Eco*K and *Eco*A, like the M polypeptides, exhibit extraordinarily high divergence. In addition, the S genes have

```
                                                                                                              100
K      MLWALNIWTGPHSNGLITMMNKSNFEFLKGVNDFTYAIACAAENNYPDDPNTTLIKMRMFGEATAKHLGLLLNIPPCENQHDLLRELGKIAFVDDNILSV

                                                                                                              200
K      FHKLRRIGNQAVHEYHNDLNDAQMCLRLGFRLAVWYYRLVTKDYDFPVPVFVLPERGENLYHQEVLTLKQQLEQQVREKAQTQAEVEAQQQKLVALNGYI
R124                                                           MTHQTHTIAESNNFIVLDKYIKAEPTGDSYQSESDLERELIQDLRNQG

                                                                                                              300
K      AILEGKQQETEAQTQARLAALEAQLAEKNAELAKQTEQERKAYHKEITDQAIKRTLNLS-----EEESRFLIDAQLRKAGWQADSKTLRFSKGARPEPGV
A                                MAELN--LSNLTEAD------IITKCVMPAILNAG----------WDNTTQIRQEVKLRDGKVI-----------V
R124   YEFISVKSQS-----AMLANVREQLQNLNGVVFN--DSEWRRFTEQYLDNPSDGILDKTRKIHIDYICDFIFDDERLENIYLIDKKNL-----------M

                                                                                                              400
K      NKAIAEWPTGKDETGNQGFADYVLF-VGLKPIAVVEAKRNNIDVPARLNESYRYSKCFDNGFLRETLLEHYSPDEVHEAVPEYETSWQDTSGKQRFKIPF
A      RGKVAARRTVKS-------ADIVLYHKPGIPLAVIEAKANKHEIGKGMQQGIEYARLLD---------------------------------------VPF
R124   RNKVQIIQQFEQAGSHANRYDVTIL-VNGLPLVQIELKKRGVAIREAFNQIHRYSK---------------------------ESFNSENSLFKYLQLF

                                                                                                              500
K      CYSTNGRE--YRATMKTKSGIWYRDVRDTRNMSKALPEWHR-PEELLEML--------GSEPQKQNQWFADNPGMSELGLRYYQEDAVRAVEKAIVKGQQ
A      VFATNGDGFIFRDATAAEGECLEKQI--------TLDDFPS-PAELWQKFCLWKGYTQAQLPVITQDYYDDGSGKSP---RYYQLQAINKTIEAVSNGQN
R124   VISNGTDTRYFANTTKRDKNSFDFTMNWAKSDNTLIKDLKDFTATCFQKHTLLNVLVNYSVFDSSQTLLVMRPYQIAATERILWKIKSSFTAKNWSKPES

          I                      Ia                                                                           600
K      EILLAMATGTGKTRTAIAMMFRLIQSQRFKRILFLVDRRSLGEQA----LGAFED--TRINGDTFNSIFDIKGLTDKFPEDSTKIHVATVQSLVKRTLQS
A      RVLLVMATGTGKTYTAFQIIWRLWKSKNKKRILFLADRNILVDQTKNNDFQPFGTAMTKVSGRTIDPAYE--------------IHLALYQAITGPEEDQ
R124   GGYIWHTTGSGKTLTSFKAARLATELDFIDKVFFVVDRKDLDYQTMKE--------YQRFSPDSVNGSENTAGLKRNLDKDDNKIIVTTIQKLNNLMKAE
          ATP-binding
                  II                     III                                                                  700
K      DEPMPVAR--YDCIVVDEAHRGYILDKEQTEGELQFRSQLDYVSAYRRILDHFECGKNRSHRHPGATYCADFRRAGLPLYLPYRGYRRFSDRPGSAYSDH
A      KAFKQVAPDFFDLIVIDECHRG------SASEDSAWREILDYFSSATQI--GLTATPKETHEVSSTDYFGD------PVYV----YSLKEGIEDGFLAPY
R124   SDLPVYNQQV--VFIFDECHRS-QFGEAQKNLKKKFKRYYQFGFTGTPIFPENALGSETTASVFGREL---------------HSYVITDAIRDEKVLKF

       IV                                                                                    IV               800
K      HPQRAGGGLSLQRRAGRAHQPAGEVINDTLEDDQDFEVADFNRGLVIPAFNRAVCNELTNYLDPTGS-QKTL-----VFCVTNAHA-----DMVVEE---
A      KVVRVDIDVDLQ---GWRPTKGQTDLNGEVIDDRIYNQKDFDRTMVIDERTELVARTITDYLKRTNPMDKTI-----VFCNDIDHA-----ERMRRA---
R124   KVDYNDV---------RPQFKSLETETDEKKLSAAENQQAFLHPMRIQEITQYILNNFRQKTHRTFPGSKGFNAMLAVSSVDAAKAYYATFKRLQEEAAN
```

```
                                                                                    V    900
K      -------LRAAFKKKYPQLEHDAIIKITGDADKDARKVQTM---------------------------------ITRFNKERLPNIVVTVDLLTTGVDIP
A      -------L-VNLNPEQVKKNDKYVMKITGDDEIGKAQLDNF---------------------------------INP--KKPYPVIATTSELMTTGVDAK
R124   KSATYKPLRIATIFSFAANEEQNAIGEISDETFDTSAMDSSAKEFLDAAIREYNSHFKTNFSTDSNGFQNYYRDLAQRVKNQDIDLLIVVGMFLTGFDAP

               VI                                                                                  1000
K      SICNIVFLRKVRSRILYEQMKGRATRLCPEVNKTSFKIFD----------CVDIYS------------T--LESV-----DTMRPVVVRPKVELQTLVN
A      TCKLVVLDQNIQSMTKFKQIIGRGTRIDERYGKLWFTILDFK-------KATELFA------------DERFDGIPEKVMDTTPEDIADPESDFEEKLE
R124   TLNTLFVDKNLRYHGLM-QAFSRTNRIYDA-TKTFGNIVTFRDLERSTIDAITLFGDKNTKNVVLEKSYTEYMEGFTDAATGEAKRGFMTVVSELEQRFP

                                                                                                   1100
K      EITDSE------------TYKITEADGRSFAEHSHEQLVAKLQRIIGLATFNRDRSETIDKQVRRLDELCQ-DAAGVNFNGFAS-----RLREKGPHWS
A      EISEHDEEQVTGVDEPPAPPYQVTDTDDVGPLPEEDEKKIRKFHVNGVAVGVIAQRVQYYDADGKLVTESFKDYTRKTLLKEYASLDDFTRKWQDADRKE
R124   DPTSIESEKEKKDFVKLFGEYLRAENILQNYDEFATLKALQQID-LSDPVAVEKFKAEHYVDDEKFAELQTIRLPADRKIQDYRSAYNDIRDWQ---RRE

                                                                                                   1200
K      AEVFNKLPGFIAR------LEKLKTDINNLN--------DAPIFLDIDDEVVSVKSL---------------------YGDYDTPQDFLEAFDSLVQRS
A      AIIHELEQQGIIWEVL---AEEVGKDLDPFDMLCHVVYGQPPLTRKERAENVRKRNY---------------------FTKYSEAAQ--AVLDNLLDKY
R124   KEAEKKEKSTTDWDDVVFEVDLLKSQEINLDYILGLIFEHNRQNKGKGEMIEEVKRLIRSSLGNRAKEGLVVDFIQQTNLDDLPDKASIIDAFFTFAQR-

                                          1250
K      PNAQPALQAVINRPRDLTRKGLVELQEWFDRQHFEESSLRKAWKETRMKISPPG
A      ADAGVQEIESIQVLKLKPFDSMGTLPEIIKTGFGDRNGYNQALSELENEIYQLPPRSA
R124   -EQQREAEALIKEENLNEDAAKRYIRTSLKREYATENG--TELNETLPKLSPLNPQYKTKKQAVFRKSSRLLRSLKA
```

Fig. 1. Alignment of the R polypeptides of the *Eco*K, *Eco*A and *Eco*R124 R-M systems using the PILEUP Program with a gap weight of 3.0 and a gap length weight of 0.1. Underlined are the regions (Ia, Ib, II, III, IV, V and VI) identified in *Eco*K and *Eco*R124 by Gorbalenya and Koonin (1991) as the seven conserved regions characteristic of 'DEAD box' proteins of superfamily II; those identified with confidence are emphasized by bold type. The bold type is extended to those sequences in *Eco*A confidently predicted to be regions I, Ia, II, IV, V and VI. Two alternative alignments of motif III in *Eco*A are given in Fig. 3.

regions of base composition uncharacteristic of *E. coli*, which may have been acquired by lateral transfer (Dila *et al.*, 1990).

The finding of a type I R-M system (*Eco*R124) on a plasmid may be important to this argument in two respects. First, plasmids are an obvious vehicle for lateral transfer. Second, it may be that plasmid-borne sequences evolve at faster rates than genes on the chromosome. If the divergence between *Eco*K and *Sty*SB can be taken as a calibration point for the rate of evolution, and if the rate of evolution had been relatively constant, the level of divergence between the type Ia and type Ib families is such that it would require that the sequences diverged very early in the history of Gram-negative bacteria, or perhaps even earlier. However, if sequences on plasmids evolve faster, and if the R-M genes have spent part of their history on plasmids, the time of divergence could have been far more recent. This may seem more reasonable than invoking lateral transfer among very divergent species.

Divergence of EcoA *and* EcoE

The levels of sequence divergence within families are also particularly interesting in pointing to the possible evolutionary mechanisms underlying their divergence. For example, within the Ia family the extent of intraspecific divergence between *Eco*K and *Eco*B (from *E. coli* strain B) is unexpectedly high (Sharp *et al.*, 1992). Therefore, we have determined the nucleotide sequences encoding the M and R polypeptides of *Eco*E, a second member of the type Ib family.

The nucleotide sequences of the *hsdM* and *hsdR* genes of *Eco*E (not shown; EMBL/GenBank/DDBJ Nucleotide Sequence Data Library accession number L18759) predict polypeptide sequences with high identity to those of *Eco*A (Figs 2 and 3). The gene sequences are sufficiently similar that the numbers of nucleotide substitutions per synonymous (K_S) and per non-synonymous site (K_A) can be estimated (this is not true for interfamily comparisons). The K_S values for both *hsdR* and *hsdM* (Table 1) suggest that the divergence of these genes, despite being members of the same family, could predate the separation of *E. coli* and *Salmonella* from their common ancestor (see also Fig. 4). The K_S value for *hsdR*, the gene remote from *hsdS*, is even higher than that for *hsdM*, which in turn is very much higher than that estimated for the comparison of the *hsdM* genes of *E. coli* K-12 with that of B. The K_S values for *hsdM* and *hsdR* either approach or exceed the value of 1.5 at which sites are almost saturated with silent substitutions and hence these K_S values have large standard errors. For this reason we shall ignore the difference in K_S values for M and R and stress that for both genes the figures exceed the average

```
            .         .         .         .         .         .         .         .         .       100
A  MSISSVIKSLQDIMRKDAGVDGDAHGLGQLSWLLFLKIFDAQEEALELEQDNYQYPIPQRYLWRSWAANAQGITGDSLLEFVNDDLFPALKNLTAPIDKN
E  MSISSVIKSLQDIMRKDAGVDGDAQRLGQLSWLLFLKIFDTQEEELELEQDDYQFPIPQRYLWRSWAANSEGITGDALLEFVNDDLFPTLKNLTAPIDKN
   ************************  ************** *** ****** ** **************  ***** *********** ***********

            .         .         .         .         .         .         .         .         .       200
A  PRGYVVKQAFSDAYNYMKNGTLLRQVINKLNEIDFTSASERHLFGDIYEQILKDLQSAGNAGEFYTPRAVTRFMVDRVDPKLGESIMDPACGTGGFLACA
E  PRGFVVKQAFSDAYNYMKNGTLLRQVINKLNEIDFSSSQERHLFGDIYEQILRDLQSAGNAGEFYTPRAVTRFMVNRIDPKLGESIMDPACGTGGFLACA
   *** ******************************* *  ************* ********************** * **********************

            .         .         .         .         .         .         .         .         .       300
A  FDHVKNKYVKSVADHQTLQQQIHGVEKKQLPHLLATTNMLLHGIEVPVQIRHDNTLNKPLSSWDEQLDVIVTNPPFGGTEEDGIEKNFPAEMQTRETADL
E  FDHVKDNYVKTTEDHKTLQQQIYGVEKKQLPHLLCTTNMLLHGIEVPVQIRHDNTLNKPLSSWDEQVDVIVTNPPFGGTEEDGIEKNFPAEMQTRETADL
   *****  ***   ** ****** *********** ******************************* *********************************

            .         .         .         .         .         .         .         .         .       400
A  FLQLIVEVLAKNGRAAVVLPDGTLFGEGVKTKIKKLLTEECNLHTIVRLPNGVFNPYTGIKTNLLFFTKGQPTKEIWFYEHPYPAGVKNYSKTKPMKFEE
E  FLQLIIEVLADKGRAAVVLPDGTLFGEGVKTKIKKLLTEECNLHTIVRLPNGVFNPYTGIKTNILFFTKGQPTKEVWFYEHPYPDGVKNYSKTKPMKFEE
   ***** ****  *************************************************** *********** ********* **************

            .         .         .         .         .         .         .         .       490
A  FQAEIDWWGNEADGFASRVENEQAWKVSIDDVIARNFNLDIKNPHQAETVSHDPDELLAQYAKQQAEIQTLRNQLRDILGAALSV-KEVN
E  FQAEIDWWGNEADDFASREENNQAWKVGIDDIIARNFNLDIKNPYQGETISHDPDELLAQYQTQQAEIGELRNQLRDILGAALAGNKGAN
   ************* **** ** ***** *** ************ * ** ***********  *****  *************   *  *
```

Fig. 2. Alignment of the M polypeptides of *Eco*A and *Eco*E. Amino acid identities are indicated by an asterisk. The *Eco*A polypeptide sequence was published previously (Sharp *et al.*, 1992).

```
                .         .         .         .         .         .         .         .         .        100
A  MAELNLSNLTEADIITKCVMPAILNAGWDNTTQIRQEVKLRDGKVIVRGKVAARRTVKSADIVLYHKPGIPLAVIEAKANKHEIGKGMQQGIEYARLLDV
E  MAGVNKTHLTETDIITKFILPAVKDAGWDVMSQIRQEVKLRDGKIVVRGKLASRIKVKSADIVLYHKPNLPLAVIEAKANKHAISKGMQQGLDYASLLDV
   **  *   *** *****  **   ****   ************  **** * *  ************  ************ * ******  ** ****

                .         .         .         .         .         .         .         .       . I  200
A  PFVFATNGDGFIFRDATAAEGECLEKQITLDDFPSPAELWQKFCLWKGYTQAQLPVITQDYYDDGSGKSPRYYQLQAINKTIEAVSNGQNRVLLVMATGT
E  PFVFASNGDGFIFHDKTNPQQ--LESEIALSDFPTPEQLWQKYCAWKGFTQEQLPVISQDYFDDGSGKAPRYYQMQAINRTVDAVSAGKNRILLVMATGT
   ***** ******* * *          **  * * *** *  **** * *** ** ***** *** ****** ***** **** *   *** * ** ********

                .         .  Ia     .         .         .         .         .         .       . II   300
A  GKTYTAFQIIWRLWKSKNKKRILFLADRNILVDQTKNNDFQPFGTAMTKVSGRTIDPAYEIHLALYQAITGPEEDQKAFKQVAPDFFDLIVIDECHRGSA
E  GKTYTAFQIIWRLWKAKNKKRILFLADRNILVDQTKRNDFQPFGNAMTKVTGRTIDPAYEVHLALYQAITGPEEHQKAYKQVAPDFFDLIVIDECHRGSA
   ************** ******************* ******* ***** ********* ************ *** *********************

                .    III  .         .         .         .         .         .         .         .    400
A  SEDSAWREILDYFSSATQIGLTATPKETHEVSSTDYFGDPVYVYSLKEGIEDGFLAPYKVVRVDIDVDLQGWRPTKGQTDLNGEVIDDRIYNQKDFDRTM
E  SEDSAWREILEYFGSATQVGLTATPKETEDVSNIDYFGEPVYTYSLKEGIEDGFLAPYKVVRVDIDVDVQGWRPVKGQLDKYGEEIEDRIYNLKDFDRTL
   ********** ** **** *********  **   **** *** ************************ ***** *** *  ** * ***** ******

                .         .   IV    .         .         .         .         .         .       .  V   500
A  VIDERTELVARTITDYLKRTNPMDKTIVFCNDIDHAERMRRALVNLNPEQVKKNDKYVMKITGDDEIGKAQLDNFINPKKPYPVIATTSELMTTGVDAKT
E  VIDERTMLVAQTITDYLKRTNPMDKTIVFCNDIDHADRMRHALVVLNPEQVLKNEKYVMKITGDDDIGKAQLDNFINPKKAYPVIATTSELMTTGVDAQT
   ****** *** ************************ *** *** ****** ** ********** ************** ***************** *

                .    VI   .         .         .         .         .         .         .         .    600
A  CKLVVLDQNIQSMTKFKQIIGRGTRIDERYGKLWFTILDFKKATELFADERFDGIPEKVMDTTPEDIADPESDFEEKLEEISEHDEEQVTGVDEPPA-PP
E  CKLVVLDQNIQSMTKFKQIIGRGTRINEKHGKLWFTILDFKKATELFADPRFDGLPEKVLVVKPGDISDKDSDFNEQLDAENNADGGDNDASEAREDVAD
   ************************* *  ****************** **** ****     * ** *  *** * *       *

                .         .         .         .         .         .         .         .         .    700
A  YQVTDTDDV--GPLPEEDEKKIRKFHVNGVAVGVIAQRVQYYDADGKLVTESFKDYTRKTLLK--EYASLDDFTRKWQDADRKEAIIHELEQQGIIWEVL
E  YQVNRDKQAGNGEFHDDDENKVRKFYVNGVDVKVLAKRVQYYDSDGKLVTESFQDYTRKTMLKDSEYASLDSFVRKWQDAPRKQVIIEELEQLGILWDVL
   ***        *      ** * *** **** * * * ****** ********* ****** **  ****** * ****** **  ** **** ** * **

                .         .         .         .         .         .         .         .         .    800
A  AEEVGKDLDPFDMLCHVVYGQPPLTRKERAENVRKRNYFTKYSEAAQAVLDNLLDKYADAGVQEIESIQVLKLKPFDSMGTLPEIIKTGFGDRNGYNQAL
E  AEEVGKDLDPFDLLCHVVYGQPPLTRQERVANVRKRNYFTKYAEPAQQVLNTLLDKYADEGVQEIEDVQVLKLKPFDTLGRPIEIIKTRFGDKKAYEQAV
   ************ ************* **  *********** * ** **  ******* ******  *********  *   ***** ***    * **

          .  815
A  SELENEIYQLPPRSA
E  NELENEIYQLPPRSA -
   **************
```

Fig. 3. Alignment of the R polypeptides of *Eco*A and *Eco*E. Amino acid identities are indicated by asterisks. The putative motifs of the 'DEAD box' proteins are in bold letters. Two alternatives are suggested for region III. The first, underlined in *Eco*A, is the sequence deduced from the alignment shown in Fig. 1; the second, underlined in *Eco*E, is the sequence aligned by the PILEUP Program with region III of *Eco*R124 when the *Eco*E sequence was added to those for *Eco*A, *Eco*K and *Eco*R124 (alignment not shown).

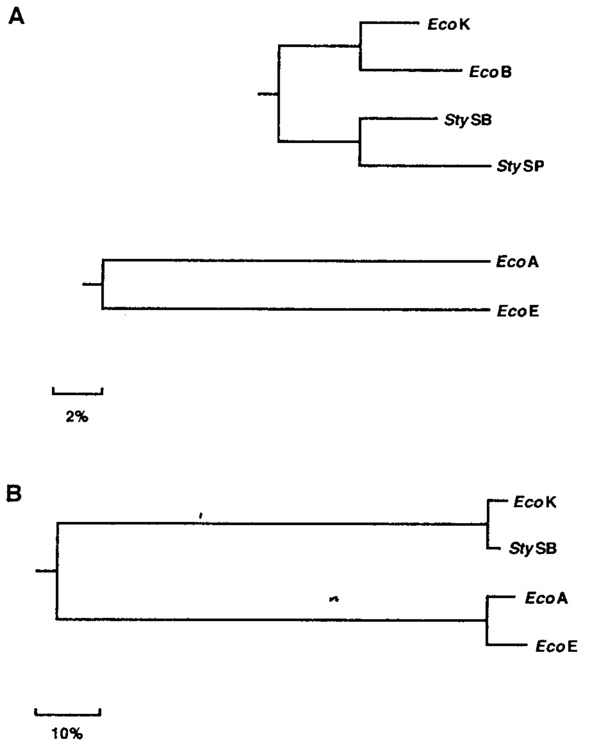

Fig. 4. Divergence between *hsdM* sequences (A) within the Ia and Ib families, and (B) between the Ia and Ib families; these trees emphasize the discrepancy between the phylogenetic relationships of the genes and the species from which they were isolated. Horizontal branch lengths are drawn to scale (both parts of (A) are drawn to a common scale); vertical separation is for clarity only. Divergence values, corrected for multiple hits, were calculated as (A) the estimated number of nucleotide substitutions per site, i.e., the weighted average of substitutions per synonymous and per non-synonymous site, by the method of Li *et al.* (1985), and (B) the estimated number of amino acid replacements per site, by the method of Kimura (1983; Equation 4.8). Relationships were estimated by the Neighbor–Joining method (Saitou and Nei, 1987), and each tree was rooted at the mid-point of the central branch.

values obtained by comparing 67 genes from *E. coli* with those from *Salmonella* (Sharp, 1991; see also Table 1). Thus, while *Eco*A and *Eco*E are both from strains of *E. coli*, the extent of silent site divergence (i.e. K_S values) at both *hsdM* and *hsdR* indicates that the divergence of these genes could predate the speciation of *E. coli* and *Salmonella* (see also Fig. 4), and is consistent with these genes having been subject to horizontal transfer.

Sequence data are not available for other genes from the strains of *E. coli* (15T⁻ and A58) from which the *hsdA* and *hsdE* genes were isolated (Fuller-Pace *et al.*, 1985). However, several gene sequences have been compared among a wide range of natural isolates of *E. coli* (Nelson *et al.*, 1991; Hall and Sharp, 1992, and references therein) and, as expected, the K_S values are typically much lower than those found for comparisons of *E. coli* and *Salmonella*. In comparisons between *E. coli* B and K-12, a parallel was drawn between *hsdM* and *gnd* (Sharp *et al.*, 1992). Both showed high K_S values and both were located immediately adjacent to loci where frequency-dependent selection for rare alleles can be invoked: *hsdS* for alternative specificities of restriction systems, and *rfb* for alternative O antigens. There is one striking difference between either *hsdM* or *hsdR* and *gnd*, namely, the high K_A values observed (Table 1) in contrast with the negligible variation in the amino acid sequences inferred from sequences of *gnd* genes (Bisercic *et al.*, 1991; Dykhuizen and Green, 1991). This may indicate that the diversity at *gnd* results only from its linkage to the *rfb* locus, while diversity at *hsdM* and *hsdR* results from direct selection on these genes. At the present time, selection for variation in sequence of both the M and R polypeptides cannot be ruled out. Although it is known that the R polypeptides of the *Eco*A and *Eco*E systems are interchangeable (Fuller-Pace *et al.*, 1985), it remains possible that hybrid enzymes function less well than native ones. The specificity polypeptides dictate the recognition of different target sequences and could have different three-dimensional structures that might warrant changes in the associated M and R polypeptides to optimize the function of the complex. The sequence variation seen in the alignments shown in Figs 2 and 3 nevertheless suggests considerable tolerance of differences at some residues in the R and M polypeptides of *Eco*A and *Eco*E.

It is quite noticeable that some regions of the R polypeptides of *Eco*A and *Eco*E are well conserved while others are very dissimilar. In one segment (residues 191–515) that includes all seven motifs conserved among

Table 1. Nucleotide sequence divergence among *hsd* genes, compared to others.

	A versus E		K-12 versus B		K-12 versus LT2	
Gene	K_A	K_S	K_A	K_S	K_A	K_S
hsdM	0.058	1.28	0.017	0.24	0.033	0.50
hsdR	0.144	1.84	–	–	–	–
hsdS[a]	0.109	1.36	0.043	0.24	0.054	0.65
Other genes[b]						
average	–	–	0.002	0.05	0.039	0.94
minimum	–	–	0.000	0.00	0.000	0.04
maximum	–	–	0.009	0.20	0.166	1.77

K_A and K_S are the estimated number of nucleotide substitutions per non-synonmyous site, and per synonymous site, respectively, after correction for multiple hits (see the *Experimental procedures*).
a. Calculated for alignable regions: *Eco*A codons 1–400, 572–589; *Eco*E 1–400, 577–594; *Eco*K 158–210, 373–465; *Eco*B 159–211, 383–474; *Sty*SB 154–206, 378–469.
b. For *E. coli* K-12 versus B, 28 genes listed in Sharp *et al.* (1992); for *E. coli* K-12 versus *S. enterica* serovar typhimurium LT2, 69 genes listed in Sharp (1991).

members of the 'DEAD box' family of proteins, the level of identity is high (89%), but in the *N*-terminal region (residues 1–190) the identity is only 72%, and in the *C*-terminal region (residues 516–810) it is 69%. Within the latter segment is a region of 40 amino acids in which only seven (17.5%) of the amino acids are identical. The conserved motifs are conspicuous, having only single differences in regions I and V and two in region III, a region that could be incorrectly defined, since the alignment (not shown) achieved when the R polypeptides of *Eco*A, E, K and R124 are all included identifies region III as a sequence overlapping the one postulated in Fig. 1. The alternative (see Fig. 3) has only one conservative amino acid difference, valine versus isoleucine. The amino acid motif SAT, commonly found in region III, is missing in both *Eco*K and *Eco*R124, though present in the first alternative for *Eco*A. The second alternative has TAT aligned with TGT of *Eco*R124; the former is found instead of SAT in the yeast spliceosome component PRP28 (Strauss and Guthrie, 1991), while SGT is found in the DNA helicase Rad3, and its human homologue ERCC2 (Weber *et al.*, 1990).

The seven motifs total 96 amino acids and with only three or four amino acid differences focus on the most conserved regions of the R polypeptides of *Eco*A and *Eco*E. An analysis of mutants is essential to document the relevance of the putative motifs to the ATP-associated activities of type I and III R-M systems. Nevertheless, the importance of the various motifs to the functions of type I R-M enzymes is supported by both the inter- and intrafamily comparisons of the inferred amino acid sequences of the R polypeptides of *Eco*A and *Eco*E.

Experimental procedures

Bacterial strains, media and microbial techniques

The general host for phages (lambda and M13) and plasmids was NM522 (*lac-pro*)Δ*hsdMS*Δ/F'*lacZ*ΔM15*lacI*^{q} (Gough and Murray, 1983). The other λ*hsd* phages used and their derivative plasmids, pFFP26 (containing the *hsdR* gene of *Eco*A) and pGC1 (containing the entire coding sequence for *Eco*E), have been described elsewhere (Fuller-Pace *et al.*, 1985). Media, general methods and tests for restriction and modification were as before (Fuller-Pace *et al.*, 1985).

DNA manipulations

Preparation, manipulation and recovery of DNA were as described previously (Midgley and Murray, 1985). Template DNA was sequenced by the dideoxy chain termination method (Sanger *et al.*, 1977) using deoxyadenosine 5-[α^{35}S]-thiotriphosphate, and the reactions were analysed on buffer gradient gels (Biggin *et al.*, 1983). Most recently, T7 polymerase and the modifications of Tabor and Richardson (1987) have been used.

Fragments of *hsd* DNA encoding *Eco*E generated by sonication of pGC1 were ligated to M13mp19 cut with *Sma*I. Sequences of cloned fragments were generated using universal primer and the sequences were compiled by the computer programs of Staden (1982). For the *hsdR* gene of *Eco*A, the *Hin*dIII–*Eco*RI restriction fragment from pFFP26 was subcloned in M13mp18 and -19 and oligonucleotide primers at intervals of about 250 bases were used to initiate sequencing reactions. All sequences were determined for both strands.

Sequence analysis

Protein sequences were aligned using the CLUSTAL Multiple Alignment Program (Higgins and Sharp, 1988) and by the PILEUP program (in the UWGCG package; Devereux *et al.*, 1984). Numbers of amino acid replacements per site between aligned sequences were estimated using the empirical method of Kimura (Equation 4.4 in Kimura, 1983). DNA sequence divergence was analysed using the approach of Li *et al.* (1985). In this method, rates of transitions and transversions are estimated separately for 0-fold, twofold and fourfold degenerate sites, and then combined with appropriate weighting to provide estimates of the numbers of nucleotide substitutions per synonymous (K_S), and per non-synonymous site (K_A). Phylogenetic relationships were estimated by the Neighbor–Joining method (Saitou and Nei, 1987).

Acknowledgements

We thank Annette Campbell and Diane Ternant for skilled assistance, Frances Fuller-Pace and present colleagues for constructive discussion, and Fiona Govan for preparing the manuscript. These studies were supported by grants from the Medical Research Council (N.E.M.) and EOLAS (P.M.S.), and a S.E.R.C. Studentship to G.M.C. This is a publication from the Edinburgh Centre for Molecular Recognition and the Irish National Centre for Bioinformatics.

References

Bickle, T.A., Brack, C., and Yuan, R. (1978) ATP-induced conformational changes in the restriction endonuclease from *Escherichia coli* K-12. *Proc Natl Acad Sci USA* **75:** 3099–3103.

Biggin, M.D., Gibson, T.J., and Hong, G.F. (1983) Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. *Proc Natl Acad Sci USA* **80:** 3963–3965.

Bisercic, M., Feutrier, J.Y., and Reeves, P.R. (1991) Nucleotide sequences of the *gnd* genes from nine natural isolates of *Escherichia coli*: evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic *gnd* locus. *J Bacteriol* **173:** 3894–3900.

Bullas, L.R., Colson, C., and van Pel, A. (1976) DNA restriction and modification systems in Salmonella. SQ a new system derived by recombination between the SB system of *Salmonella typhimurium* and the SP system of *Salmonella potsdam*. *J Gen Microbiol* **95:** 166–172.

Devereux, J., Haeberli, P., and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucl Acids Res* **12:** 387–395.

Dila, D., Sutherland, E., Moran, L., Slatko, B., and Raleigh, E.A. (1990) Genetic and sequence organization of the *mrcBC* locus of *Escherichia coli* K-12. *J Bacteriol* **172:** 4888–4900.

Doolittle, R.F., Feng, D.F., Johnson, M.S., and McClure, M.A. (1986) Relationships of human protein sequences to those of other organisms. *Cold Spring Harbor Quant Biol* **51:** 447-455.

Dykhuizen, D.E., and Green, L. (1991) Recombination in *Escherichia coli* and the definition of biological species. *J Bacteriol* **173:** 7257–7268.

Firman, K., Price, C., and Glover, S.W. (1985) The *Eco*R124 and *Eco*R124/3 restriction and modification systems; cloning the genes. *Plasmid* **14:** 224–234.

Fuller-Pace, F.V., Bullas, L.R., Delius, H., and Murray, N.E. (1984) Genetic recombination can generate altered restriction specificity. *Proc Natl Acad Sci USA* **81:** 6695–6099.

Fuller-Pace, F.V., Cowan, G.M., and Murray, N.E. (1985) *Eco*A and *Eco*E; alternatives to the *Eco*K family of type I restriction and modification systems of *E. coli*. *J Mol Biol* **185:** 65–75.

Gann, A.A.F., Campbell, A.J.B., Collins, J.F., Coulson, A.F.W., and Murray, N.E. (1987) Reassortment of DNA recognition domains and the evolution of new specificities. *Mol Microbiol* **1:** 13–22.

Gorbalenya, A.E., and Koonin, E.V. (1991) Endonuclease (R) subunits of type I and type III restriction-modification enzymes contain a helicase-like domain. *FEBS Letts* **291:** 277–281.

Gough, J.A., and Murray, N.E. (1983) Sequence diversity among related genes for recognition of specific targets in DNA molecules. *J Mol Biol* **166:** 1–19.

Gubler, M., and Bickle, T.A. (1991) Increased protein flexibility leads to promiscuous protein-DNA interactions in type Ic restriction–modification systems. *EMBO J* **10:** 951–957.

Gubler, M., Braguglia, D., Meyer, J., Piekarowicz, S., and Bickle, T.A. (1992) Recombination of constant and variable modules alters DNA sequence recognition by type Ic restriction-modification enzymes. *EMBO J* **11:** 233–240.

Hall, B.G., and Sharp, P.M. (1992) Molecular population genetics of *Escherichia coli*: DNA sequence diversity of the *celC*, *crr* and *gutB* loci of natural isolates. *Mol Biol Evol* **9:** 654–665.

Harosh, I., Naumovski, L., and Friedberg, E.C. (1989) Purification and characterisation of Rad3 ATPase/DNA helicase from *Saccharomyces cerevisiae*. *J Biol Chem* **264:** 20532–20539.

Higgins, D.G., and Sharp, P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73:** 237–244.

Hughes, S.G. (1977) Studies of plasmid encoded restriction and modification systems. Ph.D. Thesis, University of Edinburgh, Edinburgh, Scotland.

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution.* Cambridge, England: Cambridge University Press.

Lederberg, S. (1973) cited by Radding, C.M., in Molecular mechanisms of genetic recombination. *Annu Rev Genet* **7:** 87–111.

Levin, B. (1986) Restriction–modification immunity and the maintenance of genetic diversity in bacterial populations. In *Evolutionary Processes and Theory.* Karlin, S., and Nero, E. (eds). New York: Academic Press, pp. 669–688.

Li, W.-H., Wu, C.-I., and Luo, C.-C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol* **2:** 150–174.

Linder, P., Lasko, P.F., Ashburner, M., Leroy, P., Nielsen, P.J., Nishi, K., Schnier, J., and Slonimski, P.P. (1989) Birth of the D-E-A-D box. *Nature* **337:** 121–122.

Loenen, W.A.M., Daniel, A.S., Braymer, H.D., and Murray, N.E. (1987) Organization and sequence of the *hsd* genes of *Escherichia coli* K-12. *J Mol Biol* **198:** 159–170.

Meisel, A., Bickle, T.A., Krueger, D.H., and Schroeder, C. (1992) Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage. *Nature* **355:** 467–469.

Midgley, C.A., and Murray, N.E. (1985) T4 polynucleotide kinase; cloning of the gene (*pse*T) and amplification of its product. *EMBO J* **4:** 2695–2703.

Nagaraja, V., Shepherd, J.C.W., and Bickle, T.A. (1985) A hybrid recognition sequence in a recombinant restriction enzyme and the evolution of DNA sequence specificity. *Nature* **316:** 371–372.

Nelson, K., Whittam, T.S., and Selander, R.K. (1991) Nucleotide sequence polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (*gapA*) in natural populations of *Salmonella* and *Escherichia coli*. *Proc Natl Acad Sci USA* **88:** 6667–6671.

Price, C., and Bickle, T.A. (1986) A possible role for restriction in bacterial evolution. *Micro Sci* **3:** 296–299.

Price, C., Ligner, J., Bickle, T.A., Firman, K., and Glover, S.W. (1989) Basis for changes in DNA recognition by *Eco*R124 and *Eco*R124/3 type I DNA restriction and modification enzymes. *J Mol Biol* **205:** 115–125.

Saitou, N., and Nei, M. (1987) The Neighbor–Joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406–425.

Sanger, F., Nicklem, S., and Coulsen, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74:** 5463–5467.

Schmid, S.R., and Linder, P. (1992) D-E-A-D protein family of putative RNA helicases. *Mol Microbiol* **6:** 283–292.

Sharp, P.M. (1991) Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: Codon usage, map position, and concerted evolution. *J Mol Evol* **33:** 23–33.

Sharp, P.M., Kelleher, J.E., Daniel, A.S., Cowan, G.M., and Murray, N.E. (1992) Roles of selection and recombination in the evolution of type I restriction–modification systems in enterobacteria. *Proc Natl Acad Sci USA* **89:** 9836–9840.

Shine, J., and Dalgarno, L. (1974) The 3-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci* **71:** 1342–1346.

Staden, R. (1982) Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. *Nucl Acids Res* **10:** 4731–4751.

Strauss, E.J., and Guthrie, C. (1991) A cold-sensitive mRNA splicing mutant is a member of the RNA helicase gene family. *Genes Devel* **5:** 629–641.

Studier, F.W., and Bandyopadhyay, P.K. (1988) Model for how DNA restriction enzymes select cleavage sites in DNA. *Proc Natl Acad Sci USA* **85:** 4677–4681.

Sung, P., Higgins, D., Prakash, L., and Prakash, S. (1988) Mutation of lysine-48 to arginine in the yeast RAD3 protein abolishes its ATPase and DNA helicase activities but not the ability to bind ATP. *EMBO J* **7:** 3263–3269.

Suri, B., and Bickle, T.A. (1985) *Eco*A: the first member of a new family of type I restriction and modification enzymes. *J Mol Biol* **186:** 77–85.

Tabor, S., and Richardson, C.C. (1987) DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proc Natl Acad Sci USA* **84:** 4767–4771.

Walker, J.E., Saraste, M., Runswick, M.J., and Gay, N.J. (1982) Distantly related sequences in the α- and β-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J* **1:** 945–951.

Weber, C.A., Salazar, E.P., Stewart, S.A., and Thompson, L.H. (1990) ERCC2: cDNA cloning and molecular characterization of a human nucleotide excision repair gene with high homology to yeast RAD3. *EMBO J* **9:** 1437–1447.

Yuan, R., Hamilton, D.L., and Burckhardt, J. (1980) DNA translocation by the restriction enzyme from *E. coli* K. *Cell* **20:** 237–244.