Recognition Domains of Type I Restriction Enzymes

Alexander A. F. Gann

A thesis presented for the degree

of Ph.D.

Department of Molecular Biology

University of Edinburgh

December 1988

Declaration

This thesis and the work presented within it are, unless otherwise stated, my own. Many of the ideas and approaches used were devised in discussion with my supervisor, Professor Noreen Murray.

Alexander A. F. Gann December 1988

Acknowledgements

I would like to thank Noreen for many long discussions, a number of which were vital in guiding me through this work, her teaching and her kindness. I am also very grateful to Anne who has been my constant and eternally patient companion in the lab. My thanks also to: other members of the department for advice and assistance, particularly John Collins and Andrew Coulson for computing; Karen Chapman for advice on site directed mutagenesis; Graham Brown for photograph; Alan Colson, Arthur Robinson and Diana Fawcett for M13 clones; Janine for her initials; my sister Sarah for typing this when she had better things to do, my brother Simon for providing the facilities, and my sister Chog, for proof-reading on a Saturday night (serious sacrifice); Janet Panther for finishing off the typing.

Most especially I must thank Heather, Julia, Simon, Anne, Lisa, Emma, Carla and Gareth,

Abstract

Type I restriction and modification enzymes recognize asymmetric, bipartite target sequences, the specificity of which is dictated by a single subunit encoded by the *hsdS* gene. Within the K-family, the *S* genes of members with different specificities have been sequenced (Gough and Murray, 1983; Gann *et al* 1987). Comparisons of these reveal two large variable regions, each of ~450 base pairs, separated by a highly conserved region of ~100 base pairs. Recombination between the central conserved regions of two *S* genes, those of *Sty*SP and *Sty*SB, has produced a new *S* gene (*Sty*SQ) encoding a functional polypeptide that confers a novel, hybrid specificity (Fuller-Pace *et al*, 1984; Nagaraja *et al*, 1985).

In this thesis I describe the formation of a second recombinant *S* gene, *Sty*SJ, which is of reciprocal structure to *Sty*SQ. *Sty*SJ recognizes a target sequence predicted by a model wherein each S polypeptide contains two structurally independent DNA recognition domains which act together in defining an enzyme's target sequence. Site directed mutagenesis was then used to demonstrate that the variable N-terminal 150 amino acids of an S polypeptide alone constitute one DNA recognition domain. Two S polypeptides, each deleted for a single recognition domain were also produced. Though showing no enzymatic activity *in vivo*, these truncated polypeptides were capable of inhibiting the activities of complete restriction and modification enzymes from their own family, but not from another. This is interpreted as being due to the truncated S polypeptides binding other enzyme subunits, thereby disrupting the formation of functional restriction complexes.

Abbreviations

AdoMet	S-adenosyl methionine
AMPS	ammonium persulphate
bp	base pair(s)
dNTP	deoxynucleoside triphosphate
ddNTP	dideoxynucleoside 5'-triphosphate
DTT	dithiothreitol
EDTA	diaminoethanetetra acetic acid
e.o.p.	efficiency of plating
gn	standard acceleration due to gravity
H-bonds	hydrogen bonds
hsd	host specificity determinant
IPTG	isopropyl-β-D-thiogalactoside
kb	kilobase
m.o.i.	multiplicity of infection
Mr	molecular weight
$O.D._{650}$	optical density at 650nm
PEG	polyethylene glycol
p.f.u.	plaque forming units
RF	replicative form
SDS	sodium dodecyl sulphate
SSC	standard saline citrate
TEMED	N, N, N', N'-tetramethyl ethylene diamine
X gal	5-bromo-4-chloro-3-indolyl-β-D-galactoside
	deletion
ts	temperature sensitive

Amino Acids

Name	Three letter code	Single letter code
alanine	Ala	A
arginine	Arg	R
asparagine	Asn	N
aspartate	Asp	D
cysteine	Cys	C
glutamate	Glu	E
glutamine	Gln	Q
glycine	Gly	G
histidine	His	H
isoleucine	Ile	I
leucine	Leu	L
lysine	Lys	K
methionine	Met	M
phenylalanine	Phe	F
proline	Pro	P
serine	Ser	S
threonine	Thr	T
tryptophan	Trp	W
tyrosine	Tyr	Y
valine	Val	V

Contents

CHAPTER 1 : PROTEIN-DNA RECOGNITION

1) General Introduction

The ability of proteins to recognize specific nucleotide sequences within DNA molecules is essential for achieving regulation of gene expression. Initiation of transcription involves recognition of promoters by RNA polymerase (von Hippel et al, 1984; Helmann and Chamberlin, 1988). This is itself often regulated by the binding of appropriate repressor or activator proteins to other nearby sequences (Ptashne, 1986a). Sequence specific protein-DNA interactions also play central roles in site-specific recombination, restriction and modification. These processes influence the physical arrangement of genes within chromosomes and their transfer between different genomes; in effect, further levels of genetic regulation. Site-specific recombination produces precise insertions, deletions and rearrangements of DNA and depends on sequence specific recognition of the substrate by appropriate proteins (Weisberg and Landy, 1983). Restriction enzymes recognize and destroy foreign DNA entering a cell (Bickle, 1987). In this case the requirement for sequence specificity is to allow protection of the cell's own genome by a modification enzyme of identical specificity. The consequent methylation alters physical characteristics of the nucleotide sequence such that it is no longer a substrate for the restriction enzyme. Methylation induced changes of this kind may also influence the binding of regulatory proteins and hence affect gene expression (Bird, 1986).

There are inevitably different ways in which proteins interact with their target sequences, presumably influenced by functional constraints. Repressors and activators often simply bind to DNA in order to attain a defined position with respect to the other components of the transcription machinery. The DNA binding facilities of such proteins appear often to be structurally independent of regions involved in transcription regulation (Ptashne, 1986a and 1988). Other proteins, such as restriction enzymes, do not simply bind to their target sequences, but act on them. This intimate association of recognition and function may entail differences in the way they determine specificity (e.g. McClarin et al, 1986; Echols, 1986).

Before any structural information was available, it was predicted that different DNA sequences could be distinguished from one another by the pattern of potential hydrogen bonds (H-bonds) they could form (Seaman et al, 1976). Functional groups on the base pairs (bp) within a DNA double helix protrude into the major and minor grooves where they are accessible to the amino acid side chains on an appropriately positioned protein surface. The specific sequence favoured will be that which possesses a pattern of H-bond donors and acceptors exactly complementary to that provided by the protein. The importance of such H-bonding in protein-DNA recognition has subsequently been confirmed by physical and genetic analysis of a number of DNA binding proteins (Pabo and Sauer, 1984; McClarin et al, 1986). However, other features, such as sequence specific

deformations of the sugar-phosphate backbone may also contribute to the recognition process (Dickerson, 1983a). In particular, the ability of a given stretch of DNA to mould itself in complex with a protein so as to optimize contacts between the two, is, to an extent, sequence specific, and therefore an important characteristic by which cognate and non-cognate sequences can be distinguished (e.g. McClarin et al, 1986; Otwinowski et al, 1988).

The level of specificity required in a protein-DNA interaction depends in part on the function of the protein involved. Repressors and activators will probably not cause too much damage in a cell if they occasionally interact with the wrong DNA sequence, while, in the case of a restriction enzyme, such a mistake could prove fatal. Initiation of DNA replication is an event that requires extremely precise regulation, both spatially and temporally (Kornberg, 1982; Echols, 1986).

There are two levels at which specificity is achieved. Firstly, there is the inherent affinity of a protein for its target as compared to non-cognate sequences. Secondly, the invocation of some sort of co-operativity. This can be in the form of co-operative binding whereby proteins interacting simultaneously with their DNA target sequences and each other serve to stabilize binding at, and hence increase affinity for, their correct sites (e.g. Ackers et al, 1982; Ptashne, 1986a; Echols, 1986). Alternatively there can be a link between DNA binding and enzymatic function: a protein may only be

activated when bound to the correct nucleotide sequence (e.g. McClarin et al, 1986).

2) The Physical Basis for Specificity in Protein-DNA Interactions

A) Transcription Regulators : the helix-turn-helix

Many regulatory proteins bind as dimers to DNA sequences that show twofold symmetry. Each monomer of the protein interacts in an equivalent way with one half of the DNA target site (Pabo and Sauer, 1984; Hollis et al, 1988). Solution of the structure of one such complex, the DNA binding domain of bacteriophage 434 repressor bound to its operator (Anderson et al, 1985 and 1987; Aggarwal et al, 1988), revealed the details of how specificity in DNA recognition is achieved by this protein and, by analogy, many others that appear to employ a conserved region of secondary structure for this purpose (Sauer et al, 1982; Pabo and Sauer, 1984). Structures of the DNA binding domain of bacteriophage λ repressor (Pabo and Lewis, 1982), λ Cro (Anderson et al, 1981; Ohlendorf et al, 1982), the transcription activator protein (CAP) (McKay and Steitz, 1981; McKay et al, 1982) and trp repressor (Schevitz et al, 1985) had all been solved in the absence of their operator DNAs. Model building had suggested that all four could recognize these targets by inserting one α-helix (the recognition helix) from each monomer into successive major grooves along one face of the DNA. A second α-helix, which is separated from the recognition helix by a tight turn, was proposed to sit across

the major groove near the sugar-phosphate backbone. Contacts between this helix and the DNA backbone, while being important to binding and positioning of the recognition helix, were not thought to contribute significantly to specificity. Specific base pair contacts were proposed to be made by the side chains of amino acid residues on the outer (solvent exposed) face of the recognition helix. This DNA binding structure has come to be known as a helix-turn-helix motif (see Figure 1 for general situation) (above references of individual proteins; Ohlendorf et al, 1983; Pabo and Sauer, 1984). The operator DNA was thought to remain essentially B-form when bound by the proteins, though in some cases it might be bent towards the protein (Ohlendorf et al, 1982).

An increasing amount of other evidence, both genetic and biochemical, supports this model for how these proteins, and others, recognize their operators. For λ repressor, a number of mutations that either decrease or increase operator binding have been found to cluster in the helix-turn-helix region (Nelson et al, 1983; Hecht et al, 1983; Nelson and Sauer, 1985; Hecht and Sauer, 1985). Similar mutations have been described for trp repressor (Kelley and Yanofsky, 1985). Even more sophisticated approaches using mutant repressors, mutant operators, and combinations of both, examine specific amino acid base pair interactions. These have been reported for λ repressor (Hochschild and Ptashne, 1986a; Hochschild et al, 1986; Benson et al, 1988), λ Cro (Eisenbeis et al, 1985), 434 repressor (Wharton and Ptashne, 1987), trp repressor (Bass et al, 1987; Bass et al, 1988) and CAP (Ebright et al, 1984). All

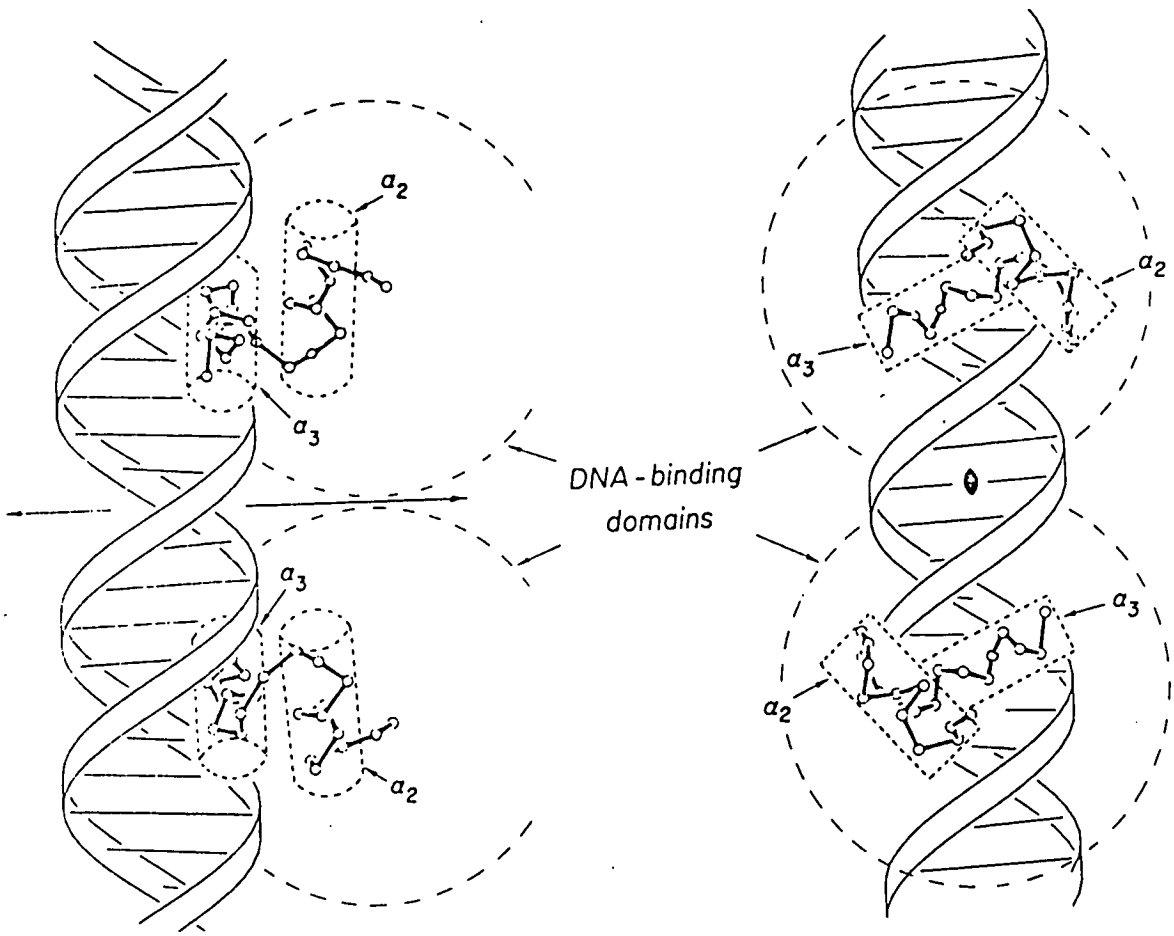


Fig. 1: Schematic diagram of the interaction between a helix-turn-helix structure and DNA. The recognition helix (α3) from each repressor dimer is inserted into succcessive major grooves along one face of the DNA double helix, while α-helix 2 lies across the major groove, contacting the sugar phosphate backbone. Taken from Pabo and Sauer, 1984.

of these studies imply that the amino acids on the solvent side of the recognition helix are major determinants of DNA recognition. Altering these can abolish, increase or change the specificity of DNA binding.

Many other proteins whose structures have not been solved also appear to use the helix-turn-helix DNA binding domain. This is based on sequence similarities to the proteins described above (Sauer et al, 1982; Pabo and Sauer, 1984) and, in some cases, is supported by genetic analysis, e.g. lac repressor (Lehming et al, 1987), CII transcription activator (Ho et al, 1988), phage P22 repressor (Wharton and Ptashne, 1985) and fnr activator (Spiro and Guest, 1987). Even RNA polymerase appears to use helix-turn-helix domains in recognition of promoters. These are found in the sigma factors, interchangeable alternatives of which confer recognition of different promoter sequences (Helmann and Chamberlin, 1988).

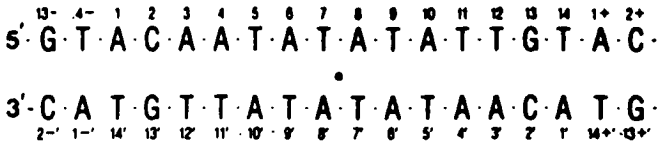
Chemical modification and protection experiments have revealed positions on the DNA which are in close proximity to the bound proteins. These data fit well with the proposed structures from the crystallographic studies (e.g. 434 repressor: Bushman et al, 1985). Clearly evident from these experiments was the symmetry of the protein DNA interaction around the centre of the operator (Johnson et al, 1978; Johnson et al, 1979; Humayun et al, 1977). More recently, a number of sophisticated approaches have been developed that reveal more details of backbone and base contacts made by various proteins

(e.g.λrepressor: Tullius and Dombroski, 1986; Brunelle and Schleif, 1987).

The structure of the DNA recognition domain of 434 repressor bound to a synthetic 14bp operator oligonucleotide demonstrated directly the details of this mechanism of DNA recognition (Anderson et al, 1985 and 1987). The protein domain consists of the N-terminal 69 amino acids of 434 repressor. The recognition helix comprises residues 28 - 36, three of which (28, 29 and 33) are glutamines (Gln) that project into the major groove of the DNA and form specific base contacts within the operator (Figure 2). The first forms two H-bonds with the adenine at position 1 of the operator (and the corresponding residue in the recognition helix of the other repressor monomer makes the equivalent contact at position 14 of the symmetric operator. See Figure 2). These are between N7 of the base and N_ϵ of the Gln, and N6 and O_ϵ. Gln 29 is in Van der Waals contact with the 5-methyl group of the thymine (T) at position 3 (and 12), and its N_ϵ can H-bond to O_6 (and perhaps N7) of the guanine (G) at position 2 (13). Gln 33 projects towards the thymine and adenine (A) at operator positions 4 and 5 (11 and 10) respectively. H-bonds occur between N_ϵ of Gln 33 and O4 of thymine, and possibly O_ϵ and N6 of adenine.

The cocrystal reveals, however, that these contacts alone do not constitute the entire mechanism of determining specificity. The affinity of the 434 operator for repressor is greatly influenced by the sequence of non-contacted bases near

a.



b.

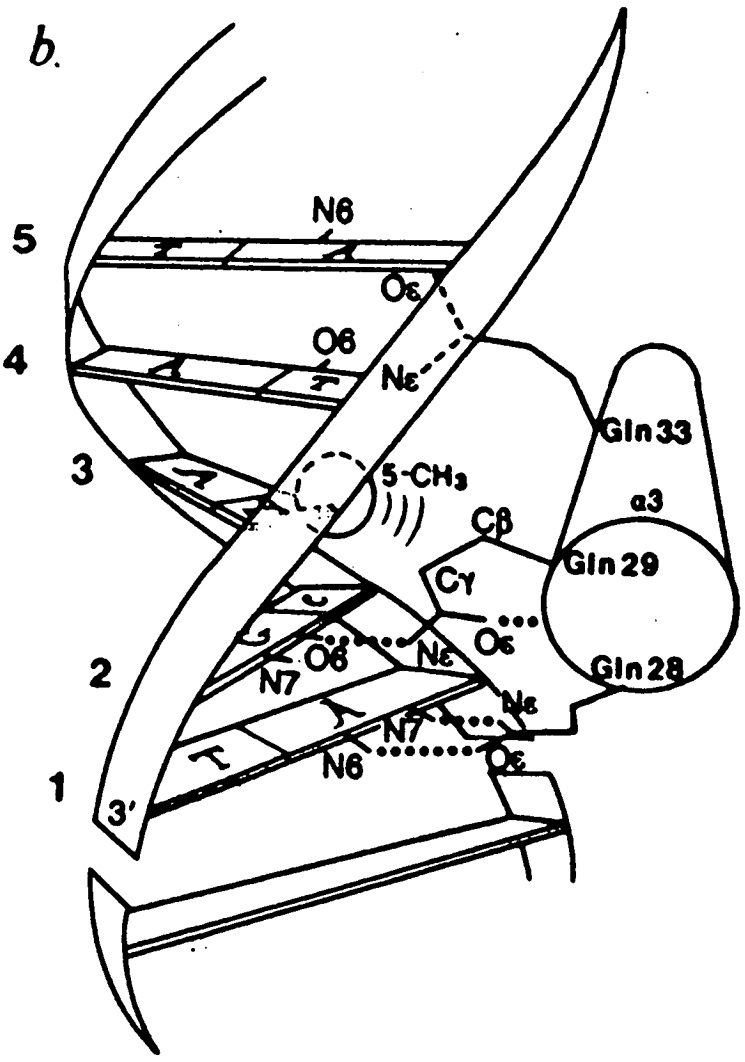


Fig. 2.a). Sequence of the double stranded 14bp oligonucleotide co-crystalized with the phage 434 repressor. The numbering scheme is that used in the text.

b). Schematic representation of α-helix 3 (recognition helix) of 434 repressor and bp 1-5 of the operator, viewed from the N-terminus of α3. Gln28, Gln29 and Gln33 are shown, with H-bonds suggested by the current model indicated by dotted lines. Functional groups on bases and on Gln33 that might participate in additional H-bonds are also shown. Taken from Anderson et al, 1987.

the operator's centre (Koudelka et al, 1987). To achieve an optimum alignment of the two operator half sites (residues 1 - 5 and 10 - 14) and the two monomers of the repressor dimer, the DNA must be overwound in the central region (residues 6 - 9). The nature of these residues influences the ease with which the overwinding can take place, and hence the affinity (over a 50x range) of repressor binding (Koudelka et al, 1988). Operators with A-T base pairs in these positions bind repressor more tightly than those with G-C base pairs. This is because, when the operator is overtwisted on interaction with the protein, runs of A-T base pairs are able to form bifurcated H-bonds (Nelson et al, 1987) which stabilize the overtwisted state. G-C base pairs merely retain the Watson-Crick pairing and are therefore less stable when overwound (Koudelka et al, 1988; Aggarwal et al, 1988).

In addition, it would be wrong to describe the contacts made between the repressor and the DNA phosphate backbone as non-specific. In the cocrystal it was seen that, though essentially acting to clamp the recognition helix in place, these contacts can only be made correctly if the backbone is able to adopt a precise structure within the complex. This is at least partially dependent on the base pair sequence within the operator (Anderson et al, 1987). This situation is even more apparent in a higher resolution structure of 434 repressor bound to a 20 bp oligonucleotide containing the same operator sequence (Aggarwal et al, 1988).

Nevertheless, that the contacts made by residues on the solvent (outside) face of the recognition helix could alone dictate the specificity of repressor binding was demonstrated in an experiment by Wharton and Ptashne (1985). Regions thought to represent the helix-turn-helix structures in the amino acid sequences of the repressors from 434 and related Salmonella phage P22 were aligned. Five residues were identified that were predicted to lie on the outer face of the recognition helices and which were different in the two repressors. These were changed in 434 repressor to the corresponding amino acids from P22 repressor. Residues on the inside face were not altered as they were thought important in correct folding of the helix against the main body of the protein. The resulting hybrid repressor, 434R [α3 P22] was shown, both in vivo and in vitro, to bind specifically to P22 operators and not those of 434. Changing just three of the five residues back to those found in 434 repressor returned the specificity to that of 434. These were at positions 27, 28 and 29, the latter two being two of the three glutamines identified as making base pair contacts in the crystal structure described above. The third glutamine, that at position 33, is common to both 434 and P22 repressors, and so is present in all the hybrids. Recently it has been shown that a heterodimer, consisting of a monomer of 434 repressor and a monomer of 434 R [α3 P22] specifically binds to a hybrid operator with one half site from that of 434 and one from P22 (Hollis et al, 1988).

This helix swap experiment demonstrates that both repressors use the outside surface of their recognition helices

as the major, if not sole, determinant of binding specificity. It also shows that both repressors insert these helices into the major groove in very similar ways, thereby allowing the same base pair contacts to be made by identical amino acids in equivalent positions in the two helices.

Whether this is generally the case is an important question in terms of whether a simple "recognition code" exists, whereby particular nucleotide sequences can be expected to be recognized by certain amino acid sequences. The early models for λ repressor (Pabo and Lewis, 1982), λ Cro (Anderson et al, 1981) and CAP (McKay et al, 1982) proposed that their recognition helices are inserted into the major groove in rather different ways. More recently, however, a very detailed mutational analysis (Hochschild and Ptashne, 1986a; Hochschild et al, 1986) has revealed that individual amino acids from the recognition helices of λ repressor and Cro can function in the context of either helix; they can be exchanged between them, not only retaining their own original base pair contacts, but also not disrupting those of the native residues around them. It therefore seems that these recognition helices must be inserted into the major groove in very similar orientations.

λ repressor and Cro recognize the same six operators within the phage genome, but do so with different orders of affinity (Johnson et al, 1979). Though the sequences of the 12 operator half sites are similar, the only two invariant positions are 2, which is always A:T, and 4, which is C:G (Gussin et al, 1983). The recognition helices of the two

repressors also have certain residues in common. Each has five solvent exposed residues, of which the first two are glutamine and serine; the other three differ in the two proteins. The mutational study shows that the two conserved amino acids contact the two invariant operator positions. The different orders of affinity shown by Cro and repressor are, to a great extent, a consequence of the contacts made between the other amino acids in each recognition helix and the less conserved base pairs in the operators (Hochschild and Ptashne, 1986a; Hochschild et al, 1986). Very recently the structure of the DNA binding domain of λ repressor bound to its operator has been solved (Jordan and Pabo, 1988). The structure confirms the description of specific amino acid base pair contacts suggested by the genetic studies.

Are all recognition helices inserted into the major groove in identical ways? Probably not. One subtle but relevant exception is the repressor of phage 16-3 from the nitrogen fixing bacterium Rhizobium meliloti (Dallmann et al, 1987). This recognizes the same operators as 434 repressor. The two proteins show very little sequence similarity except in the region of the helix-turn-helix. Here Gln 28 and Gln 29 in 434 repressor are conserved in that of 16-3. The third Gln (33), however, is replaced by asparagine (Asn), which, though chemically similar, has a shorter side chain. For the Asn to make equivalent operator contacts, the recognition helix would have to be presented at a slightly different angle, and perhaps some change in operator conformation may be required.

Recently, the structure of the phage 434 Cro protein bound to the same 14mer as in the 434 repressor complex described above (Anderson et al, 1987) has been solved (Wolberger et al, 1988). As with phage λ, the repressor and Cro proteins of 434 bind to the same six similar operators and do so with different orders of affinity (Johnson et al, 1981; Wharton et al, 1984). The sequences (Sauer et al, 1982) and three dimensional structures (Mondragon et al, JMB in press) of the 434 Cro monomer and a single 434 repressor DNA binding domain (R1-69) are very similar. However, their respective complexes with the operator DNA are rather different. The two protein monomers in each structure have different orientations relative to one another. This is caused by variations in their protein-protein contacts at the dimer interface where side chains of two residues in Cro (Phe and Tyr) are bulkier than those of the corresponding residues in repressor (Pro and Leu) (Anderson et al, 1987; Wolberger et al, 1988).

Though in complex with 434 repressor the DNA is bent, overwound near its centre and underwound at its ends, when in complex with Cro the same DNA is straight and uniformly overwound (Wolberger et al, 1988). These differences in conformation are imposed on the DNA by the different ways in which the two proteins interact with the operator, these being a result of the relative orientations of the two monomers within the two repressors: the Cro dimer demands straight DNA to maintain equivalent interactions between the helix-turn-helix from each monomer and the two operator half sites; for repressor bent DNA is necessary. Overwinding in the centre of

the operator is required by both proteins (Koudelka et al, 1987).

A number of observations demonstrate that the amino acid base pair contacts made between 434 Cro and repressor and the operators are not exactly equivalent. Both proteins bind to operators with the sequence ACAA at positions 1 - 4. In one operator, the sequence is altered to ACAG in one half site with the result that repressor now binds less well, while Cro is relatively unaffected (R.P. Wharton, Ph.D. Thesis, Harvard University, 1985; referenced in Aggarwal et al, 1988). In the repressor-operator complex described above we see that Gln 33 forms an H-bond with the T at position 4 (Anderson et al, 1987). The equivalent amino acid residue in Cro is a Leu. However, the simple change of Gln 33 to Leu in the repressor does not eliminate discrimination between A-T and G-C at position 4 of the operator (G. Koudelka and R.P. Wharton, unpublished results); i.e. it does not mimic the situation with Cro, presumably because amino acid residues at equivalent positions in the recognition helices of these two proteins are not in identical positions with respect to the base pairs of the operator. This is further demonstrated by the finding that even where residues conserved in both proteins (e.g. Gln 28 and Gln 29) are used to contact identical bases in the operator (A and C: see above and Figure 2) the interaction is not necessarily equivalent; changing Gln 28 to Ala confers on the repressor the ability to recognize an operator with an A to T change at position 1 (Wharton and Ptashne, 1987). The same amino acid substitution in the Cro protein does not enable it

to bind to the mutant operator (Wharton, unpublished results referenced in Wolberger *et al*, 1988).

A more extreme example of two repressors using their recognition helices in rather different ways is suggested by experiments carried out with mutant *lac* repressors and operators (Lehming *et al*, 1987). Boelens *et al* (1987) reached similar conclusions based on NMR studies of *lac* repressors. The recognition helix of the *lac* repressor appears to be inserted into the major groove the opposite way round to that of λ repressor. This model has residues from the N to C terminus of the *lac* repressor's recognition helix contacting bases from the centre of the operator outwards (Lehming *et al*, 1987). For λ repressor, it is the C terminus that is nearer the centre, with the N-terminal residues contacting the outer bases (Pabo and Lewis, 1982; Hochschild and Ptashne, 1986a; Hochschild *et al*, 1986).

This same study (Lehming *et al*, 1987) suggested that the *gal* and *deo* repressors from *E.coli* use the same recognition helix presentation as the *lac* repressor. It therefore appears that there are at least two different orientations in which the recognition helix can be presented; each has been adopted by a number of different repressors (with subtle variations - e.g. 434 Cro and repressor) to recognize a variety of DNA sequences.

One important feature has been employed by all of the proteins discussed so far: specificity is determined, at least to a great extent, by the interactions between amino acid side

chains on the solvent face of the recognition helix and the bases in the operator. Under these circumstances a "recognition code" is still conceptually possible, even though the variation in recognition helix presentation will ensure such a code is at least degenerate. However, solution of the structure of a trp repressor-operator complex has demonstrated that a specific DNA sequence can be recognized in a fundamentally different way, even while still mediated through a helix-turn-helix domain.

The trp repressor structure had already been solved in the absence of DNA (Schevitz et al, 1985) revealing a helix-turn-helix domain. In the cocrystal, however, it was seen that none of the specificity of its interaction with DNA is due to direct H-bonds to functional groups of the operator bases (Otwinowski et al, 1988). There are 14 direct H-bonds between each repressor monomer and operator half site, all but two of which involve the unesterified oxygens of six phosphate groups in the DNA backbone (Figure 3). Four of these phosphates accept more than one H-bond, and in one case four, so their precise location is probed for very accurately by the protein. A single possible direct H-bond between the repressor and an operator base pair has been shown, by mutation of this position in the operator, to be unimportant in defining specificity (Bass et al, 1987). Some bases, shown by mutation to be important in recognition, are contacted by H-bonds, but these are mediated by water molecules between the amino acid residues and the bases.

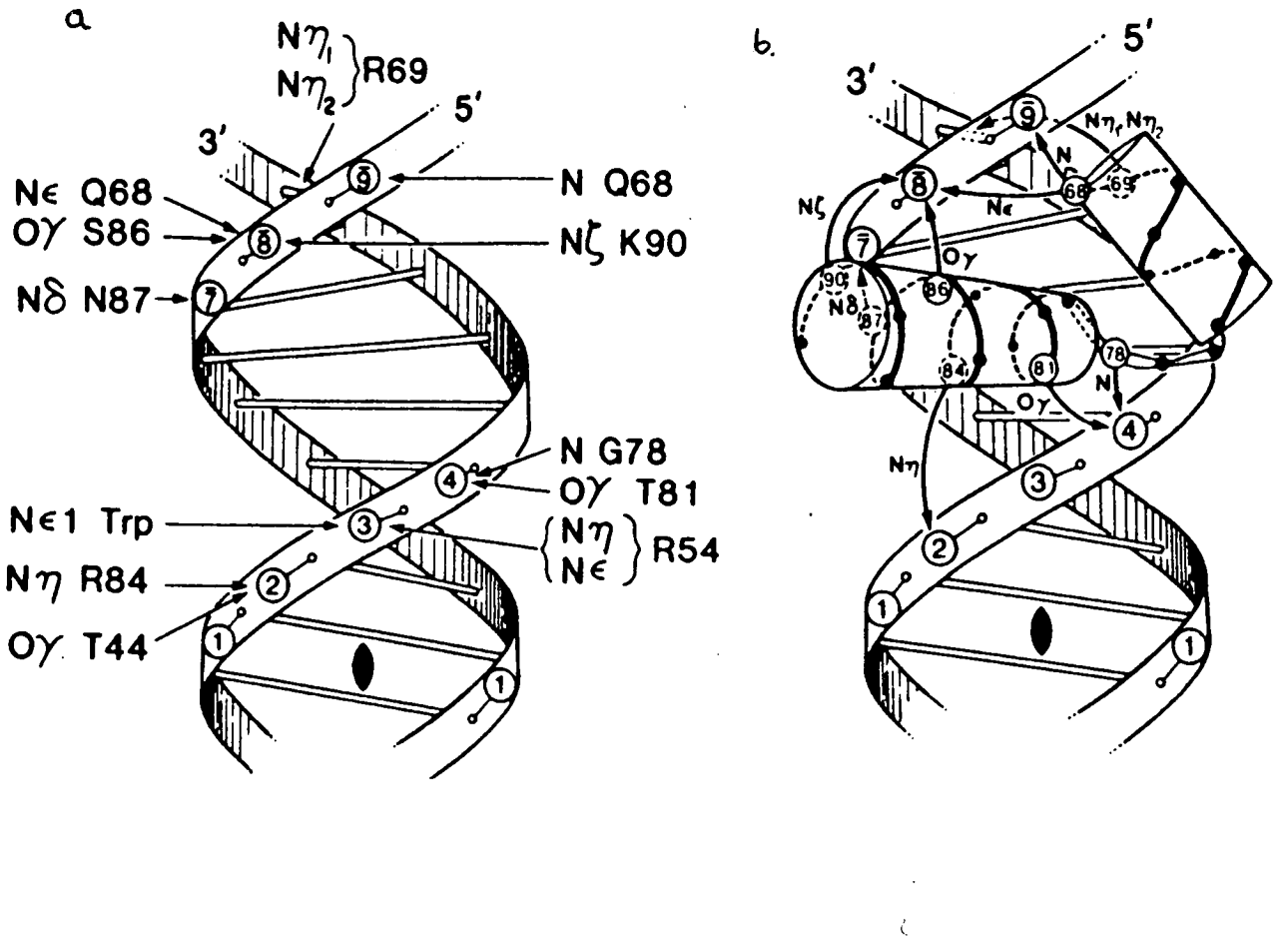


Fig. 3. Residues involved in direct contacts between trp repressor and operator. (a) Schematic view of half of the trp operator as viewed by the repressor showing the phosphates (circled numbers)and repressor functional groups (arrows) that form direct H-bonds. (b) Schematic diagram of amino acid positions in the helix-turn-helix that make direct H-bonds to the operator. The figure emphasizes that the recognition helix "points" into the major groove of the DNA, with the α-helix almost perpendicular to the DNA axis. The recognition helix does not "lie" in the major groove as shown in Figs. 1 and 2. Taken from Otwinowski et al, 1988.

The operator is segmentally bent towards the protein, thereby allowing optimum repressor-operator contacts. From model building it is clear that straight DNA would be unable to make more than a few of the twenty-four direct repressor-DNA phosphate contacts formed by the appropriately bent DNA. A number of bases known to be vital sequence determinants of repressor affinity (Bass et al, 1987; Bass et al, 1988) are located at the positions of operator bends, and are not contacted by the protein, even indirectly. It is the nature of these bases that enables the bending, and hence the resulting repressor-operator contacts, to occur. The very precise conformation the DNA needs to adopt in complex with the repressor is a sequence dependent characteristic of the trp operator and it is this that explains the specificity of the interaction (Otwinowski et al, 1988). This situation is significantly different from models of other complexes described above (Pabo and Sauer, 1984) and clearly demonstrates that the pattern of potential H-bonding presented by the base pairs is not the only characteristic of DNA that can be used by proteins to distinguish between different nucleotide sequences.

Though all use a helix-turn-helix, some of the proteins described here show additional features of their interaction with DNA which are less general. For example, not all the sequence specific interactions made between λ repressor and its operators are by residues in the recognition helix. The N-terminal 5 amino acids of this protein form an arm structure that reaches around the DNA and makes contacts in the major groove on the opposite face (Pabo et al, 1982; Eliason et al,

1985; Jordan and Pabo, 1988). These contribute to repressor's ability to distinguish between its various operators (Hochschild et al, 1986). λ Cro and trp repressors also have arms, though in these cases they appear to increase binding strength but not specificity (Caruthers et al, 1986; Otwinowski et al, 1988).

DNA binding by some proteins involved in gene regulation is itself modulated by their interaction with other molecules. trp repressor and CAP only bind their target sequences when complexed with L-tryptophan and cAMP respectively (Joachimiak et al, 1983). This is essential for their physiological roles: the tryptophan synthesizing enzymes, encoded by the operon whose expression is controlled by trp repressor, are only required in the absence of tryptophan (Yanofsky and Crawford, 1987); cAMP levels increase in a cell when glucose is in short supply, thus allowing CAP to bind to its target sites on the chromosome where it stimulates transcription of a number of operons encoding enzymes able to metabolize alternative substrates (Beckwith, 1987). In both cases it appears that ligand binding acts to induce and stabilize the correct orientation of the recognition helix, thereby enabling DNA binding to occur (Schevitz et al, 1985; Zhang et al, 1987). In the case of the trp repressor it also appears that the tryptophan molecule bound by the repressor actually interacts directly with the DNA; the nitrogen of the indole ring forms an H-bond with an operator phosphate (Otwinowski et al, 1988).

It therefore appears that, though a conserved DNA binding structure, the helix-turn-helix is not always employed in an identical way. The angle at which the recognition helix is inserted into the major groove varies and is sometimes influenced by binding of other molecules. Also, the sequence dependent characteristic of the DNA recognized is not always the same one. As a result there seems little chance of a simple, universal, recognition code. Experiments suggest that groups of proteins may operate in superimposable ways (e.g. 434 and P22 repressors (Wharton and Ptashne, 1985); *lac*, *deo* and *gal* repressors (Lehming *et al*, 1987)) but that each set is distinct from the others. The very different recognition mechanism employed by the *trp* repressor (Otwinowski *et al*, 1988) makes a general recognition code impossible to imagine. Is it reasonable to expect that direct contact of a base pair or, alternatively, recognition of its indirect effect on local DNA structure, should always be carried out by equivalent amino acid residue in a protein? In fact, the different types of variation in recognition mechanism may allow for a far greater range of possible specificities than any single model could.

B) Co-operativity

The specificity of a repressor is a measure of the affinity it has for operator compared to non-operator sequences. In a cell, repressor is either free, bound to operator sites, or bound to non-specific DNA. The length of time the protein will spend bound to its operator is dependent on its affinity for the operator and on its concentration free

in the cell (i.e. the concentration of repressor available for binding). The favourable interactions made between the protein and its operator ensure that this sequence is greatly preferred over all others ($\sim 10^6$ fold for a typical prokaryotic repressor). However, in a cell, the number of non-operator sites is far greater than operator sites. For example, in E.coli there may be a single operator site for a given repressor. As every base of the cell's chromosome is the first position of a potential binding site, there are $\sim 10^7$ such sites in the cell, i.e. there are about 10^7 times as many non-operator as operator sites, for each of which the repressor may have an affinity 10^{-6} that of its affinity for the operator. Clearly most of the repressor will spend a majority of its time bound non-specifically to DNA (von Hippel and Berg, 1985).

It has in fact been calculated that in the type of situation outlined above, only about 1% of the total repressor in a cell will be available for operator binding at any given time (Linn and Riggs, 1975). As a consequence, only 90% operator occupancy will be achieved. This is unacceptably low for realistic regulation. A figure of 99.9% occupancy is required (see Ptashne et al, 1980; Ptashne, 1986b). How is this achieved?

Obviously increasing the total concentration of repressor in the cell would suffice, but a hundredfold increase would be necessary and this is an expensive solution.

Alternatively, the specificity of the repressor could be increased by introducing additional protein-DNA contacts. However, this could result in a kinetic problem: while increasing binding to operator compared to non-operator DNA, the absolute level of non-specific binding would also increase such that dissociation from non-operator sites would be slow, and thus location of the operator would take too long. Also, if the operator affinity gets too high, then once bound the repressor would never dissociate and hence sensible regulation becomes impossible. A mutant λ repressor known to have increased affinity for both operator and non-operator sites is unable to function *in vivo*, presumably for these reasons (Nelson and Sauer, 1985).

Co-operativity solves the problem for several repressors (Johnson *et al*, 1979; Ackers *et al*, 1981; Ptashne *et al*, 1980). This involves two proteins binding not only to their operators, but also to each other in such a way that the overall complex is more stable than any of the individual interactions alone. The reason this increases specificity is that there is a much greater chance of repressors binding simultaneously to adjacent operator than non-operator sites, due to their inherent affinity for the former.

Co-operativity in lambdoid repressor binding is demonstrably vital to activity *in vivo*: proteolytic cleavage removes a domain responsible for co-operative interactions from another which, while still capable of binding a single operator as tightly as the complete repressor, is unable to bind to two

co-operatively. At normal *in vivo* concentrations, this DNA binding domain alone produces no effective repression (Sauer *et al*, 1979).

More complex arrangements involving multiple protein-DNA and protein-protein interactions operate in other systems where very high specificity is required (Echols, 1986). Initiation of DNA replication in *E.coli* and λ, and site-specific recombination by λ, are well studied examples. *E.coli* DNA replication is initiated at a single location on the chromosome, and is done so only once per cell generation (Kornberg, 1982). λ integration occurs by recombination between a single position in the phage genome and a highly preferred one on the host chromosome (Weisberg and Landy, 1983). The basis for the exceptional precision of these protein-DNA interactions is not immediately apparent from the DNA binding properties of the individual proteins that direct them. *In vitro* reconstruction experiments have shown that large protein-DNA complexes are built up at the origin of replication in *E.coli* and λ (Dodson *et al*, 1985 and 1986). These localize the initiation point by forming a structure to which other proteins essential to replication (Kornberg, 1982) are gathered, and also possibly by producing structural change in the DNA favouring unwinding (Dodson *et al*, 1986). A similar situation has also been observed in the initiation of DNA replication in Simian Virus 40 (Dodson *et al*, 1987). The relevant point is that the specificity is enormous because only the correct stretch of DNA is capable of accommodating and

forming a complex with all the different proteins involved (Ecohls, 1986).

Regulation of gene expression in eukaryotes also probably requires the co-operative interaction of a number of proteins on DNA, both to increase precision and to allow for complex levels of regulation (Wasylyk, 1988). The SV40 early promoter is a well studied example where it is known that several proteins bind to upstream sequences, probably in some co-operative manner, to enable transcription to be regulated (Jones et al, 1988). For many eukaryotic genes, regulation of expression occurs at many levels, e.g. tissue, developmental stage or sex specific (Maniatis et al, 1987). Each relevant situation can be signified by the production, activation or inhibition of various regulatory proteins. Different combinations of such proteins bound to the regulatory sequences of a gene will produce the various patterns of expression appropriate in that cell at different times. Co-operativity is here used to increase regulatory flexibility as well as specificity.

C) Type II Restriction and Modification Enzymes

Restriction and modification (R-M) enzymes not only bind to specific nucleotide sequences, but act on the DNA. They therefore represent a system in which more complex recognition mechanisms than those described for the simple repressors may operate. Particularly interesting is the possibility of comparing the recognition processes employed by enzymes which,

while of identical specificity, act on their targets in very different ways. Type II R-M systems offer such a possibility in that each consists of two separate enzymes - one an endonuclease, that cuts DNA, the other a modification enzyme that methylates the same sequence. Furthermore, systems from different organisms but which have identical specificity have been identified. Comparisons of these, known as isoschizomers, may reveal something of the flexibility available in how a specific DNA sequence can be recognized as a substrate for a certain enzymatic activity.

The EcoRI system is a well studied type II restriction and modification (R-M) system (general details of R-M systems will be given in the next chapter). The active endonuclease is a homodimer of a 276 amino acid polypeptide. It binds tightly and specifically (Kd $\sim 10^{-11}$ M^{-1}) to the DNA sequence GAATTC in the absence of Mg^{2+} (Modrich, 1979). When Mg^{2+} is present, and the target sequence is unmodified, the enzyme hydrolyses the phosphodiester bond between the G and A residues resulting in a 5′ phosphate (Connolly et al, 1984).

The structure of the enzyme dimer bound to a tridecanucleotide containing its target sequence has been solved to 3 Å resolution (Frederick et al, 1984; McClarin et al, 1986). Features of the complex involved in recognition and cleavage of the DNA are revealed and are now described. Most of what follows, unless indicated otherwise, comes from the two papers indicated above.

The two subunits of the enzyme form a globular structure about 50 Å across into one side of which the DNA is embedded. The major groove is in intimate contact with the protein, while the minor groove is open to the solvent (Figure 4).

Each subunit of the enzyme is a single domain consisting of a five strand β-sheet surrounded on either side by α-helices (Figure 5). The first three strands of the sheet form an antiparallel unit which is associated with phosphodiester bond cleavage. The other three strands and two α-helices form a parallel motif involved in subunit-subunit interactions. Two other α-helices, called "inner" and "outer", are vital to DNA recognition. Each monomer also has an arm structure made up of the N-terminal 14 amino acids of each polypeptide and a β-hairpin situated between the fourth and fifth strands of the large β-sheet. These arms encircle the DNA and clamp it into place on the enzyme surface, forming contacts with the DNA backbone which are essential for catalytic activity and binding affinity, but not specificity (Jan-Jacobson et al, 1986).

The structure of the oligonucleotide used in the complex had previously been solved (Dickerson and Drew, 1981; Dickerson, 1983b), and so changes caused by the binding of the enzyme, which turn out to be quite dramatic, can easily be seen in the cocrystal (Frederic et al, 1984). Though retaining most of the structural features of double helical DNA, including Watson-Crick base pairing, the DNA is nevertheless kinked in the recognition complex (Frederic et al, 1984). The most striking departure from B-form DNA is in the centre of the

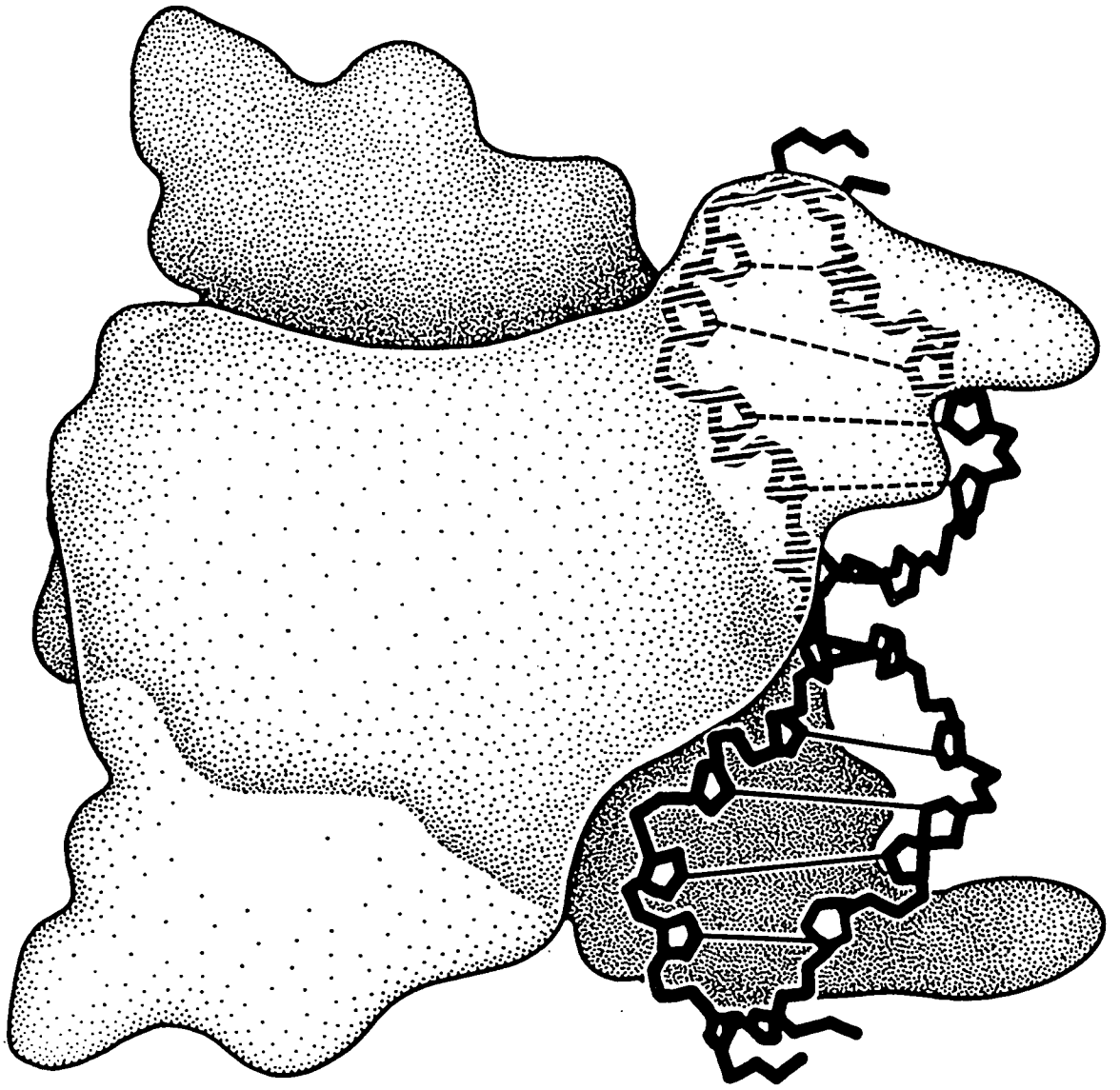


Fig. 4. Schematic diagram of the overall appearance of the EcoRI endonuclease bound to its target site as deduced frm the cocrystal. The double stranded DNA helix is embedded in one side of the EcoRI endonuclease dimer. Each protein monomer is indicated - one shaded dark and one light. The symmetry of the enzyme-DNA interaction is clear, as is the kinking of the DNA, the protein filled major groove, and the solvent exposed minor groove.

Fig. 5. Schematic diagram of one subunit of EcoRI endonuclease and both strands of the DNA in the complex. The arrows represent β strands, the coils represent α-helices, and the ribbons represent the DNA backbone. The α-helices in the foreground are the inner and outer recognition helices. They connect the third β strand to the fourth, and the fourth β strand to the fifth. The two helices also form the central interface with the other enzyme subunit. The amino terminus of the polypeptide chain is in the arm near the DNA. Taken from McClarin et al, 1986.

target hexanucleotide and is referred to as a type I neokink. It represents a 25° rotation of the upper half of the double helix relative to the lower half. This unwinding widens the major groove by approximately 3.5 Å. The base pairs do not significantly increase interplanar separation, but the base stacking contacts are clearly changed. This widening of the major groove is essential for recognition by the enzyme: it enables the insertion of four α-helices that otherwise would not fit.

The second localized change from B-form DNA is designated a type II neokink. This occurs at the two symmetrically related phosphates of the guanines in the recognition sequence. The positions and structures of the type I and II neokinks are indicated in Figure 6.

There is an extensive and intimate protein-DNA interface within the complex that is made up of both protein-DNA backbone and protein-base pair interactions. Backbone contacts are made over a region longer than the base sequence of the enzyme's target sequence. Three phosphates (3, 4 and 7 in Figure 6) previously shown by ethylation interference experiments to have the largest effect on protein binding (Lu et al, 1981), are completely buried in the protein and protected from solvent. Phosphate and sugar residues 3, 4 and 5 (on each strand) line the major groove of the recognition hexanucleotide

, which is expanded by the type I neokink. These phosphates are bound within the catalytic clefts of the protein. The clefts are formed by the loops which connect the

a.

0	1	2	3	4	5	6	7	8	9	10	11	12
T	pC	pG	pC	pG	pA	pA	pT	pT	pC	pG	pC	pG

Type II neokink — Type I neokink — Type II neokink

b.

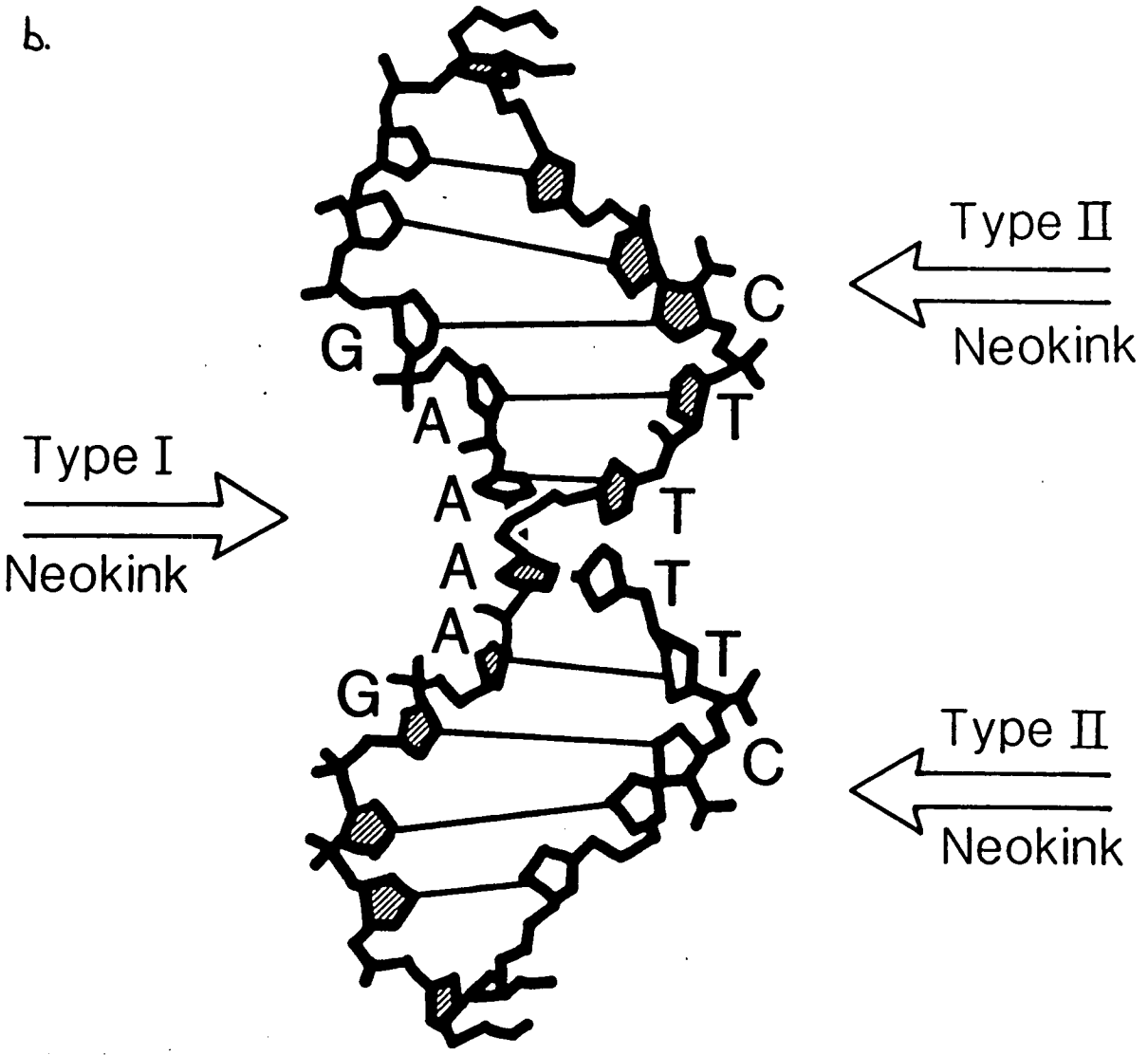


Fig. 6. a) The sequence of the tridecameric oligonucleotide used in the EcoRI cocrystal. Also shown is the location of the kinks and the base numbering scheme.

b) Schematic representation of structure of the type I and type II neokinks. The DNA is symmetric about the centre of the type I neokink.

β-strands in the antiparallel motif and the β-strand 3 to the "inner" α-helix (see Figure 5). The scissile bond is facing this side of the cleft. The other side is formed by the "inner" and "outer" α-helices from the other enzyme subunit. The cleft surface contains many basic amino acid residues that interact electrostatically with the phosphates on the DNA backbone, contributing to binding energy and protein-DNA alignment.

The direct sequence specific DNA-protein interactions are H-bonds between amino acid side chains of Glu 144, Arg 145 and Arg 200, and the purine bases of the target sequence. These amino acids are positioned in two α-helices in each protein monomer. A total of four α-helices therefore enter the widened major groove. Specificity is achieved by precise positioning of these α-helices with respect to the DNA bases.

The inner recognition helix of each monomer contains Glu 144 and Arg 145. The two such helices in the dimer form what is called the inner module which is responsible for recognition of the central AATT tetranucleotide. The symmetrically related outer helices (one from each monomer) recognize the flanking GC base pairs (see Figure 7).

The actual interactions are as follows (and as shown in Figure 8): In the outer module, one H-bond is donated by Arg 200 to the GN7 and another to the O6. In the inner module, the interactions are more complicated in that pairs of amino acid side chains interact with pairs of adjacent adenine residues;

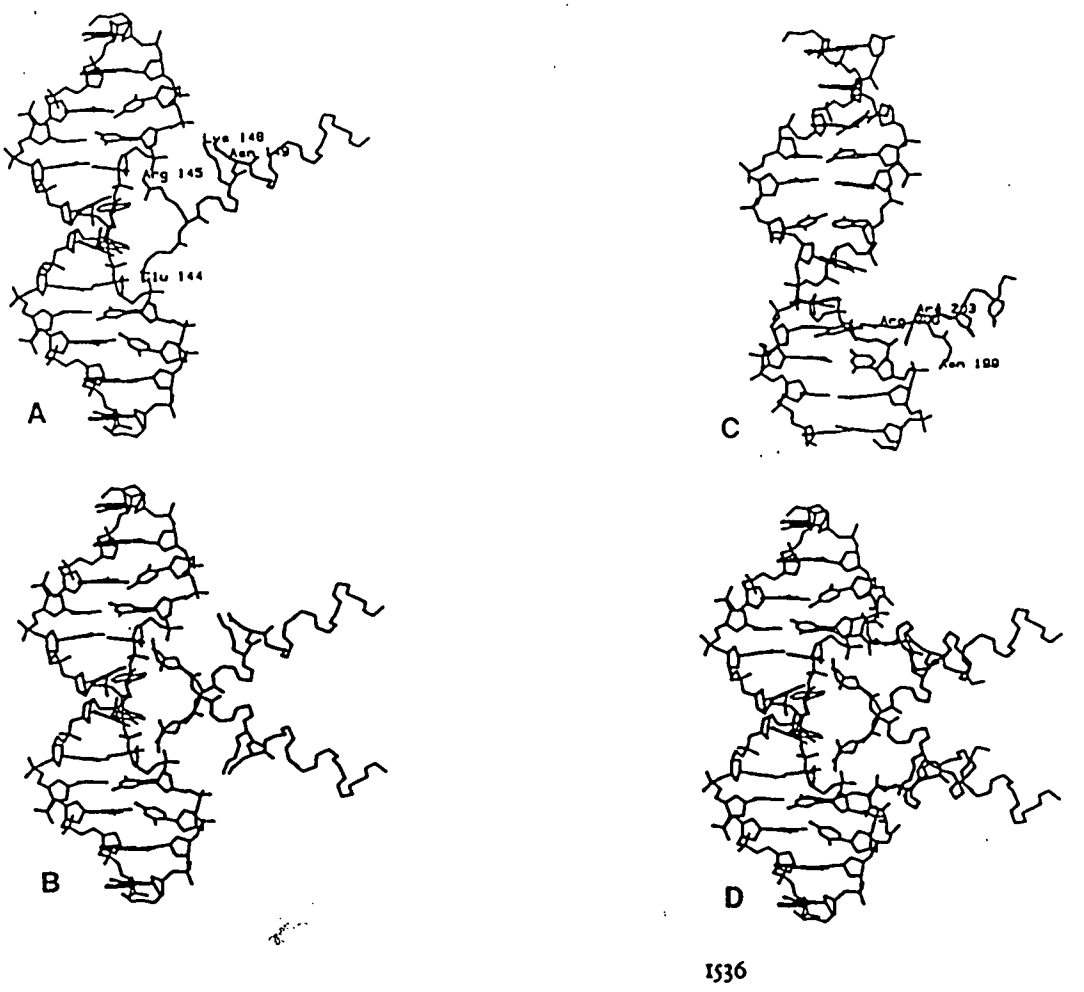


Fig. 7. Drawings showing the recognition α-helices and modules of EcoRI endonuclease interacting with the target site. (A) The inner α-helix of one monomer. (B) The inner recogniton module, containing the inner α-helix from both monomers. (C) The outer α-helix from a single monomer. (D) The outer and inner modules, known as the four helix bundle. Clearly evident is the widened major groove necessary to accommodate the four helices.

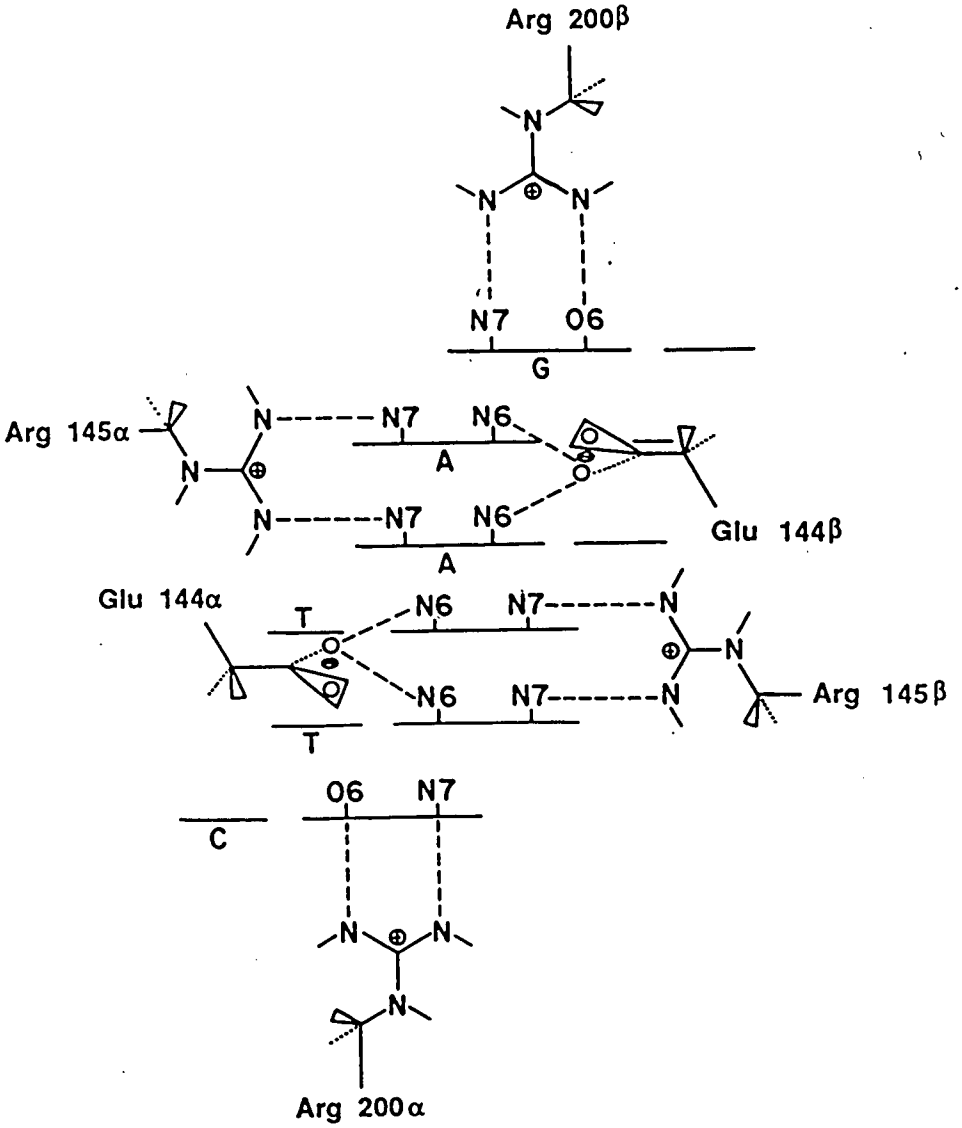


Fig. 8. A schematic representation of the 12 H-bonds that determine the specificity of EcoRI endonuclease. α and β refer to the two enzyme subunits. The positions of the bases and amino acids are such as to make the interactions between them clear, and do not reflect those found in the structure.

each pair of adenines interacts with one amino acid from each enzyme subunit (see Figure 8). The side chain of Glu 144 receives two H-bonds from the N6 amino groups of the adenines while Arg 145 donates two H-bonds to their N7 atoms.

Very evident in this recognition mechanism is the importance of precise positioning of the amino acids with respect to the base which they contact. This is particularly well illustrated by the fact that Arg 145 and Arg 200 specify different bases because of their different orientations relative to the target sequence. This positioning is defined by interactions within the protein structure and between the protein and the DNA backbone.

When bound to its cognate target sequence, therefore, EcoRI endonuclease makes twelve H-bonds to the DNA bases. Changing any one base disrupts at least one H-bond, and hence no other sequence will bind the enzyme as well as the true target. Under some conditions EcoRI can cut at sites other than GAATTC: This so called EcoRI* activity occurs at sequences which differ from the cognate one at a single position (Polisky et al, 1975; Woodhead et al, 1981). The hierarchy of EcoRI* sites, i.e. the preferential order in which the alternatives are cut, correlates well with the number of H-bonds the enzyme should in theory still be able to make.

Another very important characteristic of the protein-DNA interface is the stable array of complementary electrostatic charges on the DNA phosphate backbone and bases, and the amino

acid residues close to the DNA. The negative charges on the Glu 144 are particularly interesting in terms of the biological activity of the restriction-modification system. Displacement of the Glu 144 residues would disrupt this electrostatic complementarity, and hence disrupt the protein-DNA interaction. Glu 144 contacts the central adenine bases, where the EcoRI methylase modifies the DNA by methylating the N6 amino groups. Thus, when modified, the sequence is not only unable to donate the H-bond from AN6 to Glu 144, but also, because of the resultant displacement of this residue, the electrostatic arrangement at the interface is destabilized and hence the binding affinity is enormously reduced. The interaction between the target sequence and the endonuclease is therefore particularly sensitive to the very type of modification produced by the methylase.

The amino acids involved in specific base contacts do not themselves participate in the cleavage reaction: the recognition and cleavage sites are physically distinct. The structure discussed here is the recognition complex, which presumably represents an intermediate in the catalytic pathway. There must be some transition from this inactive recognition complex to an active complex in which the DNA can be cleaved. McClarin et al suggest that this transition is an allosteric affect which can only occur after all specific contacts between the enzyme and its target sequence have formed, and thus it acts to ensure that, when bound to non-cognate sequences, there is no activation and so the DNA is not cut. The allosteric activation therefore acts to increase the specificity of enzyme

action. Conditions that allow EcoRI* activity are presumably those that allow allosteric activation in the absence of some of the normally required specific protein-DNA contacts.

Kinetic data of Modrich and co-workers (Pers. comm. in McClarin et al, 1988) support this idea. They demonstrated that EcoRI spends more time bound to non-specific than to target sequences, but only cleaves at the latter. Also, a mutant in which Glu 111 is replaced by glycine is unable to cleave DNA, though its binds normally. Under EcoRI* conditions, however, this mutant will cut at complete EcoRI sites. Glu 111 is too far from the DNA to participate directly in either recognition or cleavage. It is thought that the mutation influences the transition from inactive to active form.

Yanofsky et al (1987) isolated sixty-two mutant strains in which EcoRI endonuclease, though completely devoid of enzymatic activity, was nevertheless present at wild type levels as judged by western blot analysis. For twenty of these, the entire endonuclease gene has been sequenced and shown to contain single missense mutations. Significant clustering (ten of the twenty mutations) occurred in regions encoding residues 139 to 144 of the protein. Some mutant enzymes were purified and their ability to dimerize tested.

Residues 139 - 144 define a critical region of the protein-DNA interface (McClarin et al, 1986). Ala 139 and Gly 140 are small residues that sit close to phosphate 7 (see

Figure 6) in the DNA backbone. Three mutations (139 Val, 139 Thr and 140 Ser) replace these with bulkier groups which would disrupt the positioning of the protein on the DNA. Glu 144, as described above, is in the inner recognition helix and forms direct contacts with the adenines in the target sequence. It also participates in the electrostatic complementarity within the complex. A non-functional mutant with Lys at this position was isolated: this substitution would disrupt both the H-bonds to the bases and the electrostatic interactions.

Several mutations change residues near the dimer interface. Simply destroying this interaction will produce a completely non-functional enzyme as only the dimer is active. Three mutant enzymes (Glu 144 -> lys, Glu 152 -> lys and Gly 210 -> Arg) were shown to be unable to dimerize. The first of these mutants has already been described in terms of how it would disrupt the protein-DNA interface; it also results in the loss of an electrostatic interaction between subunits as Glu 144 would normally interact with Arg 145 of the other monomer as part of the stabilizing of the inner recognition module.

Glu 152 is buried in the subunit interface and presumably makes a number of important hydrophobic interactions which are lost when it is replaced by lys. Gly 210 is situated at the end of the outer recognition helix close to residues in both the same and the other subunit; again dimer stability would be lost when this residue is replaced by Arg.

Another mutation was isolated which is close to neither the protein-DNA nor dimer interfaces. The crystal structure shows that it probably upsets correct positioning of the inner recognition helix. Three aromatic residues occur adjacent to one another (Phe 163, Pro 164 and Tyr 165) in the middle of β-strand 4. These interact with the residues in the inner recognition helix, thereby precisely locating it within the major groove of the DNA. Replacement of Pro 164 with Ser perhaps allows too much flexibility in this structure, resulting in loss of precision in recognition helix presentation.

The amino acids identified as making direct H-bonds to base pairs in the target sequence (McClarin et al, 1986) have been changed by site directed mutagenesis to a variety of alternative residues in an attempt to produce a functional enzyme of new (or relaxed) specificity (J. Heitman, pers. comm.; J. Rosenberg, pers. comm.). Arg 200, which contacts the outer G-C base pairs, has been systematically replaced by all other nineteen possible residues. Most of these are non-functional in vivo and, where tested, in vitro. A few, those with very conservative substitutions, still show weak, but EcoRI specific, activity: they kill a cell in which they are expressed unless the EcoRI methylase is also present. Sixteen substitutions of Glu 144 and twelve of Arg 145 have also been constructed; again none shows altered specificity.

These results emphasize the complex nature of sequence selection by restriction enzymes. The residues that make

specific contacts with the DNA target sequence are only a single component of the overall recognition process which directs the enzyme to accept only the correct sequence as a substrate. The whole enzyme structure is organized to cut a single target sequence. Altering specific contacts within the recognition complex is perhaps unlikely to alter the specificity of the entire restriction process. Even if a mutant enzyme could bind to a new target sequence - altered binding specificity - there is no reason why the recognition complex formed should act as a signal to the enzyme for transition to an active form which, as described above, is essential for cutting of the DNA.

No structure is available for the EcoRI methylase, but evidence suggests that the methylase and endonuclease recognize the target sequence in different ways. The two enzymes show no amino acid sequence similarity (Greene et al, 1981; Newman et al, 1981). While the endonuclease interacts with the DNA symmetrically as a dimer (Lu et al, 1981; McClarin et al, 1986), the methylase binds as a monomer and therefore must have an asymmetric interaction with target sequence (Rubin and Modrich, 1977).

Methylation and restriction of a set of octadeoxyribonucleotides containing modified EcoRI recognition sequences supports this idea in that there are marked differences in the effects of various base analogue substitutions on the activities of the two EcoRI enzymes (Brennan et al, 1986a and b). The various base analogues used alter functional groups in

the major and minor grooves of the target site. Their effects were interpreted on the assumption that an effect on methylation or restriction is caused by the involvement of the altered group in recognition, catalysis, or simply its proximity to the enzyme. Other indirect influences, however, could be relevant, such as analogue induced conformational changes in the oligonucleotides (or inhibition of enzyme induced changes) or upsetting of AdoMet binding by the methylase. Some alterations in the substrate were seen to interfere with enzyme activity even though not close to the bound protein, e.g. the effect of groups in the minor groove on endonuclease function. These presumably interfere with formation of the kinked DNA conformation necessary for endonuclease action (see Figure 6) (Frederic et al, 1984). In some cases removal of a functional group altogether was less disruptive than replacing it with an alternative one. This is not surprising as its removal may result only in the loss of a single DNA-protein contact, whereas replacement may upset the overall snug fit found at the protein-DNA interface causing, indirectly, loss of many favourable interactions.

An illustration of the different ways in which the two enzymes see the same target sequence is shown by changes to groups on position 5 of cytosine in the major groove and position 1 of guanine in the minor groove. The former disrupts endonuclease but not methylase activity, the latter methylase but not endonuclease (Brennan et al, 1986 b).

In all comparative studies of the substrate requirements of various type II restriction and corresponding modification enzymes, the two enzymes differ (Kaplan and Nierlich, 1975; Berkner and Folk, 1977; Modrich and Rubin, 1977; Dwyer-Hallquist et al, 1982; Bodnar et al, 1983; Lu et al, 1981; Mann et al, 1978; Marchionni and Roufa, 1978; McClelland and Nelson, 1988). Similarly, comparisons of the activities of isoschizomers HaeIII, BspI and BsuRI, all of which recognize GGCC, on oligonucleotides containing various base analogues show that these all interact with the sequence in different ways. The simplest explanation of the results is that HaeIII interacts with the major and minor grooves, BspRI with the minor and BsuRI the major grooves only (Wolfes et al, 1985). Of course, the complications in interpreting these results, as discussed above for the EcoRI methylase and endonuclease comparison, apply here too. At least with the isoschizomers, however, the enzymatic reaction is the same in each case, and so some complications may be alleviated.

Of particular interest is RsrI, an isochizomer of EcoRI (Greene et al, 1988). Not only does this enzyme recognize the same target as EcoRI and cut at the same positions within it, but also, both reactions have identical pH and Mg^{2+} concentration optima. EcoRI methylation of the target sequence protects against RsrI restriction. RsrI cross reacts strongly with EcoRI endonuclease antiserum, indicating three dimensional structural similarities between the two enzymes. Sequence of only the N-terminal 34 amino acids of RsrI is available, but this is similar to a region of EcoRI very close to its N-

terminus. Clearly the structural details of how the RsrI endonuclease recognizes its target sequence will be very interesting in the light of the detailed information already available for EcoRI (McClarin et al, 1986).

Recently Chandrasegaran and Smith (1988) have compared the predicted amino acid sequences of seventeen methylases and eight endonucleases. From this they conclude:

i) There is little significant similarity among the restriction enzymes, even isoschizomers.

ii) Endonucleases are not significantly similar to their corresponding methylases.

iii) Methylases show extensive similarities, particularly when sharing similar recognition sequences.

If sequence similarity implies evolutionary relatedness, then it might appear that the methylases evolved from a relatively small number of archetypal enzymes while the endonucleases have arisen independently of each other (in most cases) and of the methylases (Chandrasegaran and Smith, 1988).

All adenine methylases so far looked at contain the sequence N/D PP Y/F (Loenen et al, 1987; Chandrasegaran and Smith, 1988). This sequence is not found in cytosine methylases and may be involved in AdoMet binding and/or adenine/6 methyl adenine recognition. Strong support for the idea that this sequence may be involved in a direct physical interaction with methylated adenines comes from a mutant phage

P22 Mnt repressor which binds specifically to operators containing methylated adenines (Youderian et al, 1983).

The Mnt repressor of phage P22 acts in the regulation antirepressor synthesis (Suskind and Youderian, 1983). Neither the structure nor the mechanism of DNA binding are known. However, mutational evidence suggests that the N-terminal ten amino acids are important in operator recognition (Vershon et al, 1987; R.T. Sauer, pers. comm.). It has been demonstrated that a histidine to proline change at position 6 alters binding specificity (Youderian et al, 1983). The sequence recognized by this mutant has a G to A change generating a GATC sequence within the operator. This sequence is the site for Dam methylation (Marinus, 1987). Methylation of the adenine to N6 methyl adenine is in fact essential for recognition of this sequence by the mutant repressor; when unmethylated it is bound 1000x less well (Vershon et al, 1985). The histidine to proline change in the repressor alters the sequence previously implicated in DNA recognition from ARDDPHFNF to ARDDPPFNF (Youderian et al, 1983), thereby producing the DPPF tetrapeptide found in all adenine methylases (Loenen et al, 1987; Chandrasegoran and Smith, 1987). It is thought that in the wild type Mnt repressor-operator complex His 6 makes a contact with the guanine. Introducing Pro 6 presumably allows an alternative contact to be made which is dependent on the 6 methyl group on the adenine (Vershon et al, 1985). It is surely no coincidence that this amino acid sequence is apparently important in two completely different systems where recognition of methylated adenines is required?

CHAPTER 2 : RESTRICTION AND MODIFICATION SYSTEMS

1) General Introduction

Three types of restriction and modification (R-M) systems are known: types I, II and III (see Bickle, 1987 for review). Type II are the simplest and, as mentioned in the previous section, consist of two separate enzymes - an endonuclease and a methylase - with identical specificity for the DNA target sequence. An endonuclease acts as a dimer in recognizing unmethylated palindromic DNA sequences of 4 - 8 bp in length. It then cuts both strands of the DNA at defined positions, normally within the target sequence. The modification enzyme methylates two adenine or cytosine residues within the same target sequence, thereby rendering it no longer susceptible to the endonuclease (Modrich and Roberts, 1982). Some type II enzymes have been found which recognize asymmetric target sequences and cut the DNA several nucleotides to one side of that sequence. These are now classified as type IIS (Wilson, 1988).

Type I systems are the most complex, consisting of two enzymes made up of three subunits. One is a methylase, the other both a methylase and an endonuclease (Bickle, 1982; Yuan, 1981; Endlich and Linn, 1981 for reviews). Though both activities are triggered by recognition of the same target site, and modification involves methylation of adenines within that sequence, restriction actually occurs at seemingly random sites up to several kb away.

Type III systems include a DNA methylase, which is a monomer, and a restriction enzyme, consisting of this in complex with a second subunit. The DNA cleavage occurs 25 - 27 nucleotides away from their target sites (Bickle, 1982 for review).

All three types of endonuclease require Mg^{2+} for activity. Type I enzymes also require S-adenosyl methionine (AdoMet) and ATP, which is hydrolysed during the cleavage reaction. Type III endonucleases require ATP, but do not hydrolyse it. AdoMet is the methyl donor used by all the methylases (Bickle, 1987).

2) Type I Restriction and Modification Systems

A) Characteristics of the System

Type I restriction and modification enzymes are made up of three types of subunit, R, M and S, encoded by the hsdR, M and S genes (for reviews see Arber and Linn, 1969; Boyer, 1971; Meselson et al, 1972; Modrich, 1979; Yuan, 1981; Bickle 1982 and 1987). S determines the specificity of the DNA target sequence recognized and, together with M, forms a modification enzyme that methylates two adenines within this sequence. The inclusion of R subunits in this complex produces an enzyme with both methylase and endonuclease activities. How this enzyme behaves is dictated by the methylation state of the target sequence with which it interacts. If this sequence is fully methylated, the enzyme dissociates from the DNA; if

hemimethylated, the complex methylates the complementary strand; only unmethylated targets activate DNA cutting. This follows translocation of the DNA through the bound enzyme (details and references will be given below).

Hemimethylated DNA is the product of semiconservative replication of fully methylated DNA, whereas foreign DNA entering the cell will normally have completely unmethylated targets. A single species that methylates the former and restricts the latter can therefore usefully be maintained in the cell.

Type I R-M enzymes have a number of interesting characteristics, the molecular details of which are amenable to investigation. These include: recognition of specific DNA sequences; protein-protein interactions involved in binding between subunits within the complex; non-specific DNA-protein interactions; DNA translocation; the enzymatic functions themselves - methylation, restriction and DNA dependent ATPase activity.

The specific interaction between the enzyme and its target sequence is very subtle in that, not only is the target distinguished from non-target sequences, but the methylation state is identified by the enzyme. This is true of type II endonucleases also but, in that case, methylation merely blocks enzyme binding (see previous chapter). With type I enzymes, different methylation states are all bound by the enzyme, but induce alternative activities.

Another appealing characteristic of type I systems is that they can be grouped into families (Murray et al, 1982). Each family contains enzymes originally shown to be related by their ability to interchange subunits (Glover and Colson, 1969; Boyer and Roulland-Dussoix, 1969; Van Pel and Colson, 1974; Bullas and Colson, 1975; Fuller-Pace et al, 1985). More recently, molecular studies have confirmed these relationships (Murray et al, 1982): DNA encoding enzymes within a family cross hybridize, as do antibodies raised against the enzyme subunits. Nucleotide sequences also show the highly conserved nature of members within a family (see Loenen et al, 1987; Gough and Murray, 1983; Bickle, pers. comm.; own unpublished results). In contrast, enzymes from different families show very little similarity by any of these criteria, even though, in terms of organization and function, all type I enzymes are very alike (Fuller-Pace et al, 1985; Suri et al, 1985; Suri and Bickle, 1985; Price et al, 1987a). Areas of sequence conservation within or between families are helpful in allocating possible functions to different regions of the various polypeptides (see, for example, Gough and Murray, 1983). Questions concerning the evolution of these enzymes are also raised by the family groupings and these will be discussed later (see Chapter 5; also Murray et al, 1982; Nagaraja et al, 1985a; Daniel et al, 1988; Gann et al, 1987).

Three families of type I enzymes have been identified (see Bickle, 1987). Each is named after its archetypal member, these being EcoK, A and R124. Alternatively, they are

sometimes referred to as Ia, Ib and Ic respectively (Bickle, 1987). The K-family is, to date, the most extensively studied.

B) Genetic Determinants

EcoK, B and D from E.coli, and StySP and SB from Salmonella strains have been identified as members of the K-family. Each is encoded by three genes; hsdR, M and S. Phenotypes are expressed in terms of restriction (r) and modification (m) proficiency (+) or deficiency (-). In some of the experiments described below I have included descriptions of genotypes, even when these were not known at the time. These are indicated in brackets and are intended merely to clarify the results described.

The first mutations in the hsd genes were isolated by Wood (1966). He found that about half the mutants with a restriction minus phenotype (r-) were also deficient in modification (m-). He concluded, because of their high frequency, that these r-m- mutants were not double mutants, but simply mutants of a gene essential for both activities. This was confirmed by complementation tests using an F′ encoded system with an r-m+ phenotype (hsdR- M+ S+ genotype). This complemented a first step r-m- mutant (hsdR+ M+ S- or hsdR+ M- S+) but not an independent r-m+ (hsdR- M+ S+) mutant (Boyer and Roulland-Dussoix, 1969; Glover, 1970). In the same study, second step r-m- mutants, produced from r-m+ mutants, were also shown not to complement the F′ system.

The *hsdS* gene was shown to be the determinant of specificity in interstrain complementation between the *Eco*K and B systems which, though related, are of different specificity (Boyer and Roulland-Dussoix, 1969). They showed an $r^-_K\ m^+_K$ /$r^-_B\ m^-_B$ (first step mutant) diploid gave an $r^+_K\ m^+_K$ phenotype, demonstrating that the specificity determinant of the *Eco*K, but not *Eco*B, system was still functional. Also, a second step $r^-_B\ m^-_B$ mutant was produced that could complement a wild type K-system ($r^+_K\ m^+_K$) to give $r^+_{KB}\ m^+_{KB}$, indicating that the $r^-_B m^-_B$ mutant was in fact *hsdR*- *M*- *S*+. These second step $r^-_B\ m^-_B$ mutants could also complement a first step $r^-_B\ m^-_B$ (*hsdR*+ *M*+ *S*-) to give wild type B phenotype.

Hubacek and Glover (1970) isolated temperature sensitive restriction mutants of *Eco*K (r^{ts}_K). The idea was to subsequently isolate modification deficient mutants at the non-permissive temperature, thereby avoiding the lethal consequences of an r+m- phenotype. However, they found that many of their original r^{ts} were also m^{ts}, and that selection for m- derivatives of these also led to an r- phenotype in all cases. Complementation tests between both the first and second step mutants and an F' *hsdS*^+_B *M*+ *R*- gave an $r^+_{BK}\ m^+_{BK}$ phenotype in almost all cases, implying that the original *ts* mutation was in the *M* gene, as was the second mutation that produced the $m^-_K\ r^-_K$ phenotype (Hubacek and Glover, 1970). One second step mutant gave an $r^+_B\ m^+_B$ phenotype in the complementation test implying that its second mutation was in *S*.

All of these results imply that there are three genes, hsdR, M and S, encoding three enzyme subunits. S defines the specificity and together with M alone is sufficient for modification. Restriction requires all three subunits.

In vitro complementation studies (Kuhnlein et al, 1969; Hadi and Yuan, 1974) support this in showing that a purified mutant enzyme encoded by the genes hsdS+ M+ R- shows only modification activity. However, it can complement an hsdS- M+ R+ mutant, which alone shows neither methylation nor restriction, to give both activities.

Sain and Murray (1980) cloned the hsdK genes in a λ vector. They identified three polypeptides encoded by the cloned hsd genes, and established the order of the genes to be hsdR M S. They also suggested the presence of two promoters, one upstream of R and one of M. This has subsequently been confirmed by DNA sequencing and lacZ fusions (Loenen et al, 1987).

C) The Enzymes

The complete EcoK enzyme, consisting of R, M and S subunits, has been identified as a 400,000 MW complex by gel filtration and glycerol gradients (see Meselson et al, 1972). Subunits of Mr 135,000, 62,000 and 55,000 were identified by denaturing polyacrylamide gels in the estimated ratio 2:2:1. The subunit molecular weights were confirmed by examination of polypeptides produced by λ phage in which the hsdK genes had

been cloned (Sain and Murray, 1980). Deletion derivatives of these phage allowed the Mr 135,000 polypeptide to be identified as R, the Mr 62,000 as M and the 55,000 as S (Sain and Murray, 1980). This is in good agreement with the sizes predicted from the sequences of the genes (Loenen et al, 1987). Analysis of the phenotypes and polypeptides encoded by these deletion derivatives confirmed that only M and S are required for modification. An enzyme having only methylase activity has also been purified and shown to contain only the M and S subunits in the ratio 1:1 (Suri et al, 1984a).

Under non-denaturing conditions the EcoB system was shown to contain two enzyme species (Eskin and Linn, 1972 a). Both contained three subunits of about the same sizes as for EcoK. However, in one species, these occurred in the ratio 1:2:1 and, in the other, 1:1:1. A third species containing just the two smaller subunits and which has only methylase activity has also been isolated (Lautenberger and Linn, 1972).

The differences in subunit composition of EcoK and B restriction complexes may well be an artifact caused by the different purification procedures. Considering their relatedness, which allows interchange of subunits, it seems unlikely that any real difference in subunit composition exists.

Examination of mutant EcoK enzymes has enabled individual biochemical activities to be assigned to the various subunits. Mutations in hsdS produce enzymes which have been shown to lack

methylase and endonuclease activities, as well as being unable to bind DNA to filters or exhibit any DNA dependent ATPase activity (Hadi and Yuan, 1974). An hsdM mutant also lacked methylase, endonuclease and DNA binding activity. However, it did show some ATPase activity (Buhler and Yuan, 1978).

D) Reaction Mechanisms

Type I enzymes can recognize and methylate specific nucleotide sequences, translocate and cut DNA, and hydrolyse ATP when bound to DNA (see Bickle, 1987). The mechanisms are complex, reflecting the enzyme's ability not only to perform, but select subsets of these functions under the influence of the methylation state of the target sequence.

Both EcoK and B have been used as model systems for the reaction mechanisms (Yuan et al, 1980; Studier and Bandyopadhyay, 1988; Rosamond et al, 1979; Endlich and Linn, 1985). Due to their relatedness, it would seem likely that both should act in very similar ways.

Co-factors required by the complete restriction enzyme complexes, which act as both endonucleases and methylases, are Mg^{2+}, AdoMet and ATP (Meselson and Yuan, 1968; Yuan and Meselson, 1970; Vovis et al, 1974). The latter two, as well as the DNA, act as substrates and allosteric effectors (Yuan et al, 1975; Bickle et al, 1978; Habermann et al, 1972). The enzyme made up of only S and M subunits, and capable of only methylase activity does not require (and is unaffected by) ATP

(Suri et al, 1984a). The AdoMet acts as the methyl donor in the methylation reaction (Haberman et al, 1972), but its allosteric effect enables the enzyme (simple methylase or entire complex) to bind DNA, no affinity for which is seen in the absence of this co-factor (Yuan et al, 1975). The initial DNA binding is probably non-specific (initial complex) and is followed by tighter binding to the recognition sequence (recognition complex) (Yuan et al, 1975) (see Figure 9). In this complex, the EcoK enzyme is referred to as EcoK*. The initial non-specific binding probably aids the enzyme in locating its target sequence; by binding to a DNA molecule anywhere and then searching along it, a protein can reduce the dimensionality of the search by limiting it to the surface of the DNA molecule, rather than the entire volume of the cell (von Hippel et al, 1974; Berg et al, 1981). Experiments with the type II endonucleases EcoRI, HindIII and BamHI suggest that they use such a mechanism in vitro; any importance in vivo has not been demonstrated (Ehbrecht et al, 1985).

Methylation:

Methylation can be carried out by either the simple two subunit methylase or the entire restriction complex (Suri et al, 1984a). With completely unmethylated substrate DNA, the restriction enzyme is considerably less effective than the methylase and is somewhat inhibited by ATP. This has been shown for EcoK and B (Suri et al, 1984a; Haberman et al, 1972; Lautenberger and Linn, 1972). The inhibition of methylation by the restriction complex may be due to the release of AdoMet

from the enzyme following the ATP induced conformational change which is important to the mechanism by which the enzyme distinguishes between fully methylated, hemimethylated or completely unmethylated DNA (Bickle et al, 1978); its usual response to unmethylated DNA is restriction, not methylation.

Both the methylase and restriction enzyme modify hemimethylated DNA more efficiently than an unmethylated substrate. For EcoK the methylase has been shown to be a little more efficient than the restriction complex (Suri et al, 1984a). ATP stimulates the complex about twofold in this situation. Comparing the rate constants for the two enzymes on unmethylated and hemimethylated DNA, it is found that the restriction complex methylates the latter about 150-fold more efficiently than the former; for the simple methylase the difference is 35-fold (Suri et al, 1984a).

Restriction:

In the absence of ATP the restriction enzyme forms recognition complexes irrespective of the methylation state of the target sequence. Indeed, the enzyme binds almost as well to modified as to unmodified sites, though the relative stabilities of these complexes differ (Bickle et al, 1978). Binding of ATP, however, causes a conformational change in the enzyme which enables it to discriminate these different methylation states. If the site is methylated, then the complex dissociates from the DNA. When hemimethylated, the unmodified strand is efficiently methylated as described above

(Vovis et al, 1974; Vovis and Zinder, 1975; Burckhardt et al, 1981a and b). If, alternatively, the site is completely unmodified, methylation is very inefficient, and a complex series of events occurs resulting in DNA translocation and cutting at sites 0.5-7kb from the recognition site (Bickle et al, 1978; Yuan et al, 1980; Endlich and Linn, 1986a). The major steps and decisions are outlined in Figure 9.

The ability of EcoK to discriminate between different methylation states of its target sequence has been considered in detail by Burckhardt et al (1981b). They showed that binding of the restriction enzyme to methylated, hemimethylated and unmethylated target sites formed three different recognition complexes even before the ATP induced conformational change; i.e. they suggest that it is the way EcoK* sits on its target site initially that determines the enzyme's subsequent response to ATP binding. They envisage that the enzyme uses the AdoMet bound to its M subunits as probes for the presence of methylated adenines in the major groove of the DNA. When the site is fully methylated, both the M subunits are excluded from the major groove due to the steric hindrance between the AdoMet and methyl groups on the adenines. This enzyme-DNA conformation they call an open complex. With heteroduplex DNA, one M subunit can enter the major groove, while the other is excluded; this, which they call a partially open complex, results in methylation of the unmodified adenine. With an unmodified site, both M subunits can enter the major groove to produce a closed complex. This is thought to

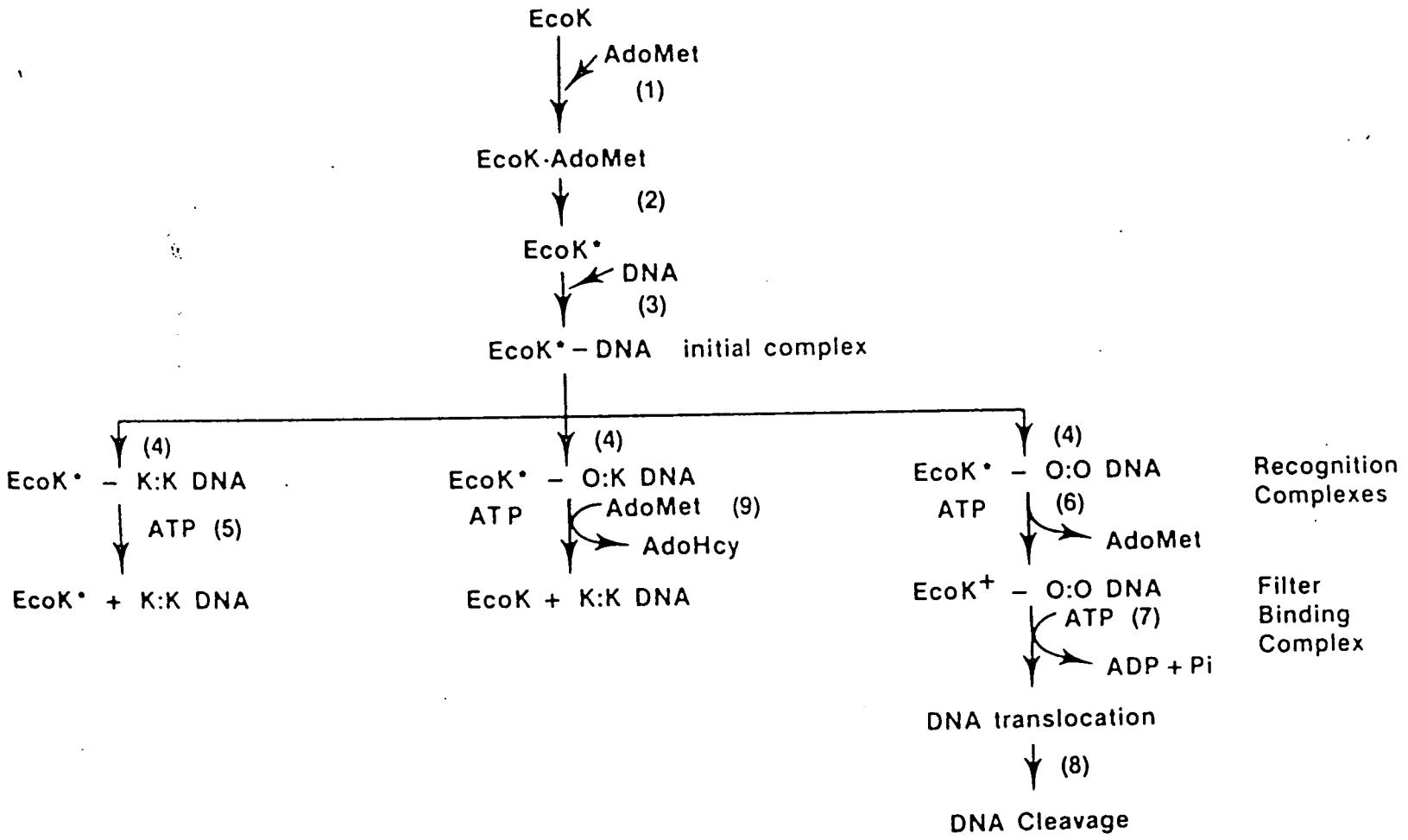


Fig. 9. The reaction mechanism of the restriction endonuclease EcoK. Steps 1-3 are identical irrespective of the methylation state of the target site. Steps 4 onwards vary depending on whether the target is fully methylated (K:K), hemimethylated (O:K) or unmethylated (O:O).

position the R subunits appropriately for DNA translocation and cleavage.

When ATP is bound by an enzyme in this latter complex, the conformational change that occurs is very large and may even represent the loss of a subunit (Bickle et al, 1978). For EcoK, this form of the enzyme is called EcoK+ (Figure 9). The recognition complex is converted to a filter binding complex, so named because it can be retained on a nitrocellulose filter. It also leads to release of AdoMet from the enzyme. The conformational change does not require ATP hydrolysis; it can be induced by non-hydrolysable analogues of ATP. Later steps in DNA translocation and cutting, however, do require hydrolysis (Bickle et al, 1978).

Electron microscopic studies on intermediate structures formed during DNA translocation and restriction have been performed for EcoK (Yuan et al, 1980) and EcoB (Rosamond et al, 1979; Endlich and Linn, 1986) and models have been put forward to account for these observations. More recently an explanation of how the sites of cleavage are selected has been based on examination of the immediate products of EcoK cleavage of the phage T7 genome (Studier and Bandyopadhyay, 1988).

The original work of Yuan et al (1980) examined EcoK digestion of unmodified plasmid pBR322 DNA. They demonstrated that addition of ATP to recognition complexes led to the production of linear DNA molecules and DNA fragments with enzyme still bound to the recognition sites. Formation of

recognition complexes in the absence of ATP allowed synchronization of the reactions; samples were analyzed at ten second intervals after addition of ATP. Several novel structures were observed:

i) Twisted loop : Supercoiled or relaxed DNA with a tightly wound loop that had its origin in an EcoK+ molecule.

ii) Regular loop : Supercoiled DNA with EcoK+ making a two point attachment with the DNA to form a loop.

iii) Double twisted loop : Two twisted loops, assumed to be formed by two enzyme molecules translocating DNA towards each other.

The kinetics of formation of these structures implies a series of events:

Recognition -----> Regular -----> Twisted -----> Linear

Complex Loops Loops + Fragments

These results were all gained using circular DNA molecules; when previously linearized, no such structures were seen. This was thought to be due to complete translocation of the DNA through the enzyme before even the first samples were taken. The appearance of the various loop structures when longer linear DNA substrates were used (e.g. λ) supports this idea. Presumably, circular DNA produces constraints on translocation which slow or stall the process (Yuan et al, 1980).

It was also established that EcoK can cut DNA on either side of its target sequence (Yuan et al, 1980). DNA substrates containing single target sites very close to either of the ends are cleaved. If translocation only occurred in one direction, it was thought that one of the DNA molecules would simply translocate right through the enzyme before the minimum length of DNA translocation necessary for cleavage could occur. This was based on the observation that cuts do not usually occur within 500 bp of a target site, and so it was assumed that this amount of DNA translocation was a prerequisite for the cutting reaction.

Initial studies of EcoB (Rosamond et al, 1979) produced slightly different results. In this case linear phage fd DNA was used as substrate and relaxed loops (presumably comparable with the regular loops of Yuan et al, 1980) were the only intermediate structures seen. Subsequently (Endlich and Linn, 1985), supercoiled/twisted loops were seen, encouraging the belief that differences seen between EcoK and B may be simply due to different experimental conditions. It was also claimed (Rosamond et al, 1979; Endlich and Linn, 1985) that EcoB can only cut DNA 5′ to its target sequence. As pointed out by Studier and Bandyopadhyay (1988), these experiments, designed to establish the direction of DNA translocation and hence the location of cutting with respect to the enzyme's target sequence, depend on seeing cutting of some DNA substrates and not others. Hence any reason for preferential cutting of alternative substrates can easily be misinterpreted as

demonstrating a unidirectional translocation. It seems most likely that EcoK translocates DNA in either direction (Yuan et al, 1980; Studier and Bandyopadhyay, 1988) and it is difficult to imagine that EcoB behaves differently. If there are any differences in the mechanisms of these enzymes, then it would be of interest to see which parental behaviour was adopted by mixed enzymes containing different combinations of subunits from the two systems.

The model of EcoK action put forward by Yuan et al (1980) envisaged the enzyme as having two DNA binding sites - one specific for its target sequence, and the other, non-specific and only available after the ATP induced conformational change (Bickle et al, 1978). DNA on either side of the recognition complex can, on random collision with the enzyme, be bound by this non-specific site, thereby producing a regular loop. The randomness of this interaction means that the bound DNA can be in any one of four different conformations (see Figure 10). The DNA is then wound past this second site while the enzyme remains tightly bound to its target site. The winding is always in the same direction, irrespective of the original conformation of the regular loop, and leads to the production of twisted loops (Figure 10). The twisting presumably occurs because the winding involves tracking the major or minor groove, or backbone, of the DNA, past the enzyme which remains at a fixed position bound to its target sequence. This would cause the DNA to rotate as it passed through the enzyme.

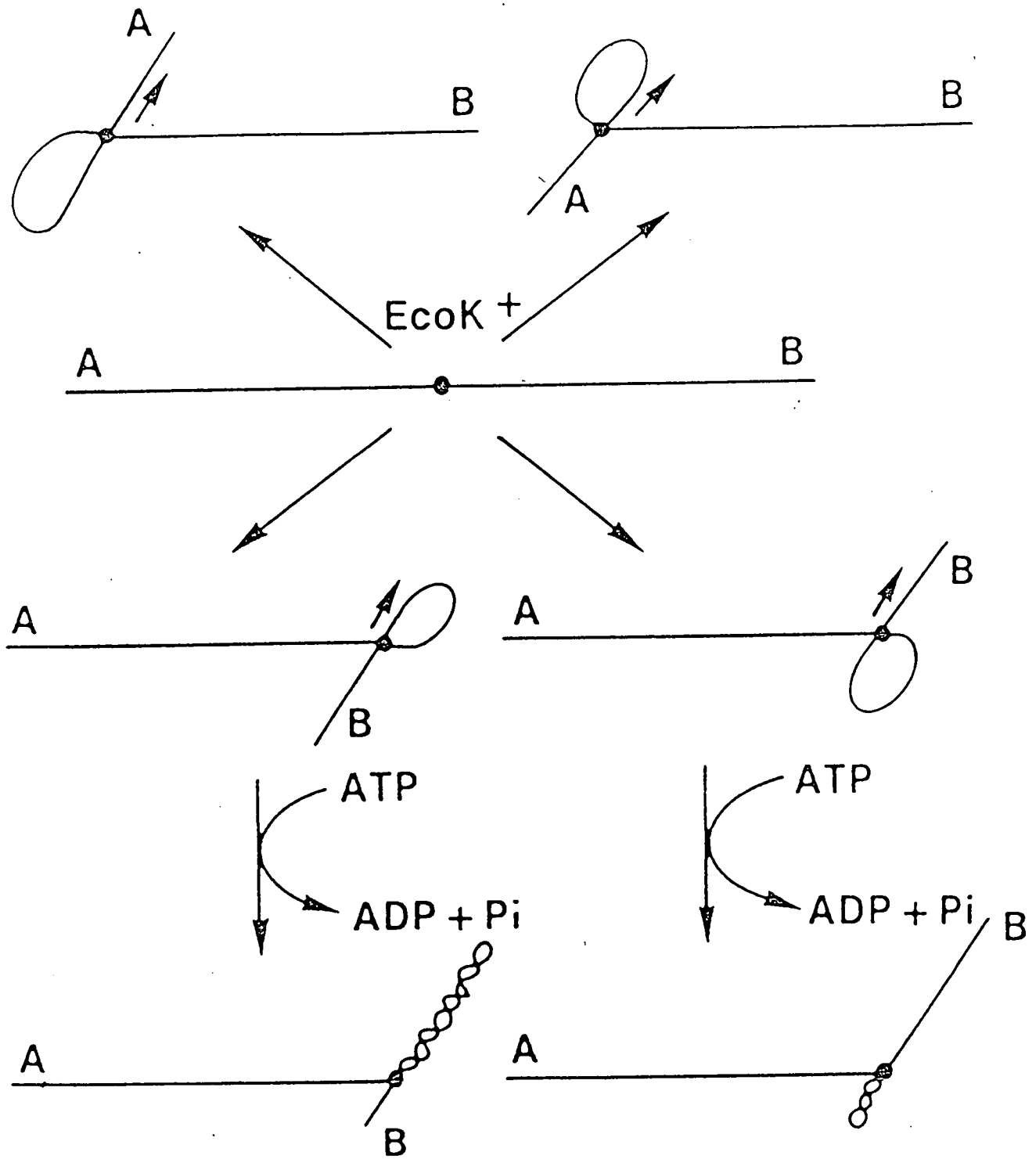


Fig. 10. The four possible conformations of regular loop formed by EcoK, and the result of subsequent DNA translocation. A and B refer to the different ends of the DNA substrate.

Whatever the mechanism for their formation, the strain induced by twisted loops may act to trigger DNA cleavage. Cutting occurs in two steps: initially the DNA is nicked in one strand, and only subsequently in the second, probably by a different enzyme molecule (Meselson and Yuan, 1968; Adler and Nathans, 1973; Eskin and Linn, 1972b).

EcoK and B enzymes show an enormous DNA dependent ATPase activity (Bickle et al, 1978). This commences prior to DNA cleavage and very likely drives DNA translocation. In vitro, this activity continues for several hours after cleavage. Endlich and Linn (1985a) have suggested that the continued ATPase activity may be due to a scanning function of the enzyme. This, they propose, occurs before restriction, enabling the enzyme to check that the DNA substrate is unmodified and has not previously been restricted. They envisage the continuation of the scanning to involve back tracking along the preformed loop, and feel that the interruption of the ATPase activity induced by a single cut within the loop supports this idea (Endlich and Linn, 1985a). Though the continued scanning and concomitant ATPase activity seems wasteful, it may be that in vivo it is short lived due to rapid degradation of the restriction fragments by cellular nucleases (Simon and Lederberg, 1973).

Studier and Bandyopadhyay (1988) have recently proposed a model for how primary sites are selected. The position of cutting produced by type I enzymes has often been thought to be random (Hartmann and Zinder, 1974; Murray et al, 1973; Horiuchi

and Zinder, 1972). However, examination of the initial products of EcoK digestion of phage T7 DNA in vitro has revealed that discrete fragments are produced (Studier and Bandyopadhyay, 1988). This DNA has four recognition sites whose positions are known from the DNA sequence (Dunn and Studier, 1983). The positions of the primary cuts occur directly between adjacent target sites and are produced after intervals of time that are proportional to the distances between those sites (Studier and Bandyopadhyay, 1988). Their model claims that each enzyme bound to a target site translocates DNA towards itself from both directions until it collides with another such enzyme bound to a neighbouring target. This collision induces DNA cutting. From the times taken for cutting to occur in the various intervals between target sites in the T7 genome, it can be concluded that:

i) Initiation of translocation is immediate.

ii) Cutting occurs immediately after collision.

iii) The rate of translocation is $\sim$200 bp/second.

At sufficiently high enzyme to substrate molar ratios, cutting appears not to need a collision between two DNA bound enzymes, presumably due to some cooperation between one DNA bound and one free enzyme. Such a process may also account for the subsequent cutting of primary restriction fragments, thereby producing the apparent random cutting so often observed in the past (Murray et al, 1973).

A number of predictions are suggested by this model. For these and previous observations to be compatible, it seems that

the outcome of a given reaction will depend on the molar ratio of enzyme to DNA, the number of recognition sites per DNA molecule, and whether the DNA is linear or circular. Linear molecules having a single site will not be cut unless a high enzyme to DNA ratio is used. Linear molecules with two or more sites (e.g. T7) will initially be cut between sites as described. If the enzyme/DNA ratio is sufficiently high, secondary cuts will occur, analogous to linear molecules with single sites. Circular molecules with several sites are essentially the same as linears with several sites; indeed, after the initial cut they will be just that. Circular molecules with a single site may be an odd case. Although there will not be two DNA bound enzyme molecules, translocation of the DNA by a single enzyme will eventually stall under some topological constraint. Such an enzyme apparently then cuts one strand of the DNA. At sufficiently high enzyme concentrations, a second molecule will cooperate in cutting the second strand. The enzyme/DNA ratio required for this is lower than that needed to make 2 cuts in linear molecules with a single site.

Primary cleavage occurring between adjacent target sites is exactly the hypothesis put forward by Brammar *et al* (1974) to explain restriction of λ *trp* phages *in vivo*. They looked at the effect of K-restriction on expression of the *trpE* gene in the phage. These experiments were done in a *recBC* host, which is deficient for the nuclease responsible for degrading the DNA fragments produced by restriction. It was found that even when positioned within the *trpE* gene, an *EcoK* site had little effect

on its expression. This is in marked contrast to a similarly placed EcoRI site (a type II enzyme known to cut within its recognition sequence) which destroys expression on infection of an EcoRI restricting host. However, when two K sites are positioned such that the trpE gene lies between them, K-restriction inhibits expression, implying that cutting occurs preferentially between them (Brammar et al, 1974).

E) DNA Recognition

The DNA sequences recognized by type I restriction and modification enzymes are of an unusual but characteristic structure (Bickle, 1987). They are asymmetric and bipartite, containing a central region of non-specific nucleotides bounded by short defined regions of three bp 5' and 4 or 5 bp 3' to the spacer: for example, EcoK recognizes 5' AAC(N6)GTGC. A complete list of known target sequences is shown in Figure 11.

One adenine in each defined component is the substrate for methylation (Kuhnlein and Arber, 1972; Vovis and Zinder, 1975; Von Ormondt et al, 1973; Roy and Smith, 1973) which occurs at its N6 position. One of these is on each strand of the DNA, and they are nine or ten nucleotides apart. They could therefore be approached in two successive major grooves on one face of the DNA (Nagaraja et al, 1985b). The spacer varies from six to eight bp and would be tucked away in the minor groove between the two defined components.

FIG. 11. Recognition Sequences of Type I Restriction Endonucleases

*Eco*K	A A C N N N N N N G T G C T T G N N N N N N C A C G	Kan *et al.* (1979)
*Eco*B	T G A N N N N N N N N T G C T A C T N N N N N N N N A C G A	Ravetch *et al.* (1978) Lautenberger *et al.* (1978)
*Eco*D	T T A N N N N N N N G T C Y A A T N N N N N N N C A G R	Nagaraja *et al.* (1985a)
*Sty*SB	G A G N N N N N N R T A Y G C T C N N N N N N Y A T R C	Nagaraja *et al.* (1985b)
*Sty*SP	A A C N N N N N N G T R C T T G N N N N N N C A Y G	Nagaraja *et al.* (1985b)
*Sty*SQ	A A C N N N N N N R T A Y G T T G N N N N N N Y A T R C	Nagaraja *et al.* (1985c)
*Sty*SJ	G A G N N N N N N G T R C C T C N N N N N N C A Y G	Gann *et al.* (1987)
*Eco*A	G A G N N N N N N N G T C A C T C N N N N N N N C A G T	Suri *et al.* (1984)
*Eco*DXXI	T C A N N N N N N N A T T C A G T N N N N N N N T A A G	Piekarowicz & Goguen (1986)
*Eco*R124	G A A N N N N N N N R T C G C T T N N N N N N N Y A G C	Price *et al.* (1987)
*Eco*R124/3	G A A N N N N N N N N R T C G C T T N N N N N N N Y A G C	Price *et al.* (1987)

* indicates methylated adenine residues.

Y indicates that either pyrimidine base may be present, and R either purine base.

The same basic pattern of target sequences is conserved between the different families (see Figure 11). Of those known so far it is noticeable that all three A-family members have seven bp spacers. Those of the K-family are six, seven or eight. In the R124 family N6 and N7 are found. Particularly interesting are EcoR124 and R124/3 whose defined components are identical and whose different specificities are therefore produced entirely because of their different spacer lengths (Price et al, 1987b). EcoK and StySP have the trimeric component in common while the tetrameric component recognized by StySP is a degenerate version of that for EcoK. This explains the observation (Bullas et al, 1980) that StySP modification protects a DNA molecule from EcoK restriction, but K-modification does not necessarily protect against SP restriction. StySQ has a hybrid target sequence comprising the trimeric component as recognized by StySP and pentameric component of StySB (Nagaraja et al, 1985a). The S polypeptide of StySQ is encoded by a recombinant gene formed by crossing over between those of StySP and SB (Fuller-Pace et al, 1984; Fuller-Pace and Murray, 1986). The significance of this will be discussed later.

The relatedness of enzymes of different specificity within the K-family (Murray et al, 1982) suggests that amino acid residues responsible for DNA recognition may be identified simply by a sequence comparison of their respective S polypeptides, this subunit being the one implicated in DNA recognition (Boyer and Roulland-Dussoix, 1969; Glover and Colson, 1969). Originally, Gough and Murray (1983) obtained

the nucleotide sequences of the S genes of EcoK, B and D. The hope was that, since each S polypeptide interacts with essentially identical M and R subunits, directing this complex to bind DNA and various other conserved co-factors and substrates, then the only variation in their predicted amino acid sequences would be associated with recognition of their different target sequences. The actual variation in the S polypeptides was far more extensive than expected (Gough and Murray, 1983); subsequent sequencing of the S genes of StySP (Fuller-Pace and Murray, 1986) and StySB (Gann et al, 1987) have shown these to conform to the pattern found for EcoK, B and D. We now believe that the extent of this variation reflects the complex nature of the recognition process and that the original expectation (Gough and Murray, 1983) was indeed correct. At the time, however, it was difficult to believe that so much variation would be needed to change the specificity of the target sequence. Figure 12 shows a schematic diagram of a generalized K-family S polypeptide. Indicated are the regions which are conserved or variable when S polypeptides of different specificity are compared (Gough and Murray, 1983). The polypeptides vary in length between 445 and 475 amino acid residues. The N-terminal 150 residues are referred to as the amino variable domain, and are encoded by the proximal variable region of the gene. The central conserved region is the next 35 amino acids. This is followed by the 150 residues of the carboxyl variable domain, encoded by the distal variable region. The C-terminal 80 amino acids represent and are encoded by the carboxyl conserved and distal conserved regions respectively.

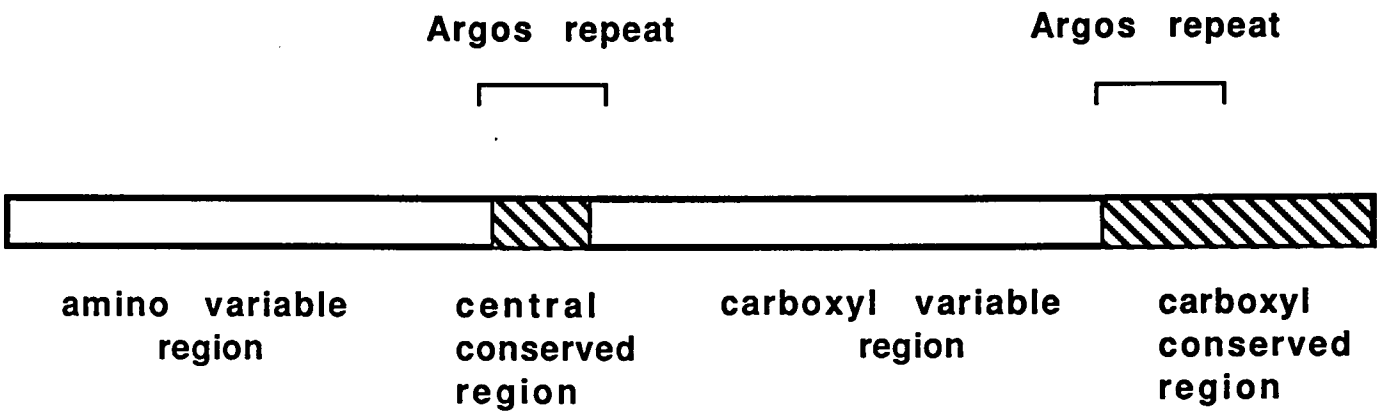


Fig 12. A schematic diagram of a K-family specificity polypeptide. The conserved regions are hatched; the variable regions are indicated as open segments. The positions of the repeated domains identified by Argos (1985) are shown.

One model originally put forward by Gough and Murray (1983) was that each variable domain represented, or contained, a recognition domain specifying one of the defined regions of the target sequence. Alternatively, they suggested that areas within the conserved regions may be the DNA recognition domains and perhaps the variable regions, different versions of which appeared as unlike one another as they are randomly selected sequences (e.g. ØX174), may actually be functionally unimportant, and hence under no sequence constraint. Subsequently, Argos (1985) proposed a model which again implicated the conserved regions in DNA recognition. His model was based on the observation that the central conserved and first half of the carboxyl conserved regions are not only conserved between all the enzymes, but are, within a single S polypeptide, similar to one another (see Figure 12) (Argos, 1985; Gann et al, 1987; Fuller-Pace and Murray, 1986). This, in conjunction with the prediction that these regions adopt a mainly α-helical structure, led him to suggest that an S polypeptide interacts with its bipartite recognition sequence as a pseudodimer, in a manner analogous to a repressor dimer binding to its symmetric operator via a helix-turn-helix domain (Pabo and Sauer, 1984). The different specificities were produced by the small number of amino acid differences between these mainly conserved regions.

The recombinant specificity StySQ is encoded by a gene produced by crossing over between the central conserved regions of the S genes of StySP and SB (Bullas et al, 1976; Fuller-Pace

et al, 1984). Its proximal variable region and first half of its central conserved region come from StySP, while the rest of the molecule originated from StySB (Fuller-Pace and Murray, 1986). The target sequence, as shown above, has the trimeric component as recognized by StySP and pentameric component of StySB (Nagaraja et al, 1985a and b). Thus the recombination event reassorted two independent DNA recognition domains within the polypeptide, each involved in recognition of one defined component of the target sequence (Fuller-Pace et al, 1984; Nagaraja et al, 1985a). The position of the crossover in the middle of the central conserved region makes the Argos model (1985) somewhat less appealing in that it limits the residues within this region that could be involved in specifying the trimeric component of the target sequence to those that occur to the left of this crossover. In the case of StySQ, there are only four residues within the central conserved region that originate from StySP and are different from the corresponding residues in StySB (Fuller-Pace and Murray, 1986). For the Argos model to be correct, these four amino acid residues would have to be responsible for determining the different trimeric components of StySP and SB (AAC and GAG respectively; see Figure 11).

However, circumstantial evidence that the variable domains are involved in defining specificity comes from the observation that the target sequences of EcoK and StySP both contain the trimeric component 5'AAC (Kan et al, 1979; Nagaraja et al, 1985b). This correlates with the fact that their S

polypeptides have very similar amino variable regions (Fuller-Pace and Murray, 1986).

CHAPTER 3 : MATERIALS AND METHODS

1) Strains

A) Bacterial Strains

See Table 1a.

B) Phage Strains

See Table 1b.

2) Enzymes and Chemicals

DNA polymerase (Klenow fragment) and T4 DNA ligase were purchased from Boehringer; DNA polymerase I from NBL Enzymes; restriction endonucleases from Boehringer, New England Biolabs, or NBL Enzymes; DNase I, RNase A and lysozyme were all from Sigma Chemical Company Ltd. T4 DNA polynucleotide kinase was a gift from S.A. Bruce (Edinburgh).

M13 sequencing primer (17-mer) and M13 hybridization probe primer were purchased from New England Biolabs; other synthetic oligonucleotides were from Oswel DNA Service (Edinburgh); deoxynucleoside triphosphates and dideoxynucleoside triphosphates from Boehringer.

Deoxyadenosine 5'-[α-^{32}P] triphosphate and deoxycytidine 5'-[γ-^{32}P] triphosphate were purchased from Amersham

Table 1a: Bacterial Strains

Strain Number	Specificity	Relevant Features	Reference/Source
NM522	K	(lac-pro) hsdMS 5 F'lacZ M15 lacI^q	Gough and Murray, 1983
BMH71-18	K	(lac-pro) hsdMS 5 F'lacZ M15 lacI^q	Gronenborn et al, 1976
		MutL derivative of BMH71-18	Kramer et al, 1984
NM661	B	hsdB genes in NM522	N. E. Murray
NM555	A	hsdMS 2 derivative of WA2899	Fuller-Pace et al, 1985
NM550	SB	hsdSB 9	Fuller-Pace et al, 1984
NM551	SQ	hsdSQ derivative of NM550	Fuller-Pace et al, 1984
AG1	SJ	hsdSJ derivative of NM550	Gann et al, 1987
AG2	SJ	F'kan^r derivative of AGI	Gann et al, 1987
AG3	SB	F'kan^r derivative of NM550	Gann et al, 1987
ED8689	K	hsdR, Phi80^s (for phage crosses)	Wilson et al, 1977
EH55	K	asn F'kan^r (source of F'kan^r)	Hansen et al. 1983
WA2574	K	ptsM (Pel$^-$) hsdS	Elliot and Arber, 1978
WA 2899	A	hsdA genes in E.coli K-12	W. Arber
NM490	B	hsdR derivative of C3000	N. E. Murray
feb 10	K	pohA10 (check for λred)	Zissler et al, 1971
NS377		nusA1 rpoB (check for nin)	Sternberg, 1976
L4001	SB	hsdSB genes in E.coli K-12	Bullas and Colson, 1975
L4002	SP	hsdSP genes in E.coli K-12	Bullas and Colson, 1975

Table 1b: Phage strains

Strain Number	Relevant Features	Reference/Source
NM63	λ*c*I26	N. E. Murray
NM143	λ*h*^{80} *imm*^{21} *nin*$^{+}$ (for crosses)	N. E. Murray
NM144	λ*h*^{82} *b*522 *imm*$^{\lambda}$ *c*I	N. E. Murray
NM243	λ*vir*	N. E. Murray
NM507	λ*imm*^{21} *c*I	N. E. Murray
NM675	λ*h*^{80} *att*^{80} *c*I857 *nin*5	N. E. Murray
NM848	λ*h*^{82} *b*522 *imm*^{21} *c*I	N. E. Murray
NM1048	λ*hsd*K genes in NM781	Sain and Murray, 1980
NM1183	λ*hsd*SB genes in NM762	Fuller-Pace *et al*, 1984
NM1185	λ*hsd*SP genes in NM762	Fuller-Pace *et al*, 1984
NM1201	λ*hsd*SQ genes in NM762	Fuller-Pace *et al*, 1984
NM1290	Δ10 derivative of NM1183	N. E. Murray
NM1291	*imm*^{21} *nin*$^{+}$ derivative of NM1290	Gann *et al*, 1987
NM1292	*hsd*SJ derivative of NM1291	Gann *et al*, 1987
NM1293	*imm*$^{\lambda}$ *c*I857 *nin* derivative of NM1292	Gann *et al*, 1987
P3	Phage sensitive to *Sty*SQ restriction	Bullas *et al*, 1976
M13 mp18	Vector for DNA sequencing	Yanish-Perron *et al*, 198
M13 mp19	Vector for DNA sequencing	Yanish-Perron *et al*, 198

International; deoxyadenosine 5'-[α-^{35}S] thiotriphosphate from New England Nuclear.

Acrylamide and bis-acrylamide were supplied by BDH Ltd; TEMED and LOW e.e.o. agarose were from Sigma; standard agarose from Miles Laboratory Ltd; ethidium bromide from BDH.

Ampicillin (Penbritin) and kanamycin were purchased from Beecham Pharmaceuticals; vitamin B_1, DTT, 2-mercaptoethanol and IPTG were all from Sigma; X-gal was from Boehringer; nitrocefine was from Glaxo.

Nitrocellulose filters were purchased from Schleicher and Schuell; HP5 film from Ilford; Cronex intensifier screens and X-ray film from Du Pont Ltd.

3) Media

L-Broth: 10g Difco Bacto tryptone, 5g Difco Bacto yeast extract, 10g NaCl, distilled H_2O to 1 litre; adjusted to pH 7.2 with NaOH before autoclaving.

L-Agar: 10g Difco Bacto tryptone, 5g Difco Bacto yeast extract, 10g NaCl, 15g Difco agar, distilled H_2O to 1 litre; adjusted to pH 7.2 with NaOH before autoclaving.

BBL-Agar: 10g Baltimore Biological Labs. trypticase, 5g NaCl, 10g Difco agar, distilled H_2O to 1 litre.

BBL Top Agar: as for BBL agar but only 6.5g Difco agar added per litre.

Minimal Agar: 4g Difco agar, distilled H_2O to 300ml. After autoclaving the following sterile solutions were added: 80ml 5x Spizizen salts, 4ml 20% glucose, 0.1ml vitamin B_1 (2mg/ml).

5x Spizizen Salts: 10g $(NH_4)_2SO_4$, 70g K_2HPO_4, 30g KH_2PO_4, 5g tri-sodium citrate dihydrate, 1g $MgSO_4.7H_2O$, distilled H_2O to 1 litre.

M9-Maltose Medium: 250ml 4x M9 salts, 15ml 20% maltose, 1ml 1M $MgSO_4.7H_2O$, distilled H_2O to 1 litre.

4x M9 Salts: 28g Na_2HPO_4, 12g KH_2PO_4, 2g NaCl, 4g NH_4Cl, distilled H_2O to 1 litre.

Phage Buffer: 3g KH_2PO_4, 7g Na_2HPO_4, 5g NaCl, 10ml 0.1M $MgSO_4.7H_2O$, 10ml 0.01M $CaCl_2$, 1ml 1% (w/v) gelatin, distilled H_2O to 1 litre.

LTB Buffer: Storage buffer for M13 phages. 0.5M Tris-HCl pH 7.5, 0.1M $MgCl_2$, 0.1M DTT.

Antibiotics: Antibiotics were used at the following concentrations: Ampicillin, 100μg/ml; Kanamycine 25μg/ml. When used in plates the antibiotic was added to molten agar immediately prior to pouring.

Xgal Indicator Plates: 20μl Xgal (30mg/ml), 20μl IPTG 20mg/ml, per 2.5ml of BBL top agar.

All media were sterilized by autoclaving at 15lb in^{-2} for 15 minutes.

4) Standard Solutions

TE Buffer: 10mM Tris, 1mM EDTA; adjusted to appropriate pH with HCl.

20x SSC: 3M NaCl, 0.3M tri-sodium citrate.

10x TBE Buffer: 890mM Tris, 890mM boric acid, 25mM EDTA.

Ethidium Bromide: 10mg/ml in distilled H_2O. Stored at 4° C, protected from light.

Non-Solubilizing Scintillation Fluid: 4g butyl-PBD made up to 1 litre in toluene.

5) Microbial Techniques

A) Preparation of Plating Cells

A fresh overnight culture was diluted 20-fold in L-broth and grown at the required temperature to mid-logarithmic phase. The cells were pelleted by spinning in a bench centrifuge at 2,000g_n for 5 minutes and resuspended in half the original volume of 10mM $MgSO_4$ before storage at 4°C.

B) Preparation of λ Plate Lysates

A single plaque was picked into 1ml of phage buffer containing a drop of chloroform and mixed. After addition of 0.1ml of the phage suspension to 0.1ml of plating cells the phage were left to adsorb to the cells for 15 minutes. BBL top agar (3ml) was added and the mixture poured onto a fresh L-agar plate. The plate was incubated (without inversion) at the

required temperature until confluent lysis was observed, usually after 6-8 hours. Approximately 3ml of L-broth were added to the plate before storage at 4°C overnight. The L-broth was decanted and a few drops of chloroform added. Cell debris was pelleted by centrifugation in a bench centrifuge at 2,000g$_n$ for 10 minutes.

C) Phage Titration

Serial dilutions of the phage stock were made in phage buffer before mixing 0.1ml of phage suspension with 0.1ml plating cells, and leaving to adsorb for 15 minutes. The mixture was plated out in 2.5ml BBL top agar onto BBL plates and incubated overnight at 37°C.

D) Spot Tests

A lawn of cells was prepared by adding 0.1ml of plating cells to 2.5ml of BBL top agar for plating out on a BBL plate, and 10μl aliquots of the phage dilutions were spotted onto the lawns. The spots were allowed to dry before incubation at 37° C, overnight.

E) Preparation of CsCl Purified Phage

A fresh overnight culture of the host bacterium was diluted 50-fold into 100ml of L-broth supplemented with 10mM $MgSO_4$, and grown at 37°C with good aeration until they reached an O.D.$_{650}$ of 0.5 (i.e. 2 x 10^8 cells/ml). Phage were added

to a m.o.i. of 0.2 and incubation at 37°C was continued. The turbidity of the culture was periodically measured until lysis occurred, usually 2-4 hours later. Chloroform (0.2ml) was added and the flask shaken at 37°C for a further 15 minutes. The lysate was clarified by centrifugation, and phage precipitated by polyethylene glycol (PEG) essentially as described by Yamamota et al (1970): sodium chloride (4% w/v) was added, followed by DNAsel and RNAsel, both to 1μg/ml. After standing at room temperature for 1 hour, 10g of PEG 6000 was added and allowed to dissolve. The lysate was left at 4°C overnight. The PEG precipitate was recovered by centrifugation at 10,200g$_n$ for 10 minutes and resuspended in 5ml of phage buffer by swirling gently at 4°C for 2-3 hours. Debris was removed by centrifugation at 2,000g$_n$ for 5 minutes before concentration of the phage on a CsCl step gradient.

Step gradients of CsCl (Thomas and Abelson, 1966) were prepared in 14ml polycarbonate tubes: 1.5ml of 31% w/w CsCl solution in phage buffer was pipetted into the tube and underlaid with 1.5ml of 45% w/w CsCl solution; finally these two steps were underlaid with 1.5ml of 56% w/w CsCl solution. The phage solution was overlayed on to the gradient and centrifuged at 140,000g$_n$ for 2 hours at 20°C. The phage band was collected by piercing the tube using a 21 gauge needle and a syringe. The resulting lysate was dialysed at 4°C against phage buffer to remove the CsCl, and stored at 4°C.

F) Construction of λ Lysogens and Dilysogens

Fresh plating cells were infected with the appropriate phage (or phage and heteroimmune helper phage) at a m.o.i. of 1-2 and allowed to adsorb. The cells were diluted 50-fold in L-broth and grown for 2-3 hours at the appropriate temperature. The resulting culture was serially diluted in L-broth and plated on L-agar plates in the presence of 10^9 p.f.u. each of two homoimmune, cI^- phages of different host ranges. Colonies which grew after overnight incubation were purified and tested for lysogeny (i.e. sensitivity to λ vir, but not λ cI).

G) Genetic Manipulation of hsd Genes

The hsdRMS genes encode type I R-M systems. In the chromosome they occur as three adjacent genes. They can be cloned into bacteriophage λ, producing λhsd phages. If such phage include at least the hsdM and S genes they are able to modify themselves while growing lytically, even in an m- host. Under some circumstances, phage encoded M and S polypeptides can also interact with R polypeptides encoded by the host cell, thereby producing a fully functional restriction enzyme (Fuller-Pace et al, 1985).

The hsd genes can be transferred between bacterial and phage chromosomes via homologous recombination. In this study, the hsdSJ genes were moved from the λhsdSJ phage to the chromosome of a cell (NM550) which contains hsdSB DNA (Fuller-Pace et al, 1984). The λhsdSJ phage is att-, and so cannot

integrate into the host chromosome by site-specific recombination. However, it can integrate into the hsd region via homologous recombination. Lysogens of this type were made, using an imm$^{\lambda}$ cI857 derivative of the original λhsdSJ phage. Cured derivatives of the lysogenic strain were isolated by virtue of their no longer being ts. These were then screened for the presence of the StySJ R-M system determinants, which are left in the chromosome when the prophage excises taking with it the remnants of the chromosomal hsdSB genes in place of hsdSJ.

H) Bacterial Conjugation

Strain EH55, the F′Kan donor, was grown to O.D.$_{650}$ 0.5 at 37°C. 5ml of this was then mixed with 5ml of an overnight culture of the recipient strain, and incubated at 37°C for 1 hour without shaking. The mixture was then serially diluted in minimal medium. Recipient cells containing the F′ were then selected by plating out on minimal plates containing kanamycin (25 g/ml). EH55 cannot grow on minimal medium, and the recipient strain is kanamycin sensitive if it has not got the F′.

I) Phage Crosses

Freshly prepared plating cells were co-infected at a m.o.i. of 5 of each of the parental phages. After 15 minutes adsorbtion at room temperature the infected cells were diluted 100-fold in pre-warmed L-broth and grown at 37°C, with

aeration, for 1.5 hours. Chloroform was added and, after centrifugation to remove debris, the supernatant was titred on a permissive host for total progeny and on a selective host for the required recombinant. Single plaques were picked and tested as appropriate. After purification by single plaque isolation, phage stocks were prepared as described earlier.

6) DNA Techniques

A) Ethanol Precipitation of DNA

DNA in solution was precipitated by addition of 0.1 volumes of 3M sodium acetate and 2 volumes of ethanol. The DNA was sedimented by centrifugation at 10,000g$_n$ for 10 minutes, washed with 70% v/v ethanol, and repelleted. The pellet was dried under vacuum and resuspended in the appropriate volume of TE buffer pH 7.5.

B) Preparation of DNA from Phages

High titre lysates were prepared by CsCl gradient as described above and, after collection of the phage band, the CsCl was removed by dialysis against TE pH 8.0. The phage protein was extracted 3 times with an equal volume of phenol pre-equilibrated with TE: the phenol and aqueous phases were mixed gently by tube inversion, and then separated by centrifugation at 5,000g$_n$. The lower phenol layer was removed and discarded. The DNA was dialysed against TE at 4°C for 24 hours, with several buffer changes. The concentration of the

DNA was determined by measuring the O.D. at 260nm (an O.D. of 1 is equivalent to 50μg/ml).

C) Large-Scale Preparation of Plasmid DNA

This method is based on that of Clewell and Helinski (1969). A fresh overnight culture of the plasmid-containing cells was diluted 100-fold into 150ml of L-broth containing the appropriate antibiotic, and grown overnight at 37°C, with aeration. The cells were harvested (6,500g$_n$ for 10 minutes), resuspended in 7ml of lysis solution, and left on ice for 5 minutes. 14ml of alkaline SDS was added, followed by a 10 minute incubation on ice. After addition of 10.5ml of 3M potassium acetate pH 4.8, and a further 5 minutes on ice, the precipitated protein, dodecyl sulphate, and chromosomal DNA was removed by centrifugation at 6,500g$_n$ for 10 minutes at 4°C. The supernatant was poured through glass wool to remove any remaining precipitate. Plasmid DNA was precipitated by addition of 15ml of isopropanol, and pelleted by centrifugation at 6,500g$_n$ for 10 minutes. The pellet was washed with 70% ethanol, pelleted, and dried under vacuum for 30 minutes. The DNA was dissolved in TE pH 7.5 to a volume of 9.4ml, and then CsCl (to 0.95g/ml-) and ethidium bromide (to 0.6mg/ml-) were added. The final density of the solution should be 1.55g/ml- . The CsCl solution was transferred to a 10ml "quick-seal" polyallomer tube and centrifuged at 90,000g$_n$ for 48-60 hours at 18°C. Two bands were visible under UV light: the upper band consisted of nicked and linearized plasmid DNA and fragmented chromosomal DNA, and the lower band of supercoiled plasmid DNA

which was removed using a 21 gauge hypodermic needle inserted through the side of the tube. Ethidium bromide was removed by 4 extractions with isopropanol saturated with NaCl-saturated TE. Two volumes of H_2O were added to the (lower) aqueous phase before the DNA was precipitated with ethanol. The DNA was dissolved in 500µl TE pH 8.0, and any residual protein was extracted twice with phenol equilibrated with TE. The DNA in the aqueous phase was ethanol precipitated and redissolved in 500µl of TE pH 8.0. The concentration of DNA was determined by measuring the O.D. at 260nm.

Lysis Solution: 25mM Tris-HCl pH 8.0, 10mM EDTA pH 8.0, 1% glucose.

Alkaline SDS: 0.2M NaOH, 1% SDS.

D) Rapid Large Scale Preparation of Plasmid DNA

A fresh overnight culture of the plasmid containing strain was diluted 100-fold in 50ml of L-broth containing the appropriate antibiotic and grown overnight at 37°C with aeration. The cells were harvested (5,000gn for 10 minutes), resuspended in 3.5ml of lysis solution and put on ice. 8mg of lysozyme, dissolved in 0.5ml of lysis solution, was added to the cells, and the mixture left on ice for 10 minutes. After addition of 8ml of freshly prepared alkaline SDS solution and gentle mixing, the mixture was left on ice for a further 20 minutes. 5ml of 3M sodium acetate, pH 5.2, was then added with gentle stirring and, after a further 10 minutes on ice, the precipitated protein and chromosomal DNA was removed by

centrifugation at 10,000g$_n$ for 15 minutes at 4°C. Remaining protein was then extracted from the supernatant with phenol/chloroform, and DNA precipated with ethanol. After drying, the DNA was resuspended in 0.5ml H_2O, 5μl of RNase (10mg/ml) was added, and the mixture incubated at 37°C for 20 minutes. Protein was then extracted with phenol/chloroform; this was repeated three or four times until the interface was clean. Following ethanol precipitation, the plasmid DNA was resuspended in 100-500μl of TE pH 7.5.

E) Preparation of M13 Replicative Form (RF) DNA

A 100-fold dilution of a fresh overnight culture of NM522 was grown to an O.D.$_{650}$ of 0.2. A single M13 plaque was picked into 1.5ml of cells and the culture was shaken at 37°C for 5-6 hours. The culture was transferred to an Eppendorf tube and spun for 5 minutes at 11,600g$_n$ in a microcentrifuge to pellet the cells. The supernatant was titred (titres were 10^{11} p.f.u. ml) and used to infect a 50ml culture of early log phase cells at a final concentration of 10^4 p.f.u./ml . The culture was grown with aeration at 37°C for 16-18 hours before the cells were pelleted and the supernatant titred.

An overnight culture of NM522 was diluted 100-fold into 500ml of L-broth and grown to an O.D. of 0.1. Phage were added to 10^{11} p.f.u./ml and the culture was grown for a further 2 hours at 37°C. The cells were sedimented by centrifugation at 6,500g$_n$ for 10 minutes and the RF DNA was

prepared as described for the purification of plasmid DNA (Section C).

F) Plasmid "Miniprep"

(Ish-Horowitz and Burke, 1981)

An overnight culture was harvested in a microcentrifuge tube at 11,600gn for 5 minutes, and the cells resuspended in 100µl of lysis solution. After incubation for 5 minutes at room temperature, 200µl of alkaline SDS was added gently mixed and left on ice for 5 minutes. Precooled 3M sodium acetate (150µl) was added, mixed gently, and the tube returned to ice for a further 5 minutes. The resulting precipitate was removed by a 5 minute centrifugation at 11,600gn and the DNA in the supernatant was precipitated with ethanol. The DNA pellet was dissolved in 50µl of TE pH 8.0.

Solutions: See Section C.

G) Restriction Endonuclease Digestion of DNA

Digestion of DNA with restriction enzymes was normally carried out in a volume of 20µl containing 0.5-1µg of DNA, under conditions recommended by the suppliers. Reactions were stopped after incubation at 37°C for 2 hours by phenol extraction. The DNA was resuspended in an appropriate volume of TE pH 8.0.

H) Ligation of DNA

DNA was ligated using T4 DNA ligase, in a volume of 10μl containing 50mM Tris-HCl pH 7.5, 10mM $MgCl_2$. 0.2mM spermidine, 10mM DTT, 1mM ATP, 25-150ng DNA, and 1-2 Weiss units of T4 DNA ligase. Incubation was at 16°C overnight, or at 22°C for 2 hours.

I) Agarose Gel Electrophoresis

The concentration of agarose varied, depending on the size of fragments, between 0.7 and 1.3% w/v. Fragments of DNA were analysed by separation on agarose gels in 1 x TBE buffer. DNA samples (usually 0.2-0.5μg) were loaded mixed with 3μl of 5x Ficoll loading dye. Electrophoresis of gels was carried out at either 11V cm^{-1} for 2 hours, or 1.25V cm^{-1} for 20 hours. The DNA was visualized over a long-wave UV light transilluminator after staining for 20 minutes in a 1μg/ml solution of ethidium bromide and destaining in distilled H_2O for 20 minutes.

5x Ficoll Loading Dye: 20% Ficoll 400 in H_2O, with bromophenol blue dye.

J) Isolation of DNA Fragments from Agarose Gels

The region of the gel containing the band was cut out using a scalpel and placed in dialysed tubing, closed at each end, and containing 0.5ml TE pH 7.5. The DNA was eluted from the agarose by electrophoresis at 10V cm^{-1} for ~20 minutes.

Reversing the direction of electrophoresis for ~10 seconds released the DNA from the sides of the dialysis tubing. The TE was then placed in an Eppendorf tube, and ethidium bromide removed by extracting with TE saturated butan-1-ol. Protein was extracted once with TE saturated phenol, the DNA precipitated with ethanol, and then resuspended in an appropriate volume of TE.

K) Transfection and Transformation of Competent Cells

Cells were normally made competent for the uptake of DNA using a modification of the procedure of Mandel and Higa (1970). A fresh overnight culture was diluted 50-fold and grown, with aeration, at 37 C to an $O.D._{650}$ of 0.7. The cells were harvested at 2,000g_n for 5 minutes at 4°C and resuspended in an equal volume of 100mM $MgCl_2$. The cells were spun again, and resuspended in a half volume of 100mM $MgCl_2$. After pelleting for a third time, the cells were resuspended in a tenth volume of 100mM $CaCl_2$. DNA was added to 200µl of competent cells in a 5ml glass tube. After 10-30 minutes on ice the cells were "heat-shocked" at 42°C for 2 minutes. For transfection, the cells were then plated out in 2.5ml of BBL top agar and incubated at 37°C overnight. Transformation of cells with plasmid DNA required the addition of 1ml of L-broth to the tube after heat-shock, and incubation at 37°C for 1 hour to allow expression of antibiotic resistance. Aliquots of 10µl and 100µl were spread on L-agar plates containing the appropriate antibiotic, and incubated at 37°C overnight.

L) In Vitro Packaging

In vitro packaging mixes, namely Freeze Thaw Lysate (FTL) and Sonicated Extract (SE) were kindly donated by Heather Houston.

The packaging reaction mixture was prepared by adding reagents in the following order:

Buffer A	7µl
DNA	1-2µg (in maximum of 5µl)
Buffer M1	1µl
SE	6µl
FTL	10µl

The mixture was incubated at 25°C for 2 hours and subsequently diluted with 0.5ml of phage buffer. The number of phage produced was tested by standard titring.

Buffer A: 20µl 1M Tris-HCl pH 8.0; 2µl 0.5M EDTA pH 7.5; 3µl 1M $MgCl_2$; 975µl H_2O.

Buffer M1: 6µl 0.5M Tris-HCl pH 7.5; 300µl 100mM Putrescine / 50mM Spermidine 9µl 1M $MgSl_2$; 75µl 0.1M ATP; 1µl β-mercaptoethanol; 110µl H_2O.

M) Transfer of DNA from Plaques to Nitrocellulose

(Benton and Davies, 1977)

Phage recombinants were plated in BBL top agar on dry agar plates. After incubation at 37°C overnight the plates were cooled at 4°C for 1 hour to prevent damage to the top agar during transfer. A nitrocellulose filter was placed on the

agar and left for 1 minute. The filter was removed and placed, plaque side uppermost, on blotting paper, saturated with denaturation buffer, for 2 minutes. The filter was transferred to a beaker containing neutralization buffer for 2 or 3 minutes, rinsed briefly in 2x SSC, and blotted dry before baking at 80°C under vacuum for 2 hours.

Denaturation Buffer: 0.5M NaOH, 1.5M NaCl.

Neutralization Buffer: 0.5M Tris-HCl pH 7.4, 3M NaCl.

N) Radiolabelling of Double-Stranded Probes by Nick-Translation and Hybridization to Filters

Deoxycytidine 5'-[α-^{32}p] triphosphate (10μCi) was added to 20μl of 1x dNTP buffer, 1μl of DNAse I (2 x 10^{5}mg/ml-), 1μl DNA Polymerase I (1 unit/μl-) and 0.5-1.0μg of DNA (in ~2μl). After incubation at 16°C for 1-3 hours the reaction was terminated by the addition of 100μl of 10mM EDTA pH 8.0, and loaded onto a column of Sephadex G-50 equilibrated with TE buffer. The DNA was eluted with TE and collected as the first peak of radiolabel (detected using a mini-monitor) in a volume of 0.5-1.0ml. Samples of 1μl on Whatman GF/C discs were dried and the activity was counted in a non-solubilizing scintillant to determine the amount of label incorporated.

Filters were prehybridized in 50ml of hybridization buffer for 30 minutes at 37°C. The radiolabelled DNA (10^{6} cpm per filter) was added to 250 g of sonicated calf thymus DNA, denatured at 95°C for 10 minutes, and immediately cooled on

ice. The probe was added to the filter in 10ml of hybridization buffer and the hybridization was carried out at 37 C with gentle agitation, overnight. The filter was washed twice in 2x SSC. 0.1% SDS for 30 minutes at 37°C, and then twice in 1x SSC, 0.1% SDS for 30 minutes at room temperature. Finally the filter was rinsed in 1x SSC, blotted dry, and placed between two sheets of plastic film. Hybridization of the probe to the filter was detected by autoradiography at -70°C.

Nick Translation Buffer: 210mM Tris-HCl pH 7.5, 21mM $MgCl_2$, 20μg/ml BSA; stored in aliquots at -20°C.

1x dNTP Buffer: 250μl nick translation buffer, 10μl 2mM dATP/dTTP/dGTP, 2.5μl 2-mercaptoethanol, 737.5μl distilled H_2O; stored at -20°C in 250μl aliquots.

20x Denhardt's Solution: 0.4% polyvinylpyrolidone, 0.4% w/v Ficoll 400, 0.4% w/v BSA.

Hybridization Buffer: 50% formamide, 4x SSC, 1x Denhardt's Solution.

O) Radiolabelling of Single-Stranded M13 DNA and Hybridization to Filters

(Hu and Messing, 1982)

M13 reverse primer (1μl) was added to 5μg (5μl) M13 single-stranded template DNA. 1μl of 10x annealing mixture was boiled for 3 minutes and slowly cooled (over 15-30 minutes) to room temperature. After addition of 10μCi deoxycytidine 5'-[α-

^{32}P] triphosphate, 1μl each of 500mM dATP, dGTP and dTTP, 1μl Klenow polymerase (1 unit/μl) and 6μl H_2O, the reaction was incubated at 15°C for 90 minutes. The reaction was terminated by the addition of 100μl of 10mM EDTA.

Filters were prehybridized for several hours at 65°C in a solution containing, 5x SSC, 50μg/ml denatured, sonicated calf thymus DNA, and 0.1% SDS. A half volume of the probe was added to the prehybridization buffer and hybridization was carried out at 65°C overnight. The filter was washed twice in 1x SSC, 0.1% SDS for 30 minutes (the first wash at 65°C, the second at room temperature), and then twice in 0.5x SSC, 0.1% SDS for 30 minutes at room temperature. Filters were placed between two sheets of plastic film, and the hybridization of the probe to the filter was detected by autoradiography at -70°C.

P) Filling Recessed 3' Ends of Double Stranded DNA

Approximately 1μg of DNA was added to 2μl of 10x Nick-Translation buffer. 2nmol of each of the nucleotide triphosphates were added and the volume made up to 25μl. After addition of 1U of Klenow polymerase, the reaction was incubated at room temperature for 30 minutes. The reaction was stopped by adding 1μl of 0.5M EDTA.

7) Dideoxy Chain Termination Sequencing of DNA

(Sanger et al, 1977 and 1980)

A) Single-Stranded Template DNA Preparation from M13 Lysates

A single plaque was picked into 1.5ml of a 100-fold dilution of an overnight culture of NM522 in a 10ml glass tube. The culture was grown with vigorous shaking at 37°C for 5.5-6 hours, and then transferred to an Eppendorf tube and clarified by centrifugation at 11,600g$_n$ for 5 minutes. The supernatant was transferred to a clean tube and 200μl of PEG/NaCl solution was added. After 20 minutes at room temperature (or overnight at 4°C) the phage were pelleted by centrifugation at 11,600g$_n$ for 10 minutes. The supernatant was discarded, the tube respun briefly and any residual PEG solution removed with tissue paper. The pellet was dissolved in 100μl of TE and extracted with 50μl of TE-equilibrated phenol. The aqueous layer was transferred to a clean tube and the DNA was precipitated with ethanol, dissolved in 30μl of TE buffer pH 8.0, and stored at -20°C.

PEG/NaCl: 20% PEG 6000, 2.5M NaCl.

B) Dideoxy Chain Termination Sequencing Reactions

The DNA templates were annealed to M13 sequencing primer in a mixture containing 8μl of template DNA, 1μl (0.2pmol) 17-

mer primer, and 1 µl TM buffer. After incubation at 60°C for 1 hour the mixture was allowed to cool, 2 µl were dispensed into each of 4 Eppendorf tubes. The appropriate termination mix (2µl) was added to each well, and finally 2 µl Klenow polymerase mix was added:

Composition of Dideoxynucleotide Sequencing Reactions

Components	T	C	G	A
Template/primer	2	2	2	2
T mix	2	-	-	-
C mix	-	2	-	-
G mix	-	-	2	-
A mix	-	-	-	2
Klenow mix	2	2	2	2

: Quantities of components are given in µl.

A Hamilton repetitive dispenser was used to dispense all the reagents used in the sequencing reaction, and the 2 µl aliquots were placed on the sides of the tubes, thereby allowing all the reactions to be started simultaneously by spinning the tubes briefly in a microcentrifuge. After 30 minutes at room temperature 2 µl of sequencing chase mix was added to each tube. This was incubated at room temperature for 30 minutes. The reactions were stored at -20°C until required. Before loading on to a separating gel, 2 µl of formamide dyes was added to each sample and the tubes placed in boiling water for 10 minutes to allow denaturation of double-stranded DNA. Approximately 2 µl of the sample was loaded.

TM Buffer: 100mM Tris, 50mM $MgCl_2$; adjusted to pH 8.5 with HCl.

50mM Stock dNTP Solutions: 312mg/10ml dTTP. 296mg/10ml dCTP. 316mg/10ml dc GTP. 295mg/10ml dATP. All made up in distilled H_2O.

10mM Stock ddNTP Solutions: 61mg/10ml ddTTP. 58mg/10ml ddCTP. 62mg/10ml ddGTP. 62mg/10ml ddATP. All made up in distilled H_2O.

Chase Mix: 0.25mM dTTP, 0.25mM dCTP, 0.25mM dC GTP, 0.25mM dATP; made up in distilled H_2O from 50mM stocks.

Klenow Polymerase Mix (per clone): 4μCi [α-^{35}S]ATP, 1.5 units Klenow polymerase, 10mM Tris-HCl pH 8.5, 10mM DTT; made up to 9μl with distilled H_2O. The appropriate quantity of mix was made up immediately before dispensing into reactions.

Formamide Dyes: 100ml deionized formamide, 2ml 0.5M EDTA, 0.1g xylene cyanol FF, 0.1g bromophenol blue. Formamide was deionized by stirring with 2g of Amberlite MB-1 resin, and filtered before storage.

Chain Termination Mixes:

Composition of Dideoxynucleotide Chain Termination Reaction Mixes for Sequencing

Components	T Mix	C Mix	G Mix	A Mix
50mM dTTP	-	2.5	2.5	2.5
50mM dCTP	2.5	-	2.5	2.5
50mM dGTP	2.5	2.5	-	2.5
10mM ddTTP	15.0	-	-	-
10mM ddCTP	-	7.5	-	-
10mM ddGTP	-	-	15.0	-
1mM ddATP	-	-	-	7.5
0.5mM dTTP	12.5	-	-	-
0.5mM dCTP	-	12.5	-	-
0.5mM dGTP	-	-	12.5	-
TE buffer	500.0	500.0	500.0	500.0
Distilled H O	500.0	500.0	500.0	500.0

C) DNA Sequencing Gels

Gels were poured between 20 x 40cm glass plates. The plates were cleaned with ethanol, separated by 0.35mm Plastikard spacers and taped together using Sellotape thermosetting tape. A buffer gradient gel (Biggin et al, 1983) allowed at least 250 bases to be read from a clone. For each gel 7ml of 2.5x TBE gel mix (to which was added 14μl 25% AMPS and 7μl TEMED), and 40ml of 0.5x TBE gel mix (to which was added 70μl 25% AMPS and 35μl TEMED) were prepared. Using a 10ml pipette, 4ml of 0.5x TBE gel mix and then 6ml 2.5x TBE gel mix were taken up; 2-3 air bubbles were drawn through to create a gradient. This was poured between the clamped plates. The remaining 0.5x TBE gel mix was used to fill the space left as the plates were gradually lowered to the horizontal. Plastikard sharkstooth combs (Bethesda Research Laboratories) were used to form loading wells.

Samples were loaded with a drawn out plastic Gilson pipette tip. The gel was run at 25-30W in 0.5x TBE buffer for approximately 2.5 hours - until the bromophenol blue dye was within 3cm of the bottom of the gel. The notched plate was carefully prised off and the gel was fixed in a solution of 10% methanol, 10% acetic acid. It was then drained, transferred to damp blotting paper and covered with Saranwrap plastic film. The gel was dried on a vacuum gel drier at 80^{o}C. The Saranwrap was then removed and the gel placed in direct contact with X-ray film overnight at room temperature.

40% Acrylamide Stock: 38g acrylamide, 2g N,N'-methylene bisacrylamide; made up to 100ml in distilled H_2O and deionized by stirring with 5g Amberlite MB-1 resin. Filtered before storage at 4°C, protected from light.

0.5x TBE Gel Mix (per gel): 6ml 40% acrylamide, 1ml 20x TBE, 17g urea; made up to 40ml with distilled H_2O.

2.5x TBE Gel Mix (per gel): 1.5ml 40% acrylamide, 1.25ml 20x TBE, 4.25g urea, 2g sucrose, 0.5mg bromophenol blue; made up to 10ml with distilled H_2O.

Gel Fix: 10% methanol, 10% acetic acid.

8) Site-Directed Mutagenesis

A) Phosphorylation of 5' Ends of DNA with T4 DNA Polynucleotide Kinase

Polynucleotide Kinase (PK) was used to add 5' phosphate groups to unphosphorylated oligonucleotides which were then used in SDM reactions. Approximately 100pm of oligonucleotide were incubated in 1 x PK buffer with 1U of T4 DNA polynucleotide Kinase at 37°C for 30 minutes. The substrate providing phosphate was ATP at 20mM. The reaction was stopped by incubation at 65°C for 10 minutes.

PK Buffer: 3µl 1M tris-HCl (pH 8), 1.5µl 0.2M Mg Cl_2, 1.5µl 0.1M DTT, H_2O to 30µl total volume.

B) Labelling 5' Ends of Oligonucleotides with 5'-[γ^{32}P]-ATP using T4 DNA Polynucleotide Kinase

Adenosine 5'-γ-[^{32}P] Triphosphate (Amersham, 3000Ci/mMol, 50 ci) was used to label 100pm of oligonucleotide in the presence of 1 x PK buffer with 1U of PK at 37°C for 30 minutes.

C) Screening M13 Plaques by Hybridization with Mutogenic Oligonucleotides

Plaques arranged in asymmetric grids of approximately a hundred were blotted on to nitrocellulose filters for 1 minute. The filters were then baked at 80°C for 2 hours in a vacuum oven. They were then prehybridized in 20ml 6x SSC, 10x Denhardts and 0.2% SDS at 67°C for 1 hour. After rinsing in 50ml 6x SSC for 1 minute at room temperature, the filters were placed in hybridization solution (^{32}P labelled oligonucleotide, 7x 10^6 cpm/μg, 6x SSC and 10x Denhardts) and left at room temperature overnight.

The filters were washed at room temperature in 6x SSC for 10 minutes (with two changes of solution), dried and autoradiographed using X-ray films at -70°C for 1-6 hours. Using a 45mer oligonucleotide with 4 mismatches (3xA-C 1xG-T) to the wild type sequence, this single wash was sufficient to distinguish mutant and wildtype plaques.

D) Double Primer Mediated Site Directed Mutagenesis

This is essentially the method as described by Zoller and Smith (1983). The mutagenic oligonucleotide and recombinant M13 template are shown in Figure 20. The two primers, the mutagenic oligonucleotide and the universal sequencing primer, were 5' phosphorylated as described above. These were then annealed to the single-stranded recombinant M13 mp 18 template. 1pmol of template DNA, 10pmol of mutagenic oligonucleotide and 10pmol of universal M13 primer were added to 1μl of solution A in a total of 10μl. The mixture was heated to 100°C for 3 minutes, and then cooled slowly to room temperature.

10μl of solution C (containing T4 DNA ligase) was added to the annealed DNA, followed by 2.5U of Klenow polymerase. Following incubation at 15°C overnight, the DNA was used to transform BMH71-18, a repair deficient (mutL) strain. The transformed cells were then plated out on a lawn of NM522.

Solution A: 0.2M tris-HCl pH 7.5, 0.1M Mg Cl , 0.5M NaCl, 0.01M DTT.

Solution B: 0.2M tris-HCl pH 7.5, 0.1M Mg Cl .

Solution C: 1μL solution B; 1μL 10mM dCTP; 1μL 10mM dGTP; 1μL 10mM dATP; 1μL 10mM dTTP; 1μL 10mM ribo-ATP; 1μL 0.1M DTT; 1.5μL T4 DNA Ligase (2U/μL); To 10.5μL with H_2O.

CHAPTER 4 : RESULTS

1) Reassortment and Identification of DNA Recognition Domains within the Specificity Polypeptides

Introduction:

StySQ is a recombinant specificity system. Its specificity gene is the product of a recombination event between the central conserved regions of the S genes of StySP and SB (Fuller-Pace et al, 1984). Its target sequence is made up of the defined 5′ trimeric and 3′ pentameric components of the StySP and SB recognition sequences respectively (Figure 11; Nagaraja et al, 1985a and b). The implication is that each S polypeptide contains two independent DNA recognition domains, each involved in specifying one defined component of the target sequence. Whether it is the large variable regions within these polypeptides that are the recognition domains (Gough and Murray, 1983; Fuller-Pace and Murray, 1983) or whether in fact it is the small differences within the otherwise conserved regions that are the critical residues in defining specificity (Gough and Murray, 1983; Argos, 1985) is not known (see Chapter 2, E).

To clarify the situation, two experiments were designed. In the first, a second recombinant S gene was constructed whose target sequence should be predictable, based on the model of there being two DNA recognition domains. In the second, the potentially critical residues within the central conserved

region of the StySQ S gene were changed by site directed mutagenesis. In this way, an S gene was produced which encodes only the amino variable region of StySP, with the rest of the polypeptide being identical to that of StySB. If the amino variable region is entirely responsible for determining the trimeric component of the target sequence, then this polypeptide will have the same specificity as StySQ. If, alternatively, the specificity is that of StySB, then it would imply that the determinants of specificity occur within the first half of the central conserved region. A specificity different from either StySP or SB, or a non-functional polypeptide, would suggest that the alterations disrupted the protein generally, or that residues involved in DNA recognition are found in both the variable and conserved regions.

A) Construction of a Recombinant Specificity Gene

The recombinant I chose to make was of reciprocal structure to StySQ; i.e. containing the proximal half from the S gene of StySB and the distal half from StySP. The resultant S gene was predicted to encode a polypeptide that would specify recognition of a sequence comprising the trimeric component of the StySB target and the tetramer from StySP: GAG(N_6)GTRC (see Recognition Sequence Figure 11 in Chapter 2, E).

The new specificity gene was produced by in vivo recombination between a λ phage carrying the proximal half of the S gene of StySB including the central conserved region, and a plasmid containing the central conserved region and distal

half of StySP. The frequency of recombinants was expected to be very low as their formation requires a double crossover, one of which must be within a 70 bp stretch of the central conserved regions. Previous experiments in similar systems, but where recombinant production was simply screened for, failed (A. Gann and F. Fuller-Pace, unpublished observations), as did repetition of the type of transduction experiments from which StySQ arose (Bullas pers. comm.). Therefore, to facilitate the isolation of the desired recombinant, a starting phage was used such that recombination events that generate the complete S gene also increase the length of the phage genome. Enrichment for phage with this linked characteristic would therefore increase the level of recombinant phage within the population.

The starting phage, λ hsdSBΔ10, is a deletion derivative of the λ hsdSB phage which carries on 11 kb HindIII fragment encoding the hsdM and S genes of the StySB system (see Figure 13a) (Fuller-Pace et al, 1984). The deletion (Δ10) removes about 5 kb of DNA extending from within, or just distal to, the central conserved region of the S gene to somewhere between the downstream BamHI and EcoRI sites (Figure 13). The plasmid, pAG2, includes an 8.5 kb HindIII fragment from the hsdSP region. This insert runs from 30 bp upstream of the central conserved region (Fuller-Pace and Murray, 1986) to a HindIII site downstream of S (see Figure 13b).

The phage and plasmid share homology in the central conserved regions of their S genes (Fuller-Pace and Murray,

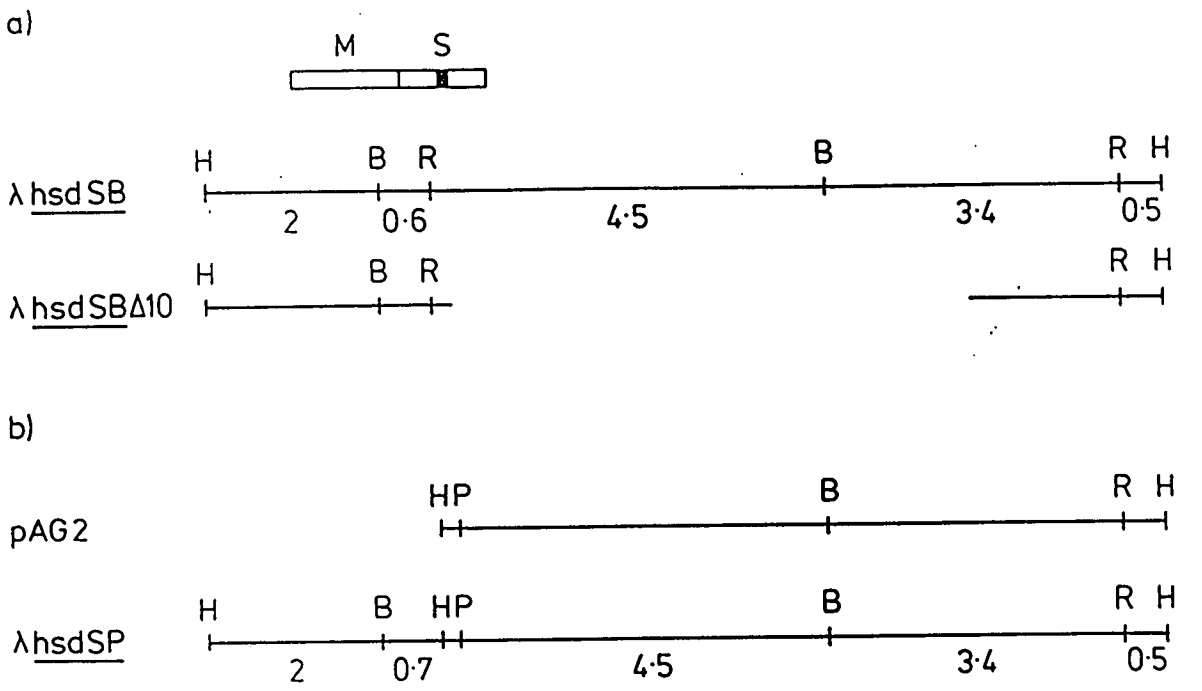


Fig. 13. Restriction maps of the DNA carried by the λhsdSB and SP phages. The positions of the M and S genes are indicated, the shaded area representing the central conserved region (Fig. 12).

(a) Shows the extent of the deletion carried by λhsdSB 10.

(b) Shows the DNA from λhsdSP that is present in the plasmid pAG2. The HindIII (H), BamHI (B), EcoRI (R) and PstI (P) targets within the cloned sequences are indicated. The distances between restriction targets are indicated in kb; that between H and P is 162bp.

1986), as well as in regions downstream of these genes (Figure 13). A crossover between the central conserved region of the S gene retained by the phage and that of the S gene on the plasmid, in conjunction with a second such event in the downstream region beyond the end of the phage deletion, would generate a phage with a recombinant S gene and a 5 kb larger genome. Since the phage and plasmid were always propagated in a bacterial host (NM522) deleted for the hsdM and S genes the chromosome was not a possible source of hsdS DNA.

To enable recombination to occur, a plate lysate of λ hsdSB Δ10 was prepared on NM522 containing pAG2. A nin+ derivative of the phage was used, as functions that stimulate recombination between phage and plasmids are encoded in this region of the genome (Lutz et al, 1987).

Phage with larger genomes were enriched for by their preferential growth on a pel- bacterial strain (WA2574) (Emmons et al, 1975; Elliot and Arber, 1978). Previous control experiments showed that λ hsdSBΔ10 plated with an efficiency of only 10^{-5}-10^{-6} that of λ hsdSB (Figure 13a), which is a phage identical in size to the expected recombinant. The population of phage grown on NM522 (pAG2) also plated with this very low efficiency, implying that the desired recombinants, if present, must indeed be very rare. The phage that did grow on this strain were then probed with single stranded recombinant M13 DNA containing a fragment specific to the distal variable region of the S gene of StySP. Plaques identified with this probe (~30%) were presumed to have rescued the StySP DNA in

this region from pAG2. Ten of these plaques were purified, reprobed and amplified on NM522.

It was possible for the starting phage (λ*hsd*SB410) to pick up, via a single crossover, the entire pAG2 plasmid. This would produce a phage with a genome of ~52 kb, which can still be packaged. Such a phage would have certain characteristics that would reveal the presence of the complete plasmid within its genome:

i) Presence of the ColE1 origin of replication enables an *imm*^{21} phage to grow on a strain lysogenic for a phage conferring immunity to phage 21. This is due to the genome being independently replicated sufficiently to dilute out the phage repressor and hence overcome repression. The λ *hsd*SB410 derivative used in this experiment is *imm*^{21}, but none of the ten isolates were able to plate on an *imm*^{21} lysogen.

ii) If the complete plasmid had been rescued, the resulting phage would carry the plasmid β-lactamase gene. The product of this gene can be detected by the conversion of nitrocefine to a pink coloured product. Phage encoding this enzyme therefore produce pink plaques in the presence of nitrocefine. This was not a characteristic of the phage isolated from this experiment.

iii) None of the ten phage isolated hybridized with labelled plasmid DNA, though all did with the M13 probe containing *hsd*SP DNA.

DNA was prepared from two of the ten lysates. HindIII digests of these (Figure 14) revealed that the insert DNA had increased in size from ~6 kb (seen in λ SBΔ10) to ~11 kb.

The new recombinant specificity system was named StySJ.

B) The Recombinant Nature of the hsdSJ Specificity Gene

The position of the crossover that generated the recombinant hsdSJ S gene was localized from the nucleotide sequence obtained from both DNA strands for the region between the EcoRI and PstI sites. The former is located in the proximal variable region of StySB, the latter in the distal variable region of StySP (see Figures 13 and 15). A comparison of the sequences in this region from two SJ isolates, and the equivalent region from the S genes of StySP (Fuller-Pace and Murray, 1986), StySB (Gann et al, 1987) and StySQ (Fuller-Pace and Murray, 1986) reveals that the crossover that generated StySJ, as in the case of StySQ, occurred within the longest region of perfect homology between the parental central conserved regions (Figure 16).

C) The StySJ Specificity Polypeptide is Functional and of Novel Specificity

A simple complementation test, originally devised to confirm the relatedness of the EcoA and E systems, was used here to establish that the StySJ specificity gene encodes a functional specificity polypeptide. Active hsdR and M genes of

1 2 3 4 5 6

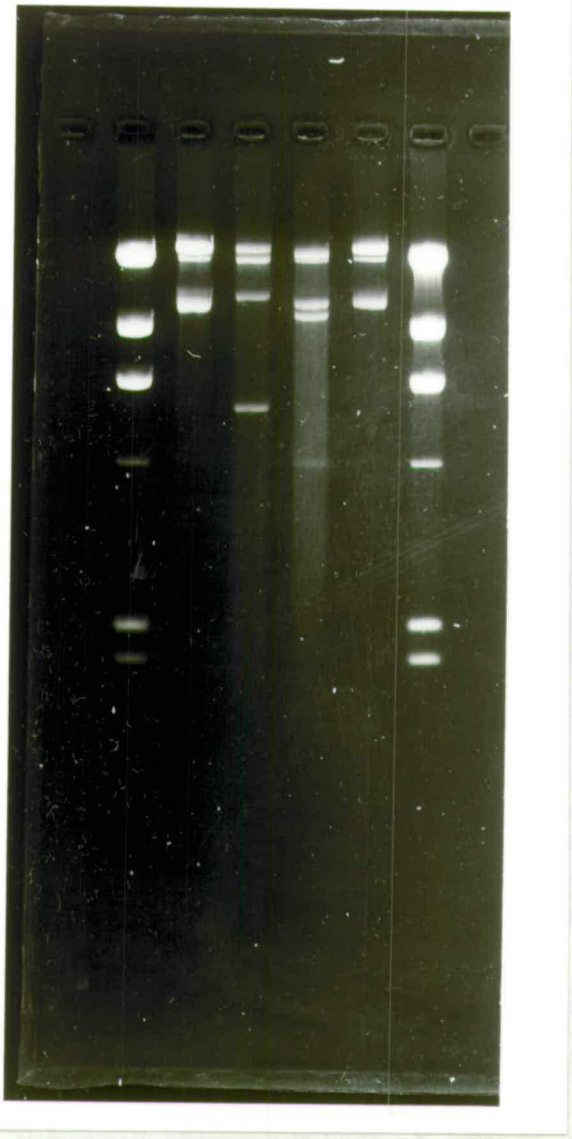


Fig. 14. HindIII digests of λhsd phage.

lane 1 and 6. λcI857. Sizes of fragments in kb are indicated.

lane 2. λhsdSB imm^{21} nin. The phage arms are 23 and 12.1kb. The hsd insert is 11kb (see Fig. 13a).

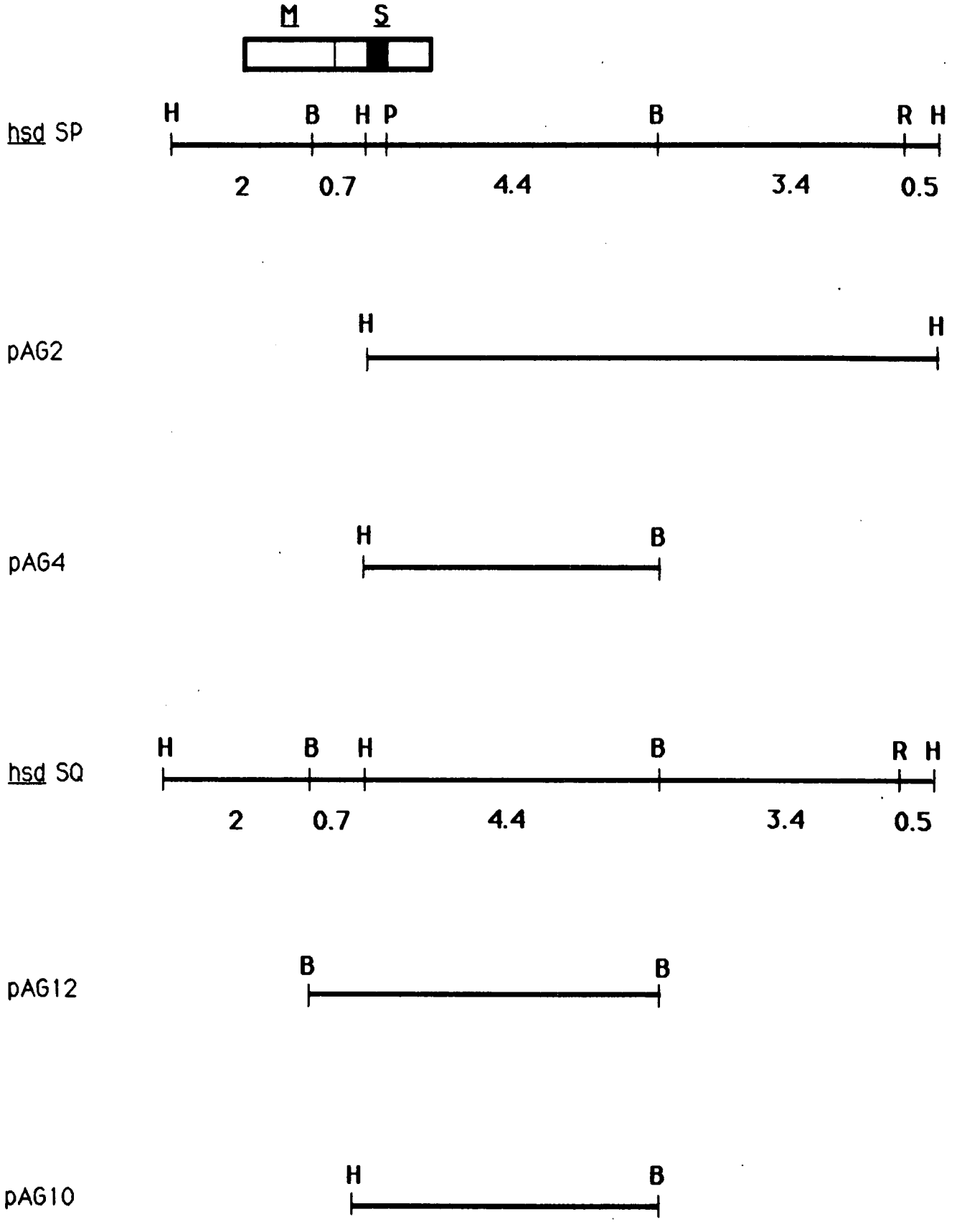
lane 3. λhsdSBΔ10 imm$^{\lambda}$ nin. The phage arms are 23 and 12.4kb. The deletion within the hsd insert has reduced its size to ~6kb (see Fig. 13a).

lane 4. λhsdSJ imm^{21} nin^{+}. The phage arms are 23, 10.4 and 4.4 kb. The presence of the nin region lengthens the right arm by 2.8kb, but this new right arm contains another HindIII site. The insert is 11kb.

lane 5. λhsdSB imm$^{\lambda}$ nin. The phage arms are 23 and 12.4kb, the insert 11kb.

Clearly evident is the increase in size of the hsd insert DNA from ~6kb in λhsdSBΔ10 (lane 3) to 11kb in λhsdSJ (lane 4).

Fig. 15. Restriction maps of the DNA carried by the λ*hsd*SP and SQ phages. The positions of the *M* and *S* genes are indicated; the shaded area represent the central conserved regions (see Fig. 12). Also shown is the DNA carried by the plasmids pAG2, pAG4, pAG10 and pAG12. Conventions and distances are as in Fig. 13.


M
S
hsd SP
H
B
H
P
B
R
H
2
0.7
4.4
3.4
0.5
pAG2
H
H
pAG4
H
B
hsd SQ
H
B
H
B
R
H
2
0.7
4.4
3.4
0.5
pAG12
B
B
pAG10
H
B

SP ATACCAATCCCGTCACTTGCTGAACAAAAAATCATCGCCGAAAAACTCGATACGCTGCTGGCGCAGGTAG

SB GTTCCTGTCCCACCTCTTGCCGAACAAAAAGTCATCGCCGAAAAACTCGATACGCTGCTGGCGCAGGTAG

SQ ATACCAATCCCGTCACTTGCTGAACAAAAAATCATCGCCGAAAAACTCGATACGCTGCTGGCGCAGGTAG

SJ GTTCCTGTCCCACCTCTTGCCGAACAAAAAGTCATCGCCGAAAAACTCGATACGCTGCTGGCGCAGGTAG

SP ACAGCACCAAAGCACGTCTTGAGCAAATCCCGCAAATCCTGAAACGTTTTCGTCAGGCGGTGTTA

SB ACAGCACCAAAGCACGTCTTGAGCAAATCCCACAAATCCTGAAACGTTTTCGCCAATCAGTGATA

SQ ACAGCACCAAAGCACGTCTTGAGCAAATCCCACAAATCCTGAAACGTTTTCGCCAATCAGTGATA

SJ ACAGCACCAAAGCACGTCTTGAGCAAATCCCGCAAATCCTGAAACGTTTTCGTCAGGCGGTGTTA

Fig. 16. Localization of the crossover that generated StySQ and SJ. The nucleotide sequences of the central conserved regions of StySP, SB, SQ and SJ specificity genes are shown. The 70bp region underlined is common to all four sequences and identifies the region in which the crossovers occurred.

the EcoK system were provided by a plasmid (pBg3) in a restriction and modification deficient host strain on which the λhsdSJ phage was plated. If the S polypeptide encoded by this phage is functional, it will associate with the resident R and M polypeptides and form an active restriction enzyme which will degrade the unmodified host chromosome. Phage encoding active S polypeptides therefore plate with a reduced efficiency in this test: λ hsdSP, SB and SQ all show a plating efficiency of 10^{-3} on NM522 (pBg3). λhsdSJ showed an equally poor plating, indicating that the new recombinant S gene encodes a functional S polypeptide.

The hsdSJ genes carried in the λ phage were transferred to the chromosome of a bacterial strain via homologous recombination (see Chapter 3, 5.6). The resulting strain, AG1, restricts unmodified or StySP, SB, SQ or EcoK modified λvir a thousandfold, demonstrating that StySJ has a specificity different from these systems. All nine other isolates from the experiment plated with an efficiency of 1 on AG1, indicating that they all confer protection against the StySJ specificity.

D) The Recombinant Nature of the StySJ Recognition Sequence

The target sequences of StySP, SB and SQ have been determined (see Chapter 2, E, Figure 11; Nagaraja et al, 1985a and b). Each of the enzymes was purified and used to methylate DNA substrates of known sequence. A computer search was then used to identify, for each enzyme, a nucleotide sequence

present in all DNA molecules that were methylated and absent from all that were not.

To discover the recognition sequence of the StySJ system, a simple in vivo strategy was devised based on the fact that phage containing a target site for a particular restriction enzyme will plate with a reduced efficiency on a strain encoding that system (Arber and Kuhnlein, 1967; Franklin and Dove, 1970).

Phage M13 was chosen for the assay for two reasons: firstly, it plates with an efficiency of 1 on an StySJ restricting strain and hence is assumed not to contain a target site for this system; secondly, by use of M13 vectors, DNA fragments of known sequence could be incorporated into the genome, resulting in phage sensitive to restriction whenever such a fragment contains an StySJ target. This can be identified by a decreased plating efficiency on an StySJ strain. Available M13 libraries of sequenced fragments were used in this screen.

Sensitivity to phage M13 requires the presence of an F plasmid in a bacterial strain. Therefore, F'Kan derivatives of AG1 (hsdSJ) and NM550 (hsdSBΔ9) were made (see Chapter 3, S.H.) This allowed comparisons of the plating efficiencies of the recombinant M13 phages on an StySJ restricting and non-restricting, but otherwise isogenic, strain.

The recombinant nature of StySJ, in conjunction with knowledge of the S polypeptide and target sequences of StySQ, enabled the prediction that the sequence recognized by StySJ would be made up of the trimeric component recognized by StySB in association with the tetrameric component of the StySP target sequence: 5′ GAG(N_6)GTRC (where R is either purine).

Initial searches of the nucleotide sequences of available M13 clones identified four which had inserts containing GAG(N_6) GTGC, a version of the predicted degenerate sequence. These clones all showed a tenfold decrease in plating efficiency on the StySJ restricting strain (see Table 2, positives 1-4). A number of clones known not to contain the candidate sequence plated with an efficiency of 1. A computer search (in collaboration with Dr. J.F. Collins) showed that GAG(N_6)GTGC was the only seven base sequence present in all four positives, even allowing for a degeneracy of the form of either purine or either pyrimidine at any position, and varying the non-specific spacer from 5 to 8.

More examples of the candidate sequence, and degenerate versions thereof, were found in the phage λ and plasmid pBR322 sequences (Sanger et al, 1982; Suttcliffe, 1979). M13 vectors carrying the appropriate regions of these were checked for their plating efficiencies (Table 2, positives 5-11; negatives 1-6). One recombinant containing two predicted StySJ targets (numbers 10 and 11 in Table) showed a hundredfold cut back, whereas all others showed only the tenfold cut back, characteristic of only one target site.

Positives					
1	A	GAG	AAAGTG	GTGC	T
2	G	GAG	CCGGAG	GTGC	T
3	C	GAG	GGAGGT	GTGC	A
4	T	GAG	CATCGT	GTGC	T
5	T	GAG	CAGATT	GTAC	T
6	A	GAG	CTGGAA	GTGC	A
7	T	GAG	ACAAAG	GTAC	G
8	T	GAG	CAGGAA	GTGC	T
9	G	GAG	GCCACG	GTAC	T
10	A	GAG	CAGGCG	GTAC	G
11	T	GAG	CACGGT	GTGC	G
Negatives					
1	G	AAG	ACCAAC	GTGC	T
2	T	GGG	GTCGAG	GTGC	C
3	T	GAA	CAGCAG	GTGC	G
4	T	GAG	CCGCTG	ATGC	T
5	T	GAG	GCGGAT	GCGC	A
6	C	GAG	GCTGCA	GTGT	A
Consensus:	N	GAG	NNNNNN	GTRC	N

Table 2

Identification of the StySJ recognition sequence. Positives (1-11) are sequences within fragments that confer sensitivity to restriction of M13 phage in which they are present. This sensitivity was seen as an approximate 10-fold decrease in plating efficiency on an StySJ restricting strain. Sequences 10 and 11 are present in the same phage, which plates with a 100-fold decrease in efficiency. Negatives (1-6) are degenerate versions of the positive sequences and did not confer sensitivity to restriction. The concensus for the StySJ sequence is shown, where R is either purine and N indicates a position at which at least 3 possible bases have been found. For clarity the sequences are written with gaps between the flanking bases, the trimeric component, the spacer, and the tetrameric component.

Positives 6-11 and negatives 4 and 6 are from phage λ (Sanger et al, 1982); positive 5 from pBR322 (Sutcliffe, 1979); positive 1 and negatives 1 and 3 from the hsdR and M genes of E.coli K-12 (Loenen et al, 1987); positive 2 and negative 5 from the hsdS gene of StySP (Fuller-Pace and Murray, 1986); positive 3 from the I-factor of Drosophila melanogaster (Fawcett et al, 1986); positive 4 from the fts region of E.coli K-12 (Robinson et al, 1984); negative 2 from M13 (Van Wezenbeck et al, 1980).

These results confirm the prediction that the recognition sequence of StySJ is 5' GAG(N_6)GTRC.

StySJ is the first recombinant produced by design (Gann et al, 1987), several isolates of which all have the same specificity. The previous recombinant, StySQ, was a single isolate from a transduction experiment (Bullas et al, 1976). The experiment has been extended (Bullas pers. comm.) without yielding any more recombinant specificities. Although StySQ certainly arose from a recombination event between the central conserved regions of StySP and SB (Fuller-Pace et al, 1984; Fuller-Pace and Murray, 1986), the presence of additional changes has not been ruled out. The isolation of this second recombinant, however, clearly demonstrates that the specificity polypeptides of the StySP and SB enzymes each contain two structurally independent recognition domains that can be reassorted to produce functional enzymes of new specificities (see Figure 17).

E) Demonstration that Recognition of the Trimeric Component of the Target Sequence is Dictated by the Amino Variable Domain

As well as reassorting the large variable regions, the recombination events that produced StySQ and SJ also reassort minor differences between the central conserved regions of StySP and SB, as shown in Figure 18 (see Chapter 2, E; Fuller-Pace and Murray, 1986; Gann et al, 1987). As a result, DNA

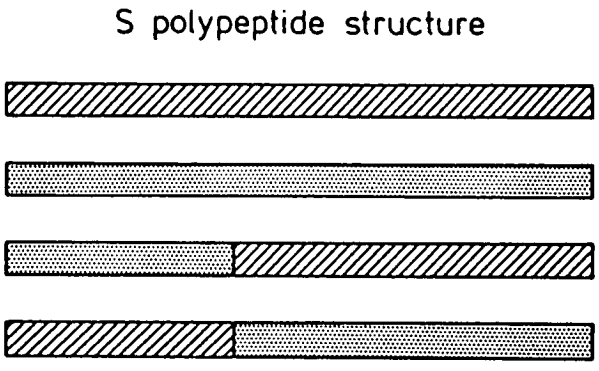


Fig. 17. Schematic diagram of wild type (stySP and SB) and the hybrid (stySQ and SJ) specificity polypeptides from the K-family, accompanied by their recognition sequences. Regions originating from stySP are hatched, those from stySB stippled.

```
    IPIPPLAEQKIIAEKLDTLLAQVDSTKARLEQIPQILKRFRQAVL

 D  VAL SPS  TL            E
 B  L L
 K                                 F
SP      S
SB  V V       V                                  S I
SQ      S                                        S I
SJ  V V       V
```

Fig. 18. The central conserved regions of the K-family S polypeptides. The amino acids, denoted by the single letter code, extend from the beginning of the first repeat to the end of the central conserved region (see Fig. 12). The uppermost line is a consensus sequence deduced from the sequences of the S genes of the five natural systems.

specificity could be a function of either the variable or the essentially conserved regions, or indeed both. All three possibilities have been suggested (Argos, 1985; Fuller-Pace and Murray, 1986; Gann et al, 1987). To ascertain which of these is true we constructed another recombinant S gene, designated StySQ*, which encodes a polypeptide whose amino variable region alone is from StySP, while the remainder of the molecule is identical to that of StySB (see Figure 19).

Between the left end of the central conserved region and the point of exchange that resulted in the formation of the recombinants StySQ and SJ, the parental genes differ in four codons (see Figure 18). The construction of StySQ* involved changing these in the S gene of StySQ, which has the proximal half of StySQ, such that they encode the corresponding residues of StySB.

These changes were made by site directed mutagenesis of an M13 template containing an ~840 bp BamHI-PstI fragment of hsdSP DNA including the central conserved region of the S gene (see Figure 15 for position of fragment within hsd region; Figure 20 for site directed mutagenesis). The four changes (3x A→G, 1x T→C) were made simultaneously using a single 45 base oligonucleotide, as indicated in Figure 20 and described in Chapter 3, 8). The wild type and mutant DNA sequences are shown in Figure 21. The changes altered the coding potential such that three isoleucines and one serine originally encoded by StySP are replaced by the three valines and one proline encoded by StySB (Figure 18). All four changes are contained within the ~250 bp

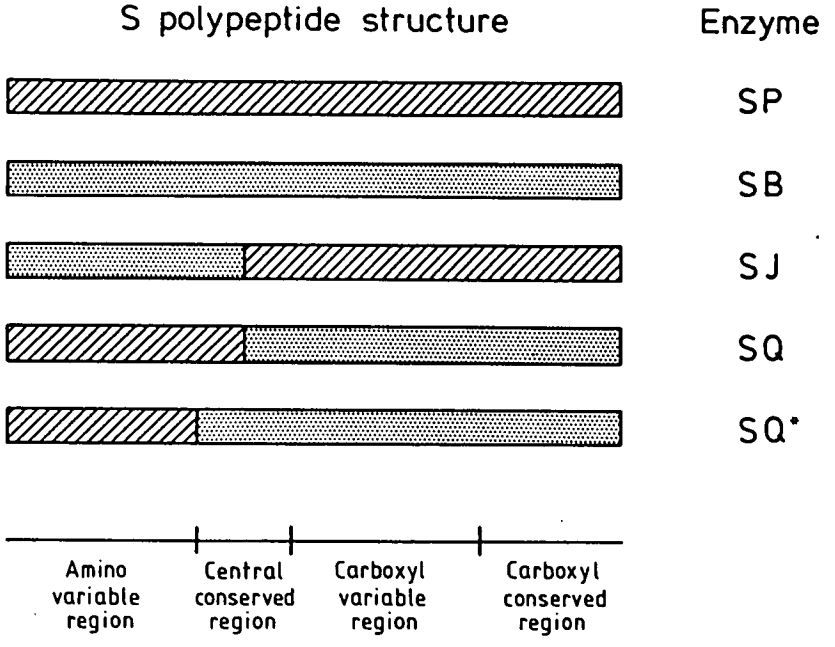


Fig. 19. Schematic diagram of wild type (StySP and SB) and recombinant (StySJ, SQ and SQ*) specificity polypeptides. Regions originating from StySP are hatched, those from StySB stippled. The bottom line indicates the positions of the variable and conserved regions in all the polypeptides (see Fig. 12).

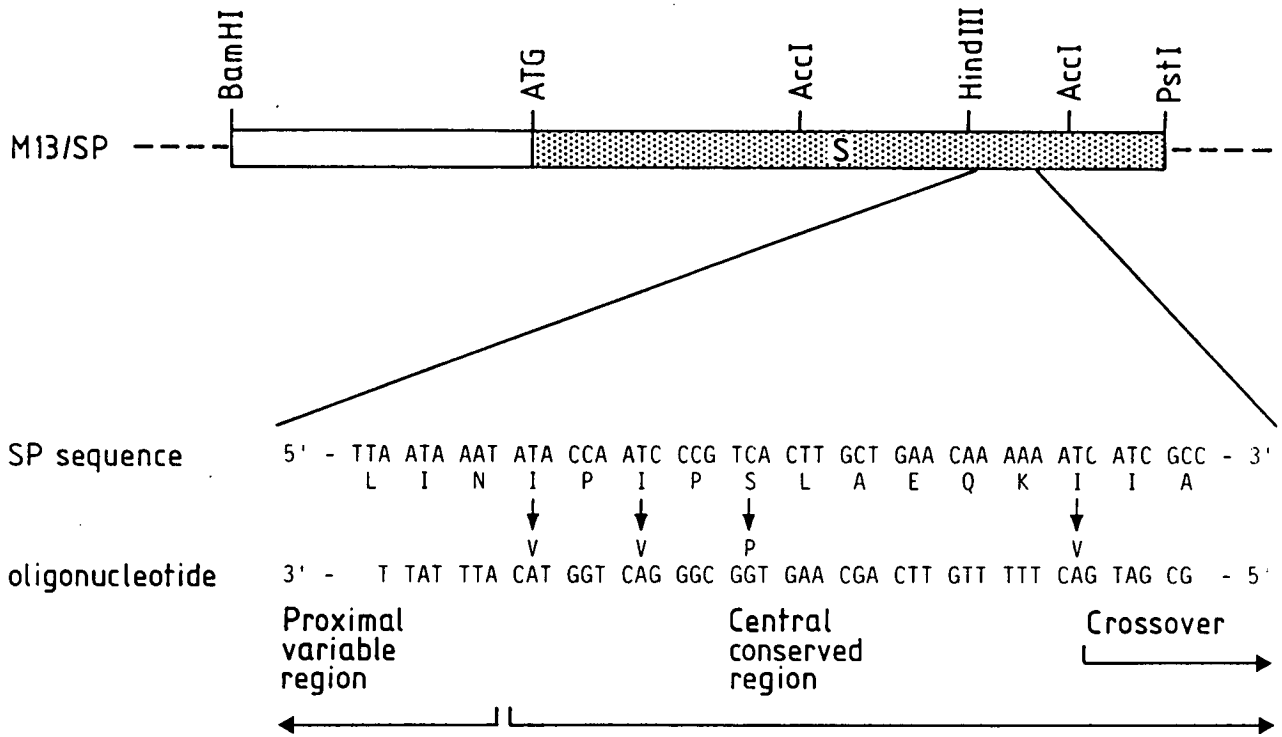

BamHI
ATG
AccI
HindIII
AccI
PstI
M13/SP
S
SP sequence
5' - TTA ATA AAT ATA CCA ATC CCG TCA CTT GCT GAA CAA AAA ATC ATC GCC - 3'
L I N I P I P S L A E Q K I I A
V V P V
oligonucleotide
3' - T TAT TTA CAT GGT CAG GGC GGT GAA CGA CTT GTT TTT CAG TAG CG - 5'
Proximal variable region
Central conserved region
Crossover

Fig. 20. Site directed mutagenesis of StySQ*. The top line shows the BamHI-PstI fragment containing part of the S gene from StySP cloned in mp18 which was used as the template for mutagenesis. At the bottom of the figure the positions of the proximal variable and central conserved regions are indicated, as is the region in which crossing over produced the recombinant S genes StySQ and SJ (See Fig. 16 and 17). The sequence of the 45 base oligonucleotide used for mutagenesis is shown along with the region of the S gene sequence to which it binds. Arrows identify the four mismatches and these changes alter four codons in StySP such that they encode the equivalent amino acids of StySB. The AccI fragment contains all the changes and was used to replace the equivalent fragment in the S gene of StySQ (in pAG12; see Fig. 22) to produce StySQ* (see Fig. 19).

mut wt

TCGA TCGA

T-C
T-C
A-G

T-C

Fig. 21. Nucleotide sequence of StySQ* mutagenesis. Gel showing the nucleotide sequence of an area within the central conserved region of StySP. On the right is the wild type (wt) sequence, and on the left the same sequence after site directed mutagenesis (Fig. 20; Chapter 4, E). The four single base changes are indicated at the side. The sequencing is in from the PstI site shown in Fig. 20.

AccI fragment (Figure 20) which, except for these changes, is identical to the equivalent AccI fragment from the S gene of StySQ.

Construction of the complete StySQ* S gene, and placing it into a system where the phenotype can be examined, involved several steps which are shown in Figure 22 and are described below.

A 5.1 kb BamHI fragment from λhsdSQ contains the entire S gene (see Figure 15). This was inserted into the unique BamHI site in the polylinker of plasmid pAG11 to produce pAG12 (step B in Figure 22). pAG11 is a derivative of pEMBL8+ (Dente et al, 1983) in which the AccI site in the polylinker has been destroyed (step A in Figure 22). This was done by cutting pEMBL8+ with AccI, filling in the resulting cohesive ends using Klenow polymerase, and ligating the blunt ends produced. The lacZ reading frame in which the polylinker is situated is disrupted, resulting in lacZ transformants of NM522, which, in the presence of IPTG and XGAL, therefore give white colonies (see Chapter 3, 6.0). Analysis of plasmid DNA, by restriction enzymes and nucleotide sequencing, identified an isolate in which the AccI site had been filled in and no other change had occurred. The removal of this AccI site was essential for the subsequent step in which the AccI fragment within the BamHI insert of pAG12 was replaced by the equivalent fragment from the M13/SP derivative following site directed mutagenesis (step C in Figure 22). This produced pAG13, which was identified as containing the mutated AccI fragment by cleavage with RsaI, a

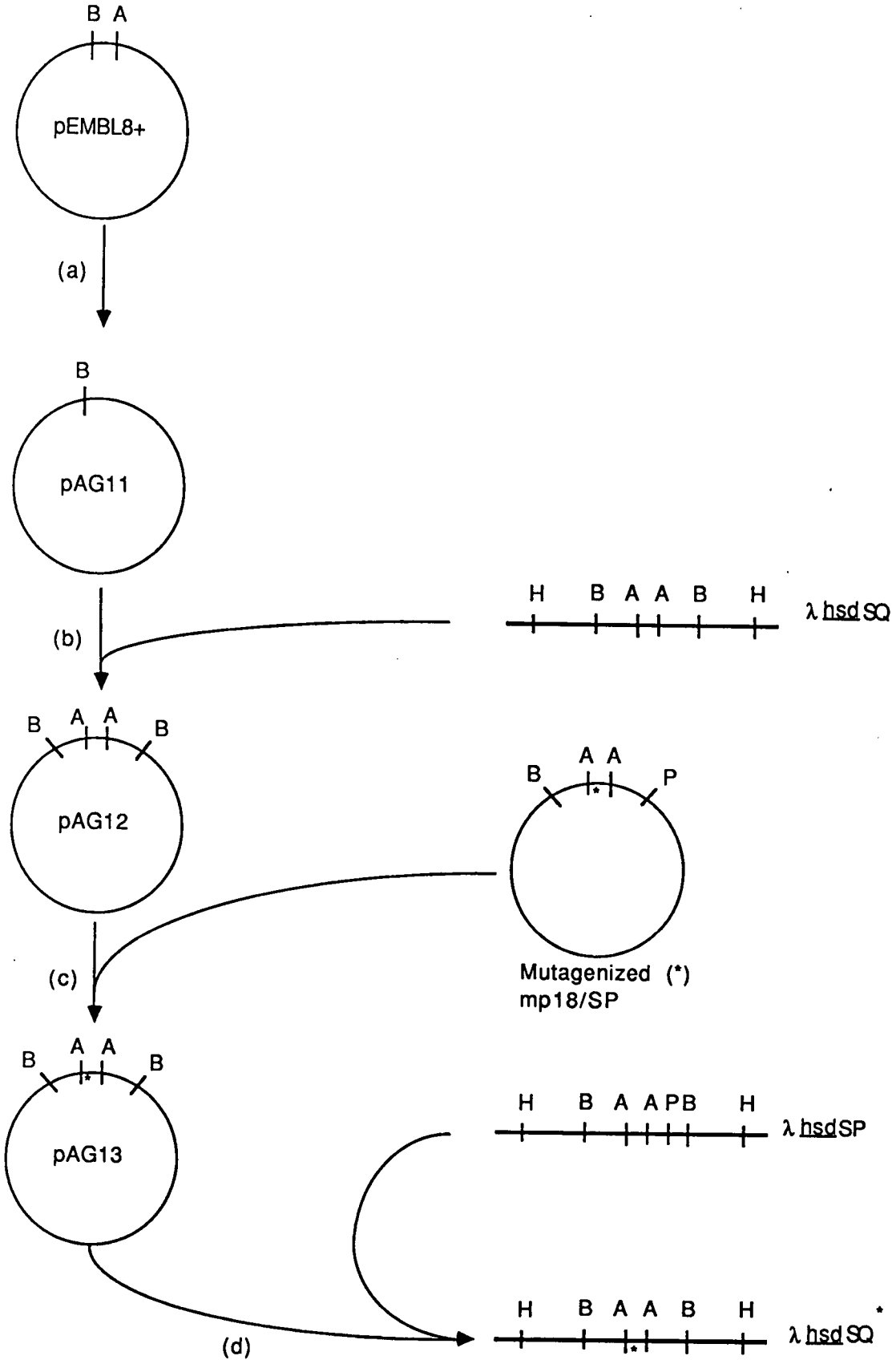

B A
pEMBL8+
(a)
B
pAG11
(b)
H B A A B H
λ hsd SQ
B A A B
pAG12
B A A P
*
Mutagenized (*)
mp18/SP
(c)
B A A B
*
pAG13
H B A A P B H
λ hsd SP
H B A A B H
*
λ hsd SQ
(d)

Fig. 22. Reconstruction of the complete S gene of StySQ*.

a) Destruction of the AccI site within the polylinker of plasmid PEMBL8$^+$ (Dente et al, 1983) by cutting with AccI, filling in the cohesive ends with Klenow polymerase, and religating.

b) Inserting the BamHI fragment from λhsdSQ (see Fig. 15) into the BamHI site in the polylinker of pAG11.

c) Replacing the AccI fragment from within the hsdSQ BamHI insert of pAG12, with the equivalent AccI fragment of hsdSP, after site directed mutagenesis of the latter (Fig. 20).

d) Replacing the BamHI fragment from within the hsd region of λhsdSP with the equivalent fragment from pAG13 to produce λhsdSQ*.

Details are given in the text (Chapter 4; A)

new site for which is created by one of the mutations, and confirmed by DNA sequencing. pAG13 therefore contains the complete StySQ* S gene.

The BamHI fragment from pAG13 was excised and used to replace the corresponding BamHI fragment from a λhsdMS SP phage (step D in Figure 22). The resultant phage, λ hsdSQ*, was identified by hybridization with recombinant M13 single strand DNA containing a region from the distal variable region of the S gene of StySB (and hence StySQ and StySQ*, but not StySP). λhsdSQ* encodes not only the new S polypeptide, but also a complete and, presumably, compatible M polypeptide. The presence of all four mutations in this phage was confirmed by subcloning and resequencing the appropriate region.

F) The StySQ* Specificity Polypeptide is Functional and has the Same Specificity as StySQ

As with the hsdSJ system, the first indication that the new S polypeptide was functional was obtained from the reduced plating efficiency of λhsdSQ* in a killing test (see Chapter 4, C). By contrast, however, reduced plating (killing) was not seen when this phage was grown on NM551(pBg3). NM551 encodes the StySQ system, and hence has an SQ modified chromosome. This strain is therefore not killed if the S polypeptide encoded by an incoming λhsd phage is of SQ specificity; λhsdSQ itself plates with an efficiency of 1 on NM551(pBg3), while λhsdSP is cut back a thousandfold. The efficient plating of

λ*hsd*SQ* on this strain therefore indicates that the specificity of this system is identical to, or a subset of, that of *Sty*SQ.

To establish that *Sty*SQ and SQ* were of identical specificity, a dilysogen of λ*hsd*SQ* and a heteroimmune helper phage (see Chapter 3, S.P) was made in NM522. This bacterial strain, deleted for *hsdM* and *S*, still has a chromosomal *hsdR* gene from the *Eco*K system. The λ*hsd*SQ* provides M and S polypeptides, and thus the dilysogen produces a complete *Sty*SQ* restriction and modification system.

Table 3 shows the plating efficiencies of unmodified, SQ modified and SQ* modified phage P3 on NM522(r-m-), NM551(r^+_{SQ} m^+_{SQ} and the *Sty*SQ* dilysogen (r^+_{SQ*} m^+_{SQ*}). These demonstrate that both systems restrict unmodified P3, but not if the phage has previously been propagated on, and hence modified by, either system. This implies that *Sty*SQ modification protects against *Sty*SQ* restriction, and *Sty*SQ* protects against *Sty*SQ. Thus *Sty*SQ and SQ* are indeed of identical specificity.

The S polypeptide structures and recognition sequences of *Sty*SP and SB, as well as all three recombinants - *Sty*SQ, SJ and SQ* - are shown in Figure 23.

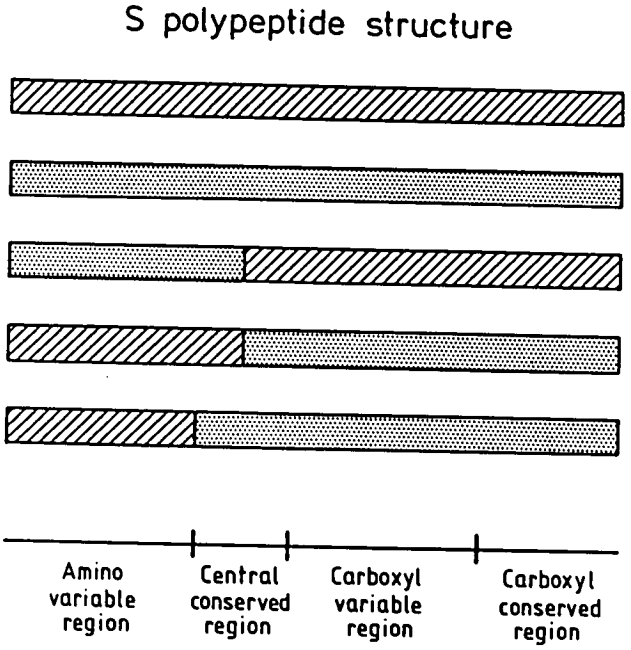
This experiment clearly demonstrates that the specificity of the trimeric component of an enzyme's target sequence is dictated entirely by the amino variable region of its specificity polypeptide. This region is therefore now referred to as the amino recognition domain (Cowan *et al*, Cell in

Table 3: Efficiencies of plating of phage P3 on StySQ and SQ* restricting hosts

	P3.0	P3.SQ	P3.SQ*
NM522	1	1	1
NM551 (SQ)	10^{-3}	1	1
Dilysogen (SQ*)	10^{-3}	1	1

P3.0 is unmodified P3; P3.SQ or SQ* is P3 modified by previous growth on an StySQ and SQ* modifying strain.

NM522 is an r^-m^- strain.



Enzyme	Recognition sequence
SP	AAC (N_6) GTRC
SB	GAG (N_6) RTAYG
SJ	GAG (N_6) GTRC
SQ	AAC (N_6) RTAYG
SQ*	AAC (N_6) RTAYG

Fig. 23. Schematic diagram of wild type and recombinant S polypeptide (as shown in Fig 19), accompanied by their recognition sequences.

press). Arguments in favour of the entire variable region being involved in recognition, rather than the recognition domain merely occurring somewhere within it, are put forward in the Discussion.

2) Effect of Deleting the Amino Recognition Domain

Introduction:

Having established that the amino variable region is alone responsible for dictating the specificity of one half of the recognition sequence, it was of interest to see what functions, if any, would be retained by a polypeptide from which this single recognition domain had been deleted. The most ambitious hope was that, in complex with M subunits, such a polypeptide would still be capable of methylating DNA, but that the target sequence would be that defined by the carboxyl recognition domain of the original S polypeptide. Alternatively, removing the amino recognition domain might leave a polypeptide which, while non-functional in terms of enzymatic activities, still folds correctly and perhaps interacts with other subunits and/or DNA.

The rationale for this experiment was based on two observations. Firstly, the independent nature of the DNA recognition domains within S polypeptides, as demonstrated by the functional recombinants previously isolated. Secondly, the DNA binding domains of several proteins have been shown to function when isolated from other parts of the polypeptides,

which themselves can operate without their normal DNA binding domains (e.g. Brent and Ptashne, 1985).

A) Construction of Genes Encoding Specificity Polypeptides Deleted for their Amino Recognition Domains

Two genes were constructed that encode polypeptides lacking amino recognition domains (ARD⁻S polypeptides). This was done taking advantage of the HindIII site that occurs ~30 bp upstream of the central conserved region and the BamHI site downstream of the S genes of StySP and SQ (see Figure 15). These fragments were inserted into the polylinker of the plasmid vector pUC13 (Vieira and Messing, 1982). This produces an inframe fusion between the lacZ gene of the vector and the remaining fragment of S gene. The resulting constructs therefore encode polypeptides comprising the first seven amino acids of β-galactosidase fused to residues 151 to the C-terminus of the StySP or SQ specificity polypeptides (see Figure 24). Expression of these fusion genes is under the control of the lac promoter.

The plasmid encoding the SP ARD⁻S polypeptide is designated pAG4; that encoding SQ ARD⁻S is pAG10 (see Figure 15 and Figure 24).

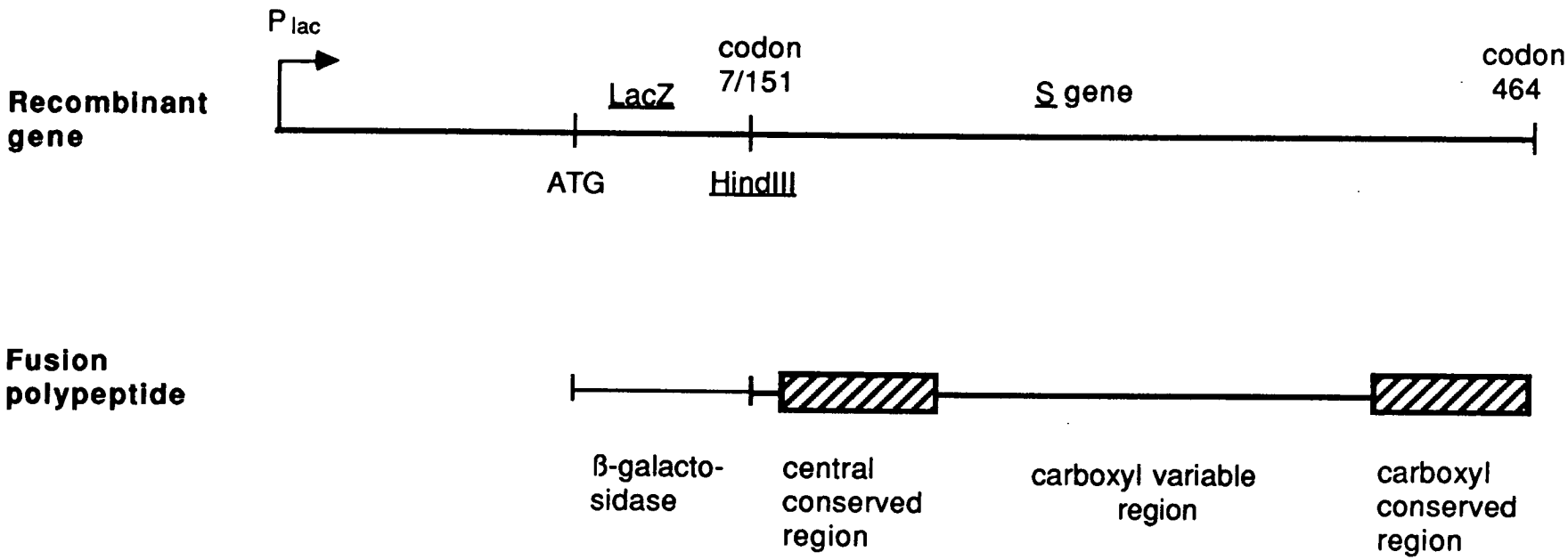

Recombinant gene
P lac
LacZ
codon 7/151
S gene
codon 464
ATG
HindIII
Fusion polypeptide
ß-galacto-sidase
central conserved region
carboxyl variable region
carboxyl conserved region

Fig. 24. Schematic diagram of the structures of the SPARD$^-$ and SQARD$^-$ specificity polypeptides. HindIII - BamHI fragments from the hsdSP and SQ regions were cloned in pUC13 plasmid vector (Veiera and Messing, 1983) to produce pAG4 an pAG10 (see Fig. 15). The HindIII junction produces an inframe fusion between the LacZ and the truncated S genes. These constructs therefore encode polypeptides containing the N-terminal seven amino acids of β-galactosidase fused to residues 151 to the C-terminus of the S polypeptides. These contain the central conserved, carboxyl variable and carboxyl conserved regions, but are deleted for the amino recognition domains, and their expression is under the control of the lac promoter situated upstream of the lacZ gene in pUC13.

B) Do the ARD⁻S Polypeptides Direct Methylation *in vivo*?

If, for example, the SP ARD⁻S polypeptide were able to direct methylation under the influence of its single recognition domain, this would be expected to be aimed at the adenine residue in sequences identical to the tetrameric component of the StySP target, i.e. 5'GTRC. If all such sequences within a genome, e.g. λ, were modified in this way, then every StySP target would be hemimethylated; no methylation of the 5' AAC component would occur. Hemimethylation, however, is sufficient to protect against restriction, this being, after all, the normal state of a cell's own genome immediately after replication (see Chapter 2). Therefore, phage grown on a bacterial strain expressing the SP ARD⁻S polypeptide, in conjunction with M subunits, should protect against subsequent StySP restriction if the fusion polypeptide directs efficient methylation of the tetranucleotide. Similarly, the phage ought to be protected against StySJ restriction as the target sequence of this system has the same tetrameric component as does StySP. In the case of the SQ ARD⁻S polypeptide, it would be protection against the StySQ and SB restriction that might be expected.

To test this, the bacterial strain NM490 was transformed with the two plasmids, pAG4 and pAG10. This strain, while being r-, encodes the M and S genes of the EcoB system, therefore providing M subunits compatible with the StySP and SQ systems. Propagation of λvir on the plasmid containing derivatives, even in the presence of IPTG (which completely

induces the lac promoter), afforded no protection against subsequent restriction by StySP, SB or SJ; a plating efficiency of 10^{-3} was seen in all cases. Similarly, no protection of λvir against StySJ restriction was seen after plating on BMH71-18. (hsdR+M+S_K^+) carrying pAG4. The presence of the ARD⁻S polypeptides also inhibited modification of λ vir by the chromosomal systems.

Although this experiment fails to detect methylation, it does not prove that no methylation is produced by ARD⁻S polypeptides. The level demanded by the assay is very high; almost all of the StySP or SQ sites in a phage need to be modified for any reduction in its restriction to be detectable in vivo. Also, if methylation is directed to all 5′ GTRC sequences for SP and 5′ RTAYG sequences for SQ, then the number of targets in the cell would be far greater than that of the usual seven base pair sequences recognized by type I enzymes. Therefore, in this system, only very efficient methylation of very large numbers of targets would be detectable. Perhaps in vitro some methylation would be seen. In many experiments of this type - i.e. where functional demands are made of artificially manipulated polypeptides - activity is seen, but at greatly reduced levels that can only be detected in vitro (e.g. Bushman and Ptashne, 1988). Also, although the ARD⁻S polypeptides are over expressed in the cell, the M subunits with which they must interact are not, and so the level of potentially active complex may be limited by this and be no higher than in a wild type situation.

C) Can any Activity of ARD S Polypeptides be Detected *in vivo*?

The ARD⁻S polypeptides do appear to fold in a sufficiently accurate manner to be capable of interacting with other enzyme subunits. The resultant effect is observed as an inhibition of wild type K-family R-M systems.

Bacterial strains, each carrying one of the chromosomally located systems *Eco*K, B, A, *Sty*SP, SB, SQ or SJ, were transformed with pAG4, pAG10 or pUC13 (vector). Table 4 shows the plating efficiencies of λ*vir* (or P3 for *Sty*SQ) on each of the transformed strains, compared to the non-restricting NM522. As can be seen, restriction by all K-family systems is inhibited, while *Eco*A is still effective. Similarly methylation is inhibited (not shown). There appears to be no sequence specificity involved in the effect, only family specificity. It therefore seems most likely that sequestering of enzyme subunits by the ARD⁻S polypeptides disrupts the formation of functional restriction complexes.

Table 4: Effect of ARD⁻S polypeptides on restriction

	Plasmids		
Bacterial strain	pUC13	pAG4	pAG10
NM522 (r^-)	1	1	1
BMH71-18 (r_K)	$3x10^{-4}$	$7x10^{-1}$	$3x10^{-1}$
NM661 (r_B)	$5x10^{-4}$	$5x10^{-1}$	$1x10^{-1}$
L4001 (r_{SB})	$1x10^{-3}$	$8x10^{-1}$	1
L4002 (r_{SP})	$5x10^{-4}$	$5x10^{-1}$	$5x10^{-1}$
NM551 (r_{SQ})	$5x10^{-3}$	1	$3x10^{-1}$
AG1 (r_{SJ})	$1x10^{-3}$	$4x10^{-1}$	1
WA2899 (r_A)	$2x10^{-3}$	$1x10^{-2}$	$1x10^{-2}$

Figures are e.o.p of λvir.0 (or P3.0 in the case of NM551).

CHAPTER 5 : DISCUSSION

The results presented demonstrate that a large region, consisting of the N-terminal 150 amino acid residues of the specificity polypeptides of K-family type I restriction and modification enzymes, constitutes a single DNA recognition domain that determines the trimeric component of the target sequence. A second domain, by analogy presumably consisting of the 150 residues of the carboxyl variable region, dictates the specificity of the second defined component of the target. These two domains are sufficiently independent to be capable of functioning in new combinations, thereby producing enzymes that recognize novel, hybrid, target sequences. Deletion of the amino recognition domain produces a polypeptide which, while apparently capable of binding other subunits, shows no enzymatic activities.

Recently, the S genes of three members of the A-family of type I enzymes have been sequenced (Cowan et al, Cell in press; P. Kannan, unpublished data). Although the overall organization and function of this second family are in many ways identical to those of K, they have been judged unrelated by genetic and molecular criteria (Murray et al, 1982; see Chapter 2). A comparison of either the nucleotide or predicted amino acid sequences corroborates this sharp distinction in showing no general sequence similarity between the families, even in regions that are conserved within either one. This is seen not only for the conserved regions of S, but, where known, for other subunits of the complex (see Loenen et al, 1987; G.

Cowan, J. Kelleher and A. Daniel, unpublished results). Nevertheless, the S polypeptides of the A-family, like those of K, contain two large variable domains of ~150 amino acids which we again believe represent two DNA recognition domains (Cowan et al, Cell in press).

Observations from both families, in addition to those described, correlate the recognition domains with the variable regions. EcoK and StySP of the K-family both recognize 5′ AAC (Kan et al, 1979; Nagaraja et al, 1985b) and show 90% identity throughout their amino variable regions (Fuller-Pace and Murray, 1986). EcoA and E from the A-family, both of which recognize 5′ GAG, have amino variable regions which show 80% sequence identity (Cowan et al, Cell in press).

A particularly satisfying observation is a 44% identity seen between the amino variable regions of StySB from the K-family, and that of either EcoA or E from the A-family (Cowan et al, Cell in press). This represents the only obvious sequence similarity between the two families and correlates with all three enzymes recognizing 5′ GAG as the trimeric component of their respective target sequences. The indication is not only that the variable regions are recognition domains, but also that, although of generally dissimilar amino acid sequence, the recognition mechanisms employed by the two families are the same.

EcoK and StySP not only recognize identical trimeric components, but also very similar tetrameric components, that

of StySP (5'GTRC) being simply a degenerate version of the sequence recognized by EcoK (5'GTGC) (see Figure 11, Chapter 2; Nagaraja et al, 1985b; Kan et al, 1979). Since the carboxyl variable regions are implicated as the recognition domains for the tetrameric component of the target sequence, those of StySP and EcoK may be expected to be rather similar. In fact they are only very slightly more alike than any two variable regions of different specificity (Fuller-Pace and Murray, 1986; Gann et al, 1987). However, the degeneracy within the StySP target sequence actually requires that the enzyme be unable to discriminate between either purine, while EcoK, by contrast, clearly can. Therefore, though similar, these two target sequences demand that the two enzymes see them in quite distinct ways, and hence there is no reason to expect their recognition domains to show much similarity. According to the scheme of Seaman et al (1976), G-C and A-T base pairs appear identical to a protein contacting them in the outer major and outer minor grooves, while they can be discriminated by contacts to the central major and minor grooves.

The recognition domains defined in the type I S polypeptides are very large, and there is no direct evidence to implicate all the residues within them in defining specificity. Nevertheless, when two from the same family specify different target sequences, it is very difficult to detect any similarity between them (Gough and Murray, 1983; Gann et al, 1987; G. Cowan and P. Kannan, unpublished observations). In contrast, the similarity found between those of identical specificity from the same family extends throughout the length of the

domains (Fuller-Pace and Murray, 1986; Cowan et al, Cell in press). Together, these observations implicate the whole variable region in recognition. However, the much lower similarity (54%) seen between recognition domains of identical specificity from different families (K and A) may reveal a more strict definition of the amino acids essential to determination of a given specificity. At the same time it must be remembered that the M subunits which S polypeptides from different families direct in methylation are quite different (G. Cowan and A. Daniel, unpublished observations; Murray et al, 1982) and consequently, the details of the precise interactions made by the two enzyme complexes and DNA may vary. Accommodating this may in turn necessitate slight variation in the recognition domains. It is therefore possible that, within the context of each enzyme, most residues in a recognition domain are important in defining specificity. In fact, if some variation is necessary in providing different enzymes with identical specificities, it merely emphasizes the subtlety involved in the recognition process. The A and K family enzymes are known to differ, for example, in their relative efficiencies of methylation of hemi and unmethylated DNA substrates (Suri et al, 1984 b).

If domains of 150 amino acids are necessary to specify recognition of nucleotide sequences of only 3, 4 or 5 bp, then it must involve a more complex mechanism than the type of simple interactions between linear segments of polypeptide and linear sequences of bases exemplified in some repressor-operator binding (Pabo and Sauer, 1984). In the case of the

type II restriction enzyme EcoRI, adjacent bases in its target sequence are contacted by residues well separated in the amino acid sequence (e.g. Arg 145 and Arg 200; see Chapter 1; McClarin et al, 1986). Also, the importance of precise presentation of these amino acids in defining specificity, and the extent of polypeptide that may be involved in this, is emphasized by the EcoRI endonuclease where different arginine residues interact with A-T or G-C base pairs, dependent in each case on their relative positions with respect to the target sequence. Defining specificity in DNA recognition can therefore involve extensive regions of polypeptide, much of it quite separate from the direct interactions occurring at the protein-DNA interface. This appears to be particularly true when the protein acts on its target sequence, rather than merely binding to it.

Type I R-M systems are more complex than type II, and their various activities may require still more sophisticated recognition mechanisms. As described in Chapter 2, type I systems consist of a single multisubunit enzyme species which can act as both a DNA methylase and an endonuclease. Complete modification of the target sequence involves methylation of one adenine within each defined component. In turn, the methylation state of the sequence dictates the bound enzyme's subsequent behaviour. When the sequence is fully methylated, the enzyme dissociates from it. When hemimethylated, the complex methylates the complementary strand. Only when unmethylated is the DNA cut. In this system, therefore, specific nucleotide sequences are recognized, and information

concerning their methylation state transmitted to the protein, where different activities are selected. The target sequence is bound initially by the enzyme irrespective of methylation state, though subsequently these must be distinguished. In contrast, modification by the EcoRI methylase simply disrupts important protein-DNA interactions between the endonuclease and its target, thereby inhibiting binding. It may only be in the light of such complexities that the extensive nature of the recognition domains of type I restriction enzymes will be understood.

Although dictating its specificity, there is no evidence that the S polypeptide alone makes direct physical contact with the target sequence. Indeed, as methylation is a function associated with M subunits, it is expected that this polypeptide will be quite intimately involved in the interaction with DNA. As mentioned in Chapter 2, the tetrapeptide D/N PP Y/F is found in all adenine methylases, including the M subunit of type I enzymes (Loenen et al, 1987); this same sequence has also been found in the region of a repressor protein that specifies an N6 methyl adenine within its operator (Youderian et al, 1983). The simplest explanation is that the tetrapeptide interacts directly with methylated adenines (Vershon et al, 1985). More precisely, the tetrapeptide, a methyl group and an adenine base can form a 'sandwich', with the methyl group between the peptide and the base. This could perhaps be achieved whether the methyl group is initially attached to either the protein or the DNA. If we also assume that two methyl groups cannot be accommodated in

this sandwich structure, then this provides us with a description of how the model proposed by Burkhardt et al (1981) for enzyme discrimination between the methylation states of target sites could operate. They proposed (see Chapter 2) that the enzyme uses methyl groups to probe the target sequence in order to ascertain whether or not it contained methylated adenines. One M subunit was envisaged as sitting over each defined component of the recognition sequence. If the sequence was unmethylated, then the polypeptide bound methyl group could be accommodated in the major groove allowing the M subunit to move close to the DNA. If both M subunits were in this state (i.e. a completely unmethylated target site), the enzyme would be in what was designated the 'closed' conformation, in which methylation is inhibited, but the R subunits are appropriately positioned for restriction to occur. If both M subunits are held off the DNA by a steric clash between protein and adenine bound methyl groups (i.e. completely methylated sequence), then the enzyme is in what was designated the 'open' complex, which rapidly dissociates. The semi-open complex is that in which one of the M subunits is in the closed and one the open position; this allows methylation but not restriction. The indication that there is a close fit between the DPPY tetrapeptide and an N6 methylated adenine, forms a structural basis for this model. Presumably, when the enzyme approaches its target, it can form a tight fit on unmodified components, but not those containing methylated adenines, where two methyl groups would clash.

The involvement of interactions between M subunits and the adenine bases conserved in all target sequences (Figure 11) implies that S only specifies the other, non-conserved, positions. This means that the 150 residue amino recognition domain actually specifies only 2 bp directly! At the same time, an arrangement in which different bases of the target sequence are contacted by residues in separate enzyme subunits further emphasizes the complex nature of the recognition mechanism employed by type I enzymes.

Although the two recognition domains within each S polypeptide are essential in dictating the target sequence, the strict definition of the length of the non-specific spacer is equally relevant to specificity. This must have some physical basis; perhaps constraints within the protein structure and important non-specific protein-DNA interactions demand that the two components of the target will only be bound when precisely positioned. A spacer of fixed length between components of a target sequence is not a characteristic inherent in merely having two recogntion domains within a protein, even when these bind to their target sequences simultaneously. The Int protein of phage has been shown to contain two such domains, each recognizing a different nucleotide sequence (Moitoso de Vargas et al, 1988). In this case the spacing of the targets is not defined. Similarly, two dimers of repressor interact with each other in their co-operative binding of two operators, thereby producing, in effect, a tetramer with two recognition domains; the operators can be spaced quite far apart and, as long as they are maintained on the same side of the helix,

binding still occurs (Hochschild and Ptashne, 1986b). The intervening DNA is presumably looped away from the protein so as to allow the two target sites to bind to the polypeptide domains. RNA polymerase is perhaps rather like type I enzymes in respect to spacing. A single polypeptide (sigma) dictates DNA recognition specificity. It contains two recognition domains that interact with two different target sites (-10 and -35 regions of the promoter). The spacing of these sites is defined and is, therefore, a component of specificity (Helmann and Chamberlin, 1988).

Within each family of enzymes, the M and R subunits are interchangeable between the various S polypeptides (see Chapter 2). Regions of conserved amino acid sequences are therefore an expected characteristic of these S polypeptides. Two such regions (the central and carboxyl conserved regions; see Figure 25) occur within those of the K-family; S polypeptides from the A-family each contain three conserved regions (Cowan and Kannan, unpublished). These regions, very highly conserved within but not between families, have been assumed to be involved in subunit/subunit interactions (Fuller-Pace and Murray, 1986; Gann et al, 1987).

Argos (1985) demonstrated that, for the K-family, there is, within each S polypeptide, a repeated sequence. This repeat is about 60 residues long and overlaps, though is not completely confined to, the regions conserved between each polypeptide. The level of similarity between these repeats is much lower than that seen between the conserved regions of

different S polypeptides (i.e. the various K-family central conserved regions are much more alike than are the Argos repeats within an S polypeptide. Similarly, the carboxyl conserved regions). I shall refer to the respective repeated sequences as the central repeat and the carboxyl repeat. Each of these sequences is made up of two components, A and B; these, but not the intervening sequence, are repeated (see Figure 25).

Argos proposed that the repeats were the DNA recognition domains, a claim which the experiments reported here demonstrate not to be true. However, if each conserved region (and hence repeat) is a binding site for other subunits, then the fact that these regions are similar to each other may not be surprising; it has been suggested that the complete EcoK restriction enzyme contains two M and R subunits per S polypeptide (Meselson et al, 1972).

Recently it has been found that S polypeptides of both the A and R124 families also have a repeat (my unpublished observations). This corresponds, in length and approximate position, to the A component of the K-family Argos repeat (Figure 25). Most significant, however, is the fact that, at the level of similarity detected between repeats within an S polypeptide, the repeat is the same in all three families; an alignment of this sequence from the central and carboxyl regions of EcoK, A and R124 reveals that all six sequences are similar, and that, at most positions, there is no tendency for identical amino acids in the two repeats within a given

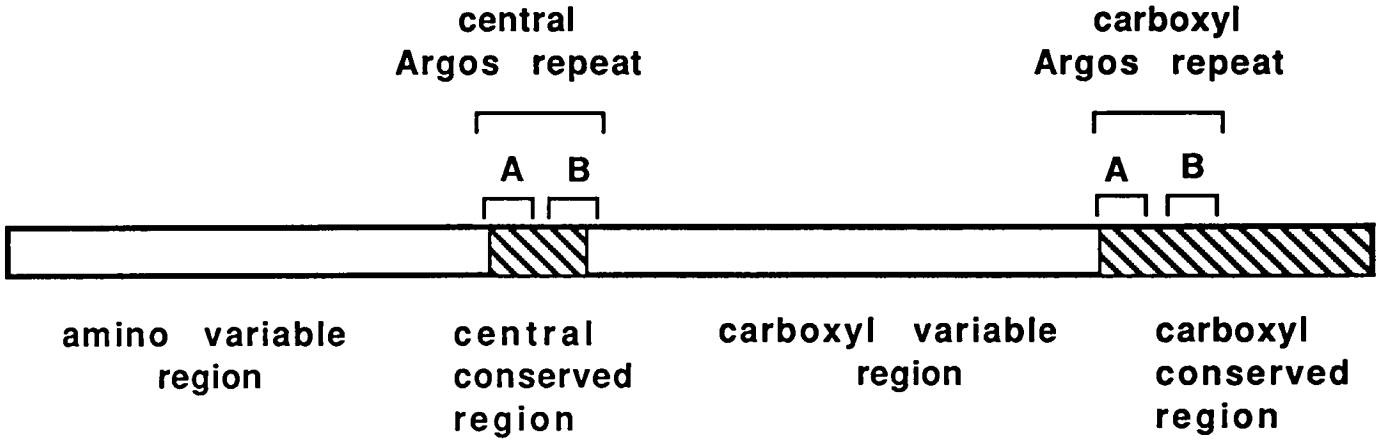


Fig. 25. Schematic diagram of a K-family specificity polypeptide, as shown in Fig. 12, but with more details of the Argos repeats. A and B represent regions that are repeated, while the intervening sequence is not. The entire Argos repeat is designated as the region from the N-terminal end of the A component to the C-terminal end of the B component.

polypeptide which are (at the same time) different from those in the other families (Figure 26). It is therefore attractive to suggest that this region is involved in some function common to all the enzymes. Certainly it has to be accepted that if the repeat within a given S polypeptide is functionally significant, then the equivalent similarity found between families must have comparable significance. The B component of the Argos repeat (Figure 25) appears to be specific to the K-family; the other families have no sequences in the equivalent regions which appear to be themselves repeated.

In the A family the repeats are actually located immediately upstream of the conserved regions, while in K they are almost completely within the conserved regions. Therefore, although the central (or carboxyl) repeats within different K-family members are much more alike than are the central and carboxyl repeats within a given polypeptide, in the A family this is not the case; the central repeats within all three known members of the A family (EcoA, E and CfrA) tend to be identical only in positions found to be generally conserved in repeats from all families (compare Figure 26 and 27). The A family carboxyl repeats are even less alike than are the central. This adds considerable weight to the idea that the conservation required within these regions is at the level of the repeat, and not the very high level found between, for example, central repeats from different K-family members. Indeed, the greater variation found in the central repeat in EcoD (Figure 18) indicates that complete conservation is not required, even in the K-family, for normal function. In this

```
          1                          24
CTA.      IPFPPLQEQE RIIIRFTQLM SLCD
CXA.      FPLIPQSEQD RIISKMDELI QTCN
CTK.      IPIPPLAEQK IIAEKLDTLL AQVD
CXK.      VLLPPVKEQA EIVRRVEQLF AYAD
CTR.      NPEKSLAIQS EIVRILDKFT ALTA
CXR.      IPVPNINEQQ RIVEILDKFD TLTN
Consensus IPLPPL*EQ- RIV**LD*L* AL-D
```

Fig. 26. An alignment of the amino acid sequences of the central (CT) and carboxyl (CX) A repeats from EcoA (A), EcoK (K) and EcoR124 (R) specificity polypeptides. The consensus indicates positions where at least three residues are identical. The asterisks indicate positions where at least one sequence from each of at least two are conserved. In total, therefore, any position indicated by an amino acid residue or an asterisk in the consensus is one where there is interfamily conservation. Of the two positions where this is not the case (10 and 23), only one (23) shows conservation between the repeats within a given S polypeptide.

```
          1                          24
CTA.   IPFPPLQEQE RIIIRFTQLM SLCD
CTE.   IPFPPNTEQA RIVGTFSKLM FLCD
CTF.   MPIPPLNEQI RIVDTIDRLM SLCD
Consensus  -P-PP--EQ- RIV-----LM -LCD
```

Fig. 27. An alignment of the amino acid sequences of the central (CT) A repeat from EcoA (A), EcoE (E) and CfrA (F). These are less conserved than the equivalent regions from K-family members (see first 24 residues fo Fig. 18). Eleven of 13 positions at which all three A-family sequences are identical are positions at which at least one repeat from another family is identical to them (compare with Fig. 26).

respect, it has also been shown that a particular glutamine residue, though conserved in the central and carboxyl conserved regions (and B repeat) of all K-family S polypeptides, can be changed to a glutamate in either or both these regions of the S polypeptide of StySP without any noticeable affect on in vivo activity of the enzyme (my unpublished results).

The Argos repeat may be the only visible remnant of a gene duplication within K-family S polypeptides. An ancestral gene, encoding a polypeptide with a single recognition domain is thought to have duplicated, resulting in the present day organization (Gann et al, 1987). This duplication presumably occurred prior to a divergence into different families, otherwise it would have had to occur in each independently. The most feasible evolutionary pathway therefore appears to be: a common ancestral gene; duplication to generate the two recognition domain species; familial divergence; generation of new specificities within the families.

The genes encoding members of the A and K family appear to be allelic, as judged by their positions in the chromosomes (Daniel et al, 1988; G. Cowan and P. Kannan, unpublished observations). This encourages a belief in their sharing a common ancestor. The R124 family, however, are plasmid encoded. It is easier to envisage that in this latter instance, the genes were transferred to a plasmid from their original chromosomal location than it is that the hsd A and K genes evolved independently and moved to identical chromosomal locations.

As I have mentioned, the hugh divergence which has occurred in the recognition domains is acceptable if it is believed that extensive variation is necessary to achieve alternative specificities, but that evolution of new specificities is an advantage to a cell harbouring an R-M system (Levin, 1986).

Type I enzymes are particularly well designed for evolution of new specificities. The methylase and endonuclease act with the specificity determined by a common component. Also, because two recognition domains act together in defining the overall specificity, reassortment of these domains (e.g. StySQ and SJ) or alteration of their relative orientations (e.g. EcoR124 and R124/3) allows new specificities to be produced from pre-existing recognition domains (Fuller-Pace and Murray, 1986; Nagaraja et al, 1985; Gann et al, 1987; Price et al, 1988). Sequence divergence within these domains presumably generates new specificity domains which can be incorporated into the pool.

REFERENCES

Ackers, G.K., A.D. Johnson and M.A. Shea (1982) Proc. Natl. Acad. Sci., U.S.A., 79, 1129-1133.

Ackers, G.K., M.A. Shea and A.D. Johnson (1982) Proc. Natl. Acad. Sci., U.S.A., 79, 1129-1133.

Adler, S.P. and D. Nathans (1973) Biochim. Biophys. Acta., 299, 177-188.

Aggarwal, A.K., D.W. Rodgers, M. Drottar, M. Ptashne and S.C. Harrison (1988) Science, 242, 899-907.

Anderson, J.E., M. Ptashne and S.C. Harrison (1985) Nature, 316, 596-601.

Anderson, J.E., M. Ptashne and S.C. Harrison (1987) Nature, 326, 846-852.

Anderson, W.F., D.H. Ohlendorf, Y. Takeda and B.W. Matthews (1981) Nature, 290, 754-758.

Arber, W. and U. Kuhnlein (1967) Pathol. Microbiol., 30, 946-952.

Arber, W. and S. Linn (1969) Ann. Rev. Biochem., 38, 467-500.

Argos, P. (1985) Embo. J., 4 (5) 1351-1355.

Bass, S., V. Sorrells and P. Youderian (1988) Science, 242, 240-245.

Bass, S., P. Sugiono, D.N. Arvidson, R.P. Gunsalus and P. Youderian (1987) Genes and Dev., 1, 565-572.

Beckwith, J. (1987) In Escherishia Coli and Salmonella typhimurium : Cellular and Molecular Biology, Ed. F.C. Neidhardt, 1444-1452.

Benson, N., P. Sugiono and P. Youderian (1988) Genetics, 118, 21-29.

Benton, N.D. and R.W. Davies (1977) Science, 196, 180-182.

Berg, O.G., R.B. Winter and P.H. von Hippel (1981) Biochem., 20, 6929-6948.

Berkner, K.L. and N.R. Folk (1977) J. Biol. Chem., 252, 3185-3193.

Bickle, T.A. (1982) Nucleases, Cold Spring Harbor, 85-108.

Bickle, T.A. (1987) In Escherichia Coli and Salmonella typhimurium : Cellular and Molecular Biology, Vol. 2, Ed. F.C. Neidhardt, 692-696.

Bickle, T.A., C. Brack and R. Yuan (1978) Proc. Natl. Acad. Sci., U.S.A., 75, 3099-3103.

Biggin, M.D., T.J. Gibson and G.F. Hong (1983) Proc. Natl. Acad. Sci., U.S.A., 80, 3963-3965.

Bird, A.P. (1986) Nature, 321, 209-213.

Bodnar, J.W., W. Zempsky, D. Warder, C. Bergson and D.C. Ward (1983) J. Biol. Chem., 258, 15206-15213.

Boelens, R., R.M. Schleef, J.M. Van Boom and R. Kaptein (1987) J. Mol. Biol., 193, 213-216.

Boyer, H.W. (1971) Ann. Rev. Microbiol., 25, 153-176.

Boyer, H.W. and D. Roulland-Dussoix (1969) J. Mol. Biol., 41, 459-472.

Brammar, W.J., N.E. Murray and S. Winton (1974) J. Mol. Biol., 90, 633-647.

Brennan, C.A., M.D. Van Cleve and R.I. Gumport (1986a) J. Biol. Chem., 261, 7270-7278.

Brennan, C.A., M.D. Van Cleve and R.I. Gumport (1986b) J. Biol. Chem., 261, 7279-7286.

Brent, R. and M. Ptashne (1985) Cell, 43, 729-736.

Brunelle, A. and R.F. Schleif (1987) Proc. Natl. Acad. Sci., U.S.A., 84, 6673-6676.

Buhler, R. and R. Yuan (1978) J. Biol. Chem., 253, 6756-6760.

Bullas, L.R., C. Colson and A. Van Pel (1976) J. Gen. Microbiol., 95, 166-172.

Bullas, L.R., C. Colson and B. Neufeld (1980) J. Bact., 141 (1), 275-292.

Burckhardt, J., J. Weisemann and R. Yuan (1981a) J. Biol. Chem., 256 (8), 4024-4032.

Burckhardt, J., J. Weisemann, D.L. Hamilton and R. Yuan (1981b) J. Mol. Biol., 153, 425-440.

Bushman, F.D., J.E. Anderson, S.C. Harrison and M. Ptashne (1985) Nature, 316, 651-653.

Bushman, F.D. and M. Ptashne (1988) Cell, 54, 191-197.

Caruthers, M.H. (1986) In Protein Structure Folding and Design, Ed. D.L. Oxender, Liss, New York, 221-228.

Chandrasegaran, S. and H.O. Smith (1988) In Structure and Expression, Vol. 1, From Proteins to Ribosomes, Eds. R.H. Sarma and M.H. Sarma, Adenine Press, 149-156.

Clewell, D.B. and D.R. Helinski (1969) Proc. Natl. Acad. Sci., U.S.A., 62, 1159-1166.

Connolly, B.A., F. Eckstein and A. Pingoud (1984) J. Biol. Chem., 259, 10760-10763.

Dallmann, G., P. Papp and L. Orosz (1987) Nature, 330, 398-401.

Daniel, A.S., F.V. Fuller-Pace, D.M. Legge and N.E. Murray (1988) J. Bact., 170, 1775-1782.

Dente, L., G. Cesareni and R. Cortese (1983) Nucl. Acids Res., 11, 1645-1655.

Dickerson, R. (1983a) Sci. Am. 249 94

Dickerson, R. (1983b) J. Mol. Biol., 166, 419-441.

Dickerson, R. and H. Drew (1981) J. Mol. Biol., 149, 761-786.

Dodson, M., F.B. Dean, P. Bullock, H. Echols and J. Hurwitz (1987) Science, 238, 964-967.

Dodson, M., H. Echols, S. Wickner, C. Alfano, K. Mensa-Wilmot, B. Gomes, J. LeBowitz, J.D. Roberts and R. McMacken (1986) Proc. Natl. Acad. Sci., U.S.A., 83, 7638-7642.

Dodson, M., J. Roberts, R. McMacken and H. Echols (1985) Proc. Natl. Acad. Sci., U.S.A., 82, 4678-4682.

Dunn, J.J. and F.W. Studier (1983) J. Mol. Biol., 166, 477-535.

Dwyer-Hallquist, P., F.J. Kezdy and K.L. Agarwal (1982) Biochem., 21, 4693-4700.

Ebright, R.H., A. Cossart, B. Gicquel-Sanzey and J. Beckwith (1984) Nature, 311, 232-235.

Echols, H. (1986) Science, 233, 1050-1056.

Ehbrecht, H.J., A. Pingoud, C. Urbanke, G. Maass and C. Gualerzi (1985) J. Biol. Chem., 260, 6160-6166.

Eisenbeis, S.J., M.S. Nasoff, S.A. Noble, L.P. Bracco, D.R. Dodds and M.H. Caruthers (1985) Proc. Natl. Acad. Sci., U.S.A., 82, 1084-1088.

Eliason, J., M.A. Weiss and M. Ptashne (1985) Proc. Natl. Acad. Sci., U.S.A., 82, 2339-2343.

Elliot, J. and W. Arber (1978) Mol. Gen. Genet., 161, 1-8.

Emmons, S.W., V. MacCosham and R.W. Baldwin (1975) J. Mol. Biol., 91, 133-146.

Endlich, B. and S. Linn (1981) In The Enzymes, 3rd edition, Ed. P.D. Boyer, 14, Part A, Academic Press, New York, 137-156.

Endlich, B. and S. Linn (1985) J. Biol. Chem., 260, 5720-5728.

Eskin, B. and S. Linn (1972a) J. Biol. Chem., 247, 6183-6191.

Eskin, B. and S. Linn (1972b) J. Biol. Chem., 247, 6192-6196.

Franklin, N.C. and W.F. Dove (1969) Genet. Res. Camb., 14, 151-157.

Frederick, C.A., J. Grable, M. Melia, C. Samudzi, L. Jen-Jacobson, B.-C. Wang, P. Greene, H.W. Boyer and J.M. Rosenberg (1984) Nature, 309, 327-331.

Fuller-Pace, F.V., L.R. Bullas, H. Delius, N.E. Murray (1984) Proc. Natl. Acad. Sci., 81, 6095-6099.

Fuller-Pace, F.V., G.M. Cowan and N.E. Murray (1985) J. Mol. Biol., 186, 65-75.

Fuller-Pace, F.V. and N.E. Murray (1986) Proc. Natl. Acad. Sci., 83, 9368-9372.

Gann, A.A.F., A.J.B. Campbell, J.F. Collins, A.F.W. Coulson and N.E. Murray (1987) Mol. Microbiol., 1, 13-22.

Glover, S.W. (1970) Genet. Res. Camb., 15, 237-250.

Glover, S.W. and C. Colson (1969) Genet. Res. Camb., 13, 227-240.

Gough, J.A. and N.E. Murray (1983) J. Mol. Biol., 166, 1-19.

Greene, P.J., M. Gupta, H.W. Boyer, W.E. Brown and J.M. Rosenberg (1981) J. Biol. Chem., 256, 2143-2153.

Greene, P.J., B.T. Ballard, F. Stephenson, W.J. Kohr, H. Rodriguez, J.M. Rosenberg and H.W. Boyer (1988) Gene, 68, 43-52.

Gussin, G., A. Johnson, C. Pabo and R. Sauer (1983) In Lambda II, Ed. Hendrix, Roberts, Stahl and Weisberg, CSH Publications, New York, 93-121.

Haberman, A., J. Heywood and M. Meselson (1972) Proc. Natl. Acad. Sci., U.S.A., 69, 3138-3141.

Hadi, S.M. and R. Yuan (1974) J. Biol. Chem., 249, 4580-4586.

Hansen, E.B., T. Atlung, F.G. Hansen, O. Skougaard and K. Van Meyenburg (1984) Mol. Gen. Genet., 196, 387-396.

Hartman, N. and N.D. Zinder (1974) J. Mol. Biol., 85, 345-356.

Hecht, M.H., H.C.M. Nelson and R.T. Sauer (1983) Proc. Natl. Acad. Sci., U.S.A., 80, 2676-2680.

Hecht, M.H. and R.T. Sauer (1985) J. Mol. Biol., 186, 53-63.

Helmann, J.D. and M.J. Chamberlin (1988) Ann. Rev. Biochem., 57, 839-872.

Ho, Y.S., M.E. Mahoney, D.L. Wulff and M. Rosenberg (1988) Genes and Dev., 2, 184-195.

Hochschild, A., J. Douhan III and M. Ptashne (1986) Cell, 47, 807-816.

Hochschild, A. and M. Ptashne (1986a) Cell, 44, 925-933.

Hochschild, A. and M. Ptashne (1986b) Cell, 44, 681-687.

Hollis, M., D. Valenzvala, D. Pioli, R. Wharton and M. Ptashne (1988) Proc. Natl. Acad. Sci., U.S.A., 85, 5834-5838.

Horiuchi, K. and N.D. Zinder (1972) Proc. Natl. Acad. Sci., U.S.A., 69, 3220-3224.

Hu, N. and J. Messing (1982) Gene, 17, 271-277.

Hubacek, J. and S.W. Glover (1970) J. Mol. Biol., 50, 111-127.

Humayun, Z., D. Klein and M. Ptashne (1977) Nucl. Acids. Res., 4, 1595-1607.

Ish-Horowitz, D. and J.F. Burke (1981) Nucl. Acids Res., 9, 2989-2998.

Jan-Jacobson, L., D. Lesser and M. Kurpiewski (1986) Cell, 45, 619-629.

Joachimiak, A.J., R.L. Kelley, P.R. Gunsalus, C. Yanofsky and P.B. Sigler (1983) Proc. Natl. Acad. Sci., U.S.A., 80, 668-672.

Johnson, A., R.J. Meyer and M. Ptashne (1978) Proc. Natl. Acad. Sci., U.S.A., 75, 1783-1787.

Johnson, A.D., B.J. Meyer and M. Ptashne (1979) Proc. Natl. Acad. Sci., U.S.A., 76, 5061-5065.

Johnson, A.D., A.R. Poteete, G. Latter, R.T. Sauer, G.K. Ackers and M. Ptashne (1981) Nature, 294, 217-233.

Jones, N.C., P.W.J. Rigby and E.B. Ziff (1988) Genes and Dev., 2, 267-281.

Jordan, S.R. and C.O. Pabo (1988) Science, 242, 893-899.

Kaplan, D.A. and D.P. Nierlich (1975) J. Biol. Chem., 250, 2395-2397.

Kelley, R.L. and C. Yanofsky (1985) Proc. Natl. Acad. Sci., U.S.A., 82, 483-487.

Kornberg, A. (1982) Supplement to DNA Replication, Freeman, San Francisco.

Koudelka, G.B., P. Harbury, S.C. Harrison and M. Ptashne (1988) Proc. Natl. Acad. Sci., U.S.A., 85, 4633-4637.

Koudelka, G.B., S.C. Harrison and M. Ptashne (1987) Nature, 326, 886-891.

Kramer, W., V. Drutsa, H.-W. Jansen, B. Kramer, M. Pflugfelder and H.-J. Fritz (1984) Nucl. Acids Res., 12, 9441-9456.

Kuhnlein, V., S. Linn and W. Arber (1969) Proc. Natl. Acad. Sci., U.S.A., 63, 556-562.

Lautenberger, J.A., N.C. Kan, D. Lackey, S. Linn, M.H. Edgell and C.A. Hutchinson III (1978) Proc. Natl. Acad. Sci., U.S.A., 75, 2271-2275.

Lautenberger, J.A. and S. Linn (1972) J. Biol. Chem., 247, 6176-6182.

Lehming, N., J. Sartorius, M. Niemoller, G. Genenger, B.V. Wilcken-Bergmann and B. Muller-Hill (1987) EMBO J., 6, 3145-3153.

Levin, B.R. (1986) In Evolutionary Processes and Theory, Eds. S. Karlin and E. Nero, Academic Press, Inc., New York, 669-688.

Lin, S. and A.D. Riggs (1975) Cell, 4, 107-111.

Loenen, W.A.M., A.S. Daniel, H.D. Braymer and N.E. Murray (1987) J. Mol. Biol., 198, 159-170.

Lu, A.-L., W.E. Jack and P. Modrich (1981) J. Biol. Chem., 256, 13200-13206.

Lutz, C.T., N.C. Hollifield, B. Seed, J.M. Davie and H.J. Huang (1987) Proc. Natl. Acad. Sci., U.S.A., 84, 4379-4383.

Mandel, M. and A. Higa (1970) J. Mol. Biol., 53, 159-162.

Maniatis, T., E.F. Fritsch and J. Sambrock (1982) In Molecular Cloning, A Laboratory Manual, CSH Publications, New York.

Maniatis, T.

Mann, M.B., R.N. Rao and H.O. Smith (1978) Gene, 3, 97-112.

Marchionni, M.A. and D.J. Roufa (1978) J. Biol. Chem., 253, 9075-9081.

Marinus, M. (1987) In Escherishia Coli and Salmonella typhimurium : Molecular and Cellular Aspects, Ed. F.C. Neidhardt, 697-702.

McClarin, J.A., C.A. Frederick, B.-C. Wang, P. Greene, H.W. Boyer, J. Grable and J.M. Rosenberg (1986) Science, 234, 1526-1541.

McClelland, M. and M. Nelson (1988) Gene, 74, 291-304.

McKay, D., I. Weber and T. Steitz (1982) J. Biol. Chem., 257, 9518-9524.

McKay, D.B. and T.A. Steitz (1981) Nature, 290, 754-758.

Meselson, M. and R. Yuan (1968) Nature, 217, 1110-1114.

Meselson, M., R. Yuan and J. Heywood (1972) Ann. Rev. Biochem., 41, 447-466.

Modrich, P. (1979) Q. Rev. Biophys., 12, 315-369.

Modrich, P. and R.J. Roberts (1982) In Nucleases, Eds. S.M. Linn and R.J. Roberts, Cold Spring Harbor Laboratory, New York, 109-154.

Modrich, P. and R.A. Rubin (1977) J. Biol. Chem., 252, 7273-7278.

Moitoso de Vargas, C.A. Pargellis, N.M. Hasan, E.W. Bushman and A. Landy (1988) Cell, 54, 923-929.

Murray, N.E., P.L. Batten and K. Murray (1973) J. Mol. Biol., 81, 395-407.

Murray, N.E., J.A. Gough, B. Suri and T.A. Bickle (1982) EMBO J., 1 (5), 535-539.

Nagaraja, V., J.C.W. Shepherd and T.A. Bickle (1985a) Nature, 316, 371-372.

Nagaraja, V., J.C.W. Shepherd, T. Pripfl and T.A. Bickle (1985b) J. Mol. Biol., 182, 579-587.

Nagaraja, V., M. Steiger, C. Nager, S.M. Hadi and T.A. Bickle (1985c) Nuc. Acids Res., 13, 389-399.

Nelson, H.C.M., J.T. Finch, B.F. Luisi and A. Klug (1987) Nature, 330, 221-226.

Nelson, H.C.M., M.H. Hecht and R.T. Sauer (1983) CSH Symp. Quant. Biol., 47, 441-449.

Nelson, H.C.M. and R.T. Sauer (1985) Cell, 42, 549-558.

Newman, A.K., R.A. Rubin, S.-H. Kim and P. Modrich (1981) J. Biol. Chem., 256, 2131-2142.

Ohlendorf, D.H., W.F. Anderson, R.G. Fisher, Y. Takeda and B.W. Matthews (1982) Nature, 298, 718-723.

Ohlendorf, D.H., W.F. Anderson, M. Lewis, C.O. Pabo and B.W. Matthews (1983) J. Mol. Biol., 169, 757-769.

Otwinowski, Z., R.W. Schevitz, R.-G. Zhang, C.L. Lawson, A. Joachimiak, R.Q. Marmorstein, B.F. Luisi and P.B. Sigler (1988) Nature, 335, 321-329.

Pabo, C.O., W. Krovatin, A. Jeffrey and R.T. Sauer (1982) Nature, 298, 441-443.

Pabo, C.O. and M. Lewis (1982) Nature, 298, 443-447.

Pabo, C.O. and R.T. Sauer (1984) Ann. Rev. Biochem., 53, 293-321.

Piekarowicz, A. and J.D. Goguen (1986) Eur. J. Biochem., 154, 295-298.

Polisky, B., P. Greene, D.E. Garfin, B.J. McCarthy, H.M. Goodman and H.W. Boyer (1975) Proc. Natl. Acad. Sci., U.S.A., 72, 3310-3314.

Ptashne, M. (1986a) Nature, 322, 697-701.

Ptashne, M. (1986b) A Genetic Switch - gene control and phage lambda, Blackwell Scientific Publications and Cell Press.

Ptashne, M. (1988) Nature, 335, 683-689.

Ptashne, M., A. Jeffrey, A.D. Johnson, R. Maurer, B.J. Meyer, C.O. Pabo, T.M. Roberts and R.T. Sauer (1980) Cell, 19, 1-11.

Price, C., T. Pripfl and T.A. Bickle (1987a) Eur. J. Biochem., 167, 111-115.

Price, C., J.C.W. Shepherd and T.A. Bickle (1987b) EMBO J., 6, 1493-1497.

Ravetch, J.V., K. Horiuchi and N.D. Zinder (1978) Proc. Natl. Acad. Sci., U.S.A., 75, 2266-2270.

Rosamond, J., B. Endlich and S. Linn (1979) J. Mol. Biol., 129, 619-635.

Roy, P.H. and H.O. Smith (1973) J. Mol. Biol., 81, 445-459.

Sain, B. and N.E. Murray (1980) Molec. Gen. Genet., 180, 35-46.

Sanger, F., A.R. Coulson, B.G. Barrell, A.J.H. Smith and B.A. Roe (1980) J. Mol. Biol., 143, 161-178.

Sanger, F., A.R. Coulson, G.F. Hong, D.F. Hill and G.B. Petersen (1982) J. Mol. Biol., 162, 729-733.

Sanger, F., S. Nicklen and A.R. Coulson (1977) Proc. Natl. Acad. Sci., U.S.A., 74, 5463-5467.

Sauer, R.T., C.O. Pabo, B.J. Meyer, M. Ptashne and K.D. Backman (1979) Nature, 279, 396-400.

Sauer, R.T., R.R. Yocum, R.F. Doolittle, M. Lewis and C.O. Pabo (1982) Nature, 298, 447-451.

Schevitz, R.W., Z. Otwinowski, A. Joachimiak, C.L. Lawson and P.B. Sigler (1985) Nature, 317, 782-786.

Seaman, N.C., J.M. Rosenberg and A. Rich (1976) Proc. Natl. Acad. Sci., U.S.A., 73, 804-808.

Simon, V.F. and S. Lederberg (1973) Biochem., 12, 3055-3063.

Spiro, S. and J.R. Guest (1987) Molec. Microbiol., 1, 53-58.

Sternberg, N. (1976) Virology, 73, 139-154.

Studier, F.W. and P.K. Bandyopadhyay (1988) Proc. Natl. Acad. Sci., U.S.A., 85, 4677-4681.

Suri, B. and T.A. Bickle (1985) J. Mol. Biol., 186, 77-85.

Suri, B., V. Nagaraja and T.A. Bickle (1984a) Curr. Top. Microbiol. Immunol., 108, 1-9.

Suri, B., J.C.W. Shepherd and T.A. Bickle (1984b) EMBO J., 3, 575-579.

Suskind, M. and P. Youderian (1983) In Lambda II, Ed. Hendrix, Roberts, Stahl and Weisberg, CSH Publications, New York, 347-364.

Suttcliffe, J.G. (1979) CSH Symp. Quant. Biol., 43, 77-90.

Thomas, C.A. and J. Abelson (1966) In Procedures in Nucleic Acid Research, Eds. G. Cantoni and D. Davies, Harper and Row, New York, 1, 553-561.

Tullins, T.D. and B.A. Dombroski (1986) Proc. Natl. Acad. Sci., U.S.A., 83, 5469-5473.

van Ormondt, H., J.A. Lautenberger, S. Linn and A. de Waard, (1973) Febs Letters, 33, 177-180.

Vershon, A.K., S.-M. Liao, W.R. McClure and R.T. Sauer (1987) J. Mol. Biol., 195, 323-331.

Vershon, A.K., P. Youderian, M.A. Weiss, M.M. Suskind and R.T. Sauer (1985) In Specificity in Transcription and Translation, Ed. R. Calendar and L. Gold, Alan Liss, Inc., New York, 209-218.

Vieira, J. and J. Messing (1982) Gene, 19, 259-268.

von Hippel, P.H., G.D. Bear, W.D. Morgan and J.A. McSwiggen (1984) Ann. Rev. Biochem., 53, 389-446.

von Hippel, P.H., A. Revzin, C.A. Gross and A.C. Wang (1974) Proc. Natl. Acad. Sci., U.S.A., 71, 4808-4812.

Vovis, G.F., K. Horiuchi and N. Zinder (1974) Proc. Natl. Acad. Sci., U.S.A., 71, 3810-3813.

Vovis, G.F. and N.D. Zinder (1975) J. Mol. Biol., 95, 557-568.

Wasylyk, B. (1988) CRC Crit. Rev. Biochem., 23, 77-120.

Weisberg, R. and A. Landy (1983) In Lambda II, Eds. Hendrix, Roberts, Stahl and Weisberg, CSH Publications, New York, 211-250.

Wharton, R.P., E.L. Brown and M. Ptashne (1984) Cell, 38, 361-369.

Wharton, R.P. and M. Ptashne (1985) Nature, 316, 601-605.

Wharton, R.P. and M. Ptashne (1987) Nature, 326, 888-891.

Wilson, G.G., V.I. Tanyashin and N.E. Murray (1977) Mol. Gen. Genet., 156, 203-214.

Wilson, G.C. (1988) Gene, 74, 281-290.

Wolberger, C., Y. Dong, M. Ptashne and S.C. Harrison (1988) Nature, 335, 789-795.

Wolfes, H., A. Fleiss and A. Pingoud (1985) Eur. J. Biochem., 150, 105-110.

Woodhead, J.L., N. Bhave and A.D.B. Malcolm (1981) Eur. J. Biochem., 115, 293-296.

Yamamoto, K.R., B.M. Alberts, R. Benzinger, L. Lawhorne and G. Treiber (1970) Virology, 40, 734-744.

Yanisch-Perron, C., J. Vieira and J. Messing (1985) Gene, 33, 103-119.

Yanofsky, C. and I.P. Crawford (1987) In Escherichia Coli and Salmonella typhimurium : Cellular and Molecular Biology, Vol. 2, Ed. F.C. Neidhardt, 1453-1472.

Yanofsky, S.D., R. Love, J.A. McClarin, J.M. Rosenberg, H.W. Boyer and P.J. Greene (1987) PROTEINS : Structure, Function and Genetics, 2, 273-282.

Youderian, P., A.K. Vershon, S. Bouvier, R.T. Sauer and M.M. Suskind (1983) Cell, 35, 777-783.

Yuan, R. (1981) Ann. Rev. Biochem., 50, 285-315.

Yuan, R., T.A. Bickle, W. Ebbers and C. Brack (1975) Nature, 256, 556-560.

Yuan, R., D.L. Hamilton and J. Burckhardt (1980) Cell, 20, 237-244.

Yuan, R. and M. Meselson (1970) Proc. Natl. Acad. Sci., U.S.A., 65, 357-362.

Zhang, R.G., A. Joachimiak, C.L. Lawson, R.W. Schevitz, Z. Otwinowski and P.B. Sigler (1987) Nature, 327, 591-597.

Zissler, J.E., E. Signer and F. Schaefer (1971) In The Bacteriophage Lambda, Ed. A.D. Hershey, CSH Publications, New York, 455-468.

Zoller, M.J. and M. Smith (1983) Meth. Enzym., 100, 468-500.