

# NUCLEOTIDE SEQUENCE SPECIFICITY OF RESTRICTION ENDONUCLEASES

Nobel Lecture, 8 December, 1978

by

HAMILTON O. SMITH

Department of Microbiology, The Johns Hopkins University School of Medicine Baltimore, Maryland, U. S. A.

## INTRODUCTION

In the past seven to eight years we have witnessed the development of a new DNA technology that has fundamentally altered our approach to modern genetics. The basic ingredients of this new technology are the cleavage-site-specific restriction enzymes: a special class of bacterial endonucleases that can recognize specific nucleotide sequences in duplex DNA and produce double-stranded cleavages. Using a collection of these enzymes, each with its own particular sequence specificity, DNA molecules may be cleaved into unique sets of fragments useful for DNA sequencing, chromosome analysis, gene isolation, and construction of recombinant DNA. The latter, combined with the concept of molecular cloning, has given birth to the new field of genetic engineering, and from this are expected many new and exciting medical and research applications.

My own role in these developments occurred primarily in the period of 1968-1970 when my colleagues and I made the chance discovery of the first of the cleavage-site-specific restriction enzymes. I should like to briefly describe this work in historical context as it leads naturally into the main part of my lecture describing our present knowledge of restriction and modification enzymes. I shall not go into the many applications as these have been reviewed elsewhere (1). However, I should like to describe in some detail the use of these enzymes as model systems for studying sequence-specific protein-DNA interactions since this is one of our major research interests.

### *Restriction and Modification in Bacteria: the Discovery of Restriction Enzymes*

The observations leading to the discovery of restriction enzymes span a period of nearly two decades and constitute a prime example of how basic research on an apparently insignificant bacteriological phenomenon has had unexpectedly far-reaching implications. The story begins in the early 1950s with some observations by Luria and Human (1a) and Bertani and Weigle (2) concerning the curious behavior of phage grown on two different strains of bacteria. Phage propagated on one strain were found to grow poorly (were "restricted") on the second, and vice-versa. However, a few phage always escaped restriction and could then grow well on the new host. They apparently had acquired some type of host-specific modification that then protected them from the restriction effects of the host. The

biochemical basis of this phenomenon remained a mystery until the early 1960s when Werner Arber and co-workers were able to show that host-specific modifications was carried on the phage DNA (3), and that restriction was associated with degradation of the phage DNA (4). In a remarkably prophetic review in 1965, Arber postulated the existence of site-specific restriction enzymes and suggested that modification might be produced by hostspecific DNA methylases (5). Thus, the notion became established that each restriction and modification (R-M) system in bacteria was composed of two enzymes with identical specificity: a restriction endonuclease that recognized short nucleotide sequences and cleaved DNA, and a modification enzyme that recognized the same sequence and modified it to protect against cleavage. In this way, the host cell DNA would be protected but foreign DNA entering from outside with improper modification would be cleaved and destroyed.

Although the existence of restriction enzymes was predicted with confidence by 1965, it was not until early 1968 that Linn and Arber (6) actually found an activity in extracts of *E. coli* B with the expected properties of such an enzyme. At the same time, Meselson and Yuan (7) reported more extensive experiments with a highly purified restriction endonuclease from *E. coli* K. Using sucrose gradient centrifugation, the latter demonstrated that their enzyme cleaved unmodified phage  $\lambda$  DNA into large fragments while modified DNA remained undegraded. An unusual feature of the enzyme was its requirement for the cofactors S-adenosylmethionine, ATP, and  $Mg^{2+}$ . Meselson and Yuan assumed from Arber's work that their enzyme was attacking the  $\lambda$  DNA at fixed sites, but were unable to confirm this by sucrose gradient analysis of the fragment species.

It is now known that restriction enzymes of *E. coli* B and K are examples of a class of restriction enzymes that do not cleave DNA at specific sites, although this fact was not appreciated for several years. Such Class I enzymes are complex, multimeric proteins that generally require ATP, S-adenosylmethionine and  $Mg^{2+}$  as cofactors and function both as restriction endonucleases and as modification methylases (8). Although they recognize specific sites in the DNA, they cleave randomly at a considerable distance from the recognition site (9, 10, 11). Because of this property, they have not proven useful as enzymatic tools for DNA analysis.

#### *Discovery of a Cleavage-Site-Specific Restriction Endonuclease in Haemophilus influenzae Rd*

In early Spring of 1968, I read the Meselson and Yuan paper with great interest. Their work imparted, in a very explicit way, a sense of biochemical reality to Arber's observations. I had at that time recently joined the faculty of the Department of Microbiology at Johns Hopkins and, with a young graduate student named Kent Wilcox, was just beginning to explore genetic recombination in *Haemophilus influenzae*, strain Rd, an efficiently transformable bacterium that we had been introduced to by Roger Herriott of the School of Hygiene. Some of our experiments employed a

viscometer as a particularly sensitive measure of endonucleolytic cleavage of DNA by cell extracts. Other experiments involved recovery of donor DNA from cells after uptake. In one such experiment we happened to use labeled DNA from phage P22, a bacterial virus I had worked with for several years before coming to Hopkins. To our surprise, we could not recover the foreign DNA from the cells. With Meselson's recent report in our minds, we immediately suspected that it might be undergoing restriction, and our experience with viscometry told us that this would be a good assay for such an activity. The following day, two viscometers were set up, one containing P22 DNA and the other, *Haemophilus* DNA. Cell extract was added to each and we began quickly taking measurements. As the experiment progressed, we became increasingly excited as the viscosity of the *Haemophilus* DNA held steady while the P22 DNA viscosity fell. We were confident that we had discovered a new and highly active restriction enzyme. Furthermore, it appeared to require only  $Mg^{2+}$  as a cofactor, suggesting that it would prove to be a simpler enzyme than that from *E. coli* K or B. From that point on, other work in the laboratory was shelved while we turned our full attention to the isolation and study of the new enzyme.

After several false starts and many tedious hours with our laborious, but sensitive viscometer assay, Wilcox and I (12) succeeded in obtaining a purified preparation of the restriction enzyme. We next used sucrose gradient centrifugation to show that the purified enzyme selectively degraded duplex, but not single-stranded, P22 DNA to fragments averaging around 100 base pairs in length, while *Haemophilus* DNA present in the same reaction mixture was untouched. No free nucleotides were released during the reaction, nor could we detect any nicks in the DNA products. Thus, the enzyme was clearly an endonuclease that produced double-strand breaks and was specific for foreign DNA. Since the final (limit) digestion products of foreign DNA remained large, it seemed to us that cleavage must be site-specific. This proved to be case and we were able to demonstrate it directly by sequencing the termini of the cleavage fragments.

#### *Sequencing the Recognition Site*

We began our sequencing efforts in late 1968 using a method that had been worked out by Bernard Weiss and Charles C. Richardson at Harvard. The method involved labeling the 5'-termini of DNA with radioactive phosphorus using T4 polynucleotide kinase and  $^{32}P$  gamma-labeled ATP, followed by digestion with pancreatic DNase and either venom phosphodiesterase to yield terminal nucleotides, or exonuclease I to yield terminal dinucleotides (13). These could then be separated by electrophoresis and identified by comparison with known marker nucleotides.

Weiss had come to Hopkins in the Fall of 1967 and occupied a neighboring laboratory in the Microbiology Department. He instructed us in the procedure and supplied us with both the kinase and the  $^{32}P$ -gamma-labeled

ATP. We started our sequencing using restriction enzyme digests of phage T7 DNA, a fortuitous choice as we later learned. In our first experiment, we found it necessary to treat with alkaline phosphatase to remove a terminal 5'-phosphoryl group from the cleavage fragments in order to obtain labeling; thus cleavage of the DNA chain produced 3'-hydroxyl, 5'-phosphoryl termini. We next examined the terminal nucleotides and found terminal  $^{32}\text{P}$  label appearing only in dGMP and dAMP. Thus, our enzyme was specific!

We were ready by early 1969 to proceed to the dinucleotide level. Unfortunately, just at this most exciting stage of the work, Wilcox received his draft notice from the Army and was forced to discontinue the work so that he could complete his formal requirements for the Master's degree. Meanwhile, I began occupying myself, for the space of several months, with the laborious preparation of the dinucleotide standards that would be necessary for identification of the terminal dinucleotides. I also prepared a supply of exonuclease I from a side fraction of a large DNA polymerase preparation generously given to me by Paul Englund. Using the exonuclease I and the standards, I proceeded to show by the Weiss and Richardson method that the terminal dinucleotide was either (5')pGpA or (5')pApA, further confirming the remarkable cleavage specificity of our restriction enzyme. I believed that extension of the analysis beyond this point was possible, but standard oligonucleotides with which to identify the longer terminal species were unavailable.

About this time, Thomas J. Kelly, Jr. joined my laboratory, and in a series of discussions we worked out an approach, using the newly available isotope  $^{33}\text{P}$  that was to prove successful. In this approach, T7 DNA was uniformly  $^{32}\text{P}$ -labeled, cleaved with the restriction endonuclease, 5'-terminally labeled with the second isotope,  $^{33}\text{P}$ , and then digested to oligomers using pancreatic DNase. The products were fractionated according to length by ion-exchange chromatography, and we then analyzed the oligonucleotides of each size class electrophoretically. Two  $^{32}\text{P}$ -terminal,  $^{33}\text{P}$ -uniformly labeled species were obtained at the dimer and trimer level, but at least six species out of a possible eight were identified at the tetramer level, so that specificity was lost at that point. The dimer and trimer species were eluted from the electrophoretic strip, digested with venom phosphodiesterase and the  $^{33}\text{P}$ -labeled nucleotides identified. In this way, the 5'-terminal dinucleotide was again confirmed as pPu-A and the trinucleotide was found to be pPu-A-C.

Three possible sequence arrangements could account for our result depending on whether cleavage was "even" or "staggered." To resolve this, we used micrococcal nuclease to release the 3'-terminal dinucleoside monophosphate and this turned out to be unique and complementary to the 5'-terminal dinucleotide. Thus, our enzyme recognized the 2-fold rotationally symmetrical six nucleotide sequence: ... (5') G-T-Py↓ Pu-A-C (3') ...  
... (3') C-A-Pu↑ Py-T-G (5') ... and produced an even duplex cleavage as indi-

cated by the arrows (14). Based on the expected occurrence of this sequence in random DNA, the enzyme would be expected to cut once every 1024 base pairs: a value in good accord with our previous observations. The most interesting feature of the sequence was its symmetry, and we speculated that this might have important implications for the restriction enzyme structure (see later sections).

In retrospect the proof of cleavage specificity was clearly our most important result. It had not been shown for the *E. coli* K and B enzymes, and, in fact, could not have been shown since these were randomly cleaving Class I enzymes. Our enzyme belonged to a different, and as we shall see, a much larger class of restriction enzymes. Such Class II enzymes (8) are cleavage-site-specific and require only  $Mg^{2+}$  as cofactor. Later studies have revealed that they are relatively simple proteins, existing typically as dimers or tetramers of a single polypeptide chain (15, 16, 17), and their corresponding modification methylases are separate proteins that exist in some cases as monomers (18). However, in 1970 when we completed the sequence work, only Class I methylases had been studied *in vitro* (8), so we turned next to the isolation of a modification methylase from *H. influenzae* Rd.

#### *Modification Methylases in H. influenzae Rd*

While we did not hesitate to call our cleavage-site-specific endonuclease a restriction enzyme—after all it was specific for foreign DNA—formal objections existed to that classification unless a modification enzyme of the same specificity could be found. In the absence of such a modifying enzyme to protect host sequences against cleavage, it would be necessary to postulate total absence of the sequence in the cell chromosome (a rather remote possibility) or, alternatively, a compartmentalization of the activity. To allay these objections, and to satisfy our own curiosity, Paul Roy (a graduate student) and I undertook a survey of the DNA methylases in *H. influenzae* Rd in late 1970 (19, 20).

We first established that *H. influenzae* Rd like many other strains of bacteria (21), contains a small percentage of methylated bases in its DNA: 5-methylcytosine occurs once per about 8 000 bases, and N-6-methyladenine is found once per about 280 bases. We realized that much of this methylation might be unrelated to R-M systems and that several methylases could be present. Arber had shown that in *E. coli* the majority of DNA methylation was not associated with R-M systems; in *E. coli* B, as little as 5% was so involved (22). With this in mind, we adopted a general approach designed to reveal the total DNA methylases of the cells. Proteins from a crude cell extract were chromatographed on phosphocellulose and assayed for ability to transfer [ $^3H$ ]methyl groups from labeled S-adenosylmethionine onto salmon sperm DNA or T7 DNA. In this way, four DNA adenine methylases were detected.

One of these methylases protected T7 DNA from cleavage by our restriction enzyme; and conversely, the sites for this enzyme in salmon

sperm DNA were destroyed by predigestion with our restriction enzyme preparation. These two results together indicated that both the restriction enzyme and the methylase shared common DNA recognition sites. As a further proof that we had isolated the modification methylase, we analyzed 3' and 5' nearest neighbors to the <sup>3</sup>H-methylated adenine residues produced by the enzyme. The results gave as the partial sequence for the methylase, the trinucleotide (5')Pu-A-C, in direct agreement with our restriction enzyme sequence.

One additional and unexpected observation came out of our methylase studies. In the salmon sperm DNA experiment in which we predigested with our restriction enzyme preparation, methyl acceptor ability was also lost for one of the other DNA methylases. This particular methylase was not active on T7 DNA unlike our restriction enzyme. It appeared that a second restriction enzyme, corresponding in specificity to this methylase, was contained in our endonuclease preparation. Our interpretation was confirmed by separation of these two restriction activities in Nathans' laboratory, and by a letter from Kenneth Murray of Edinburgh, Scotland reporting that he had purified the new enzyme and determined its recognition site sequence as (5')pA-A-G-C-T-T(23). At this point it is interesting to recall our choice of T7 DNA for the sequencing work. Since, unbeknownst to us, we had worked with a mixture of two enzymes, it was indeed fortuitous that only one was active on T7 DNA.

#### *Search for New Restriction Enzymes*

Through the work of Nathans and colleagues (24), beginning in 1969, in which they applied cleavage-site-specific restriction endonucleases to the analysis of the SV40 tumor virus genome, it became clear that these enzymes were valuable tools for DNA analysis. Their work provided an early stimulus to the search for new enzymes of differing specificities. An important consequence of our work with the *Haemophilus* restriction endonuclease was the realization that such enzymes could be readily detected in bacteria by purely biochemical procedures. This was especially true after the introduction of gel electrophoresis by Nathans for analysis of DNA restriction cleavage fragments (24) and the introduction of ethidium bromide as a fluorescent stain for DNA by Sharp *et al.* (25). Armed with easy and specific assays, other laboratories were soon reporting new restriction endonucleases, first in *E. coli* (26) and other *Haemophilus* species (25, 27, 28), and then in a variety of other bacteria. Richard J. Roberts of the Cold Spring Harbor Laboratories, who had a special interest in their possible use for DNA sequencing, spearheaded the drive to isolate new enzymes. Bacteria available from the American Type Culture Collection were systematically examined for cleavage-site-specific endonucleases that would digest phage lambda and other viral DNAs so as to produce discrete bands on electrophoretic gels. Many other laboratories joined in the effort, and today some 150 enzymes with nearly 50 different cleavage specificities are known (29).

## A Catalog of Restriction Enzymes and Their Specificities

In Table I is a current list of known restriction endonuclease cleavage specificities grouped according to type of recognition sequence. The list is taken from Roberts (29) and includes for each sequence only the prototype enzyme name\*. In many cases, other enzymes recognizing the same sequence are known. These have been called isoschizomers by Roberts and are given in his complete listing (29); e.g. isoschizomers of *Hind* II are *Chu* II, *Hinc* II, and *Mnn* I. It is important to note that among a group of isoschizomers, cleavage position within the site may vary; e.g., *Sma* I

TABLE I. CATALOG OF RESTRICTION ENDONUCLEASE SEQUENCE SPECIFICITIES†

<i>Symmetric</i> ( <i>N</i> = 6)		<i>Degenerate symmetric</i> ( <i>N</i> = 6)		<i>Symmetric</i> ( <i>N</i> = 4)	
AvaIII	ATGCAT	AccI	GT (A/C)(G/T)AC	A1uI	AG CT
BaII	TGG CCA	AvaI	C PyCGPuG	FnuDII	CG CG
BamHI	G GATC*C	HaeI	(A/T)GG CC(T/A)	HaeIII	GG C*C
BcII	T GATCA	HaeII	PuGCGC Py	HhaI	GC*G C
BgIII	A GATCT	HgiAI	G(T/A)GC(T/A) C	HpaII	C C*GG
ClaI	ATCGAT	HindII	GTPy PuA*C	MboI	GATC
EcoRI	G AA*TTC			TaqI	T CGA
HindIII	A* AGCTT	<i>Symmetric</i> ( <i>N</i> = 5)		<i>Symmetric methylated</i> ( <i>N</i> = 4)	
HpaI	GTT AAC	AsuI	G GNCC		
KpnI	GGTAC C	AvaII	G G(A/T)CC	DpnI	GAmTC
MstI	TGCGCA	BbvI	GC(T/A)GC		
PstI	CTGCA G	EcoRII	CC*(A/T)GG	<i>Symmetric</i> ( <i>N</i> = 7)	
PvuI	CGATCG	HinfI	G ANTC	EcaI	GGTNACC
PvuII	CAG CTG	<i>Asymmetric</i> ( <i>N</i> = 4,5)			
SmaI	CCC GGG	MnII	CCTC cleavage 5 to 10 bases 3' to site		
SacI	GAGCT C	HgaI	GACGCNNNNN  (3')		
SacII	CCGC GG		CTGCTNNNNNNNNNN  (5')		
SaII	G TCGAC	HphI	GGTGANNNNNNNN  (3')		
XbaI	T CTAGA		CCACTNNNNNNNN  (5')		
XhoI	CTC GAG	MboII	GAAGANNNNNNNN  (3')		
			CTTCTNNNNNNNN  (5')		
			GATGC		
		StaNI	AGACC (3')		
		†EcoPI	TCTGG (5') cleavage 24 to 26 bases 3' to site		

t Compiled from data published by R. J. Roberts (29). Names of host organisms and references for enzymes and sequences are listed in his article.

† From Reiser, J., personal communication

Sequences are 5' → 3'. They should be visualized as duplexes although the sequence of only one strand is given. Vertical lines represent cleavage positions. Asterisks (\*) indicate bases modified by the corresponding modification enzymes. An "m" represents a methyl group. Only the prototype enzyme name is given for each sequence specificity; isoschizomers exist for many sequences and are given in the Roberts reference. Bases in parenthesis indicate that either base may occupy the position; e.g., *Acc*I recognizes GT(A/C)(G/T)AC which signifies the following sequence possibilities: GTAGAC, GTATAC, GTCGAC, GTCTAC. (The first and last sequences are the same in duplex form.) Pu is purine and Py is pyrimidine.

cleaves (5') C-C-C↓G-G-G while *Xma*I cleaves (5') C↑C-C-G-G-G.

Nucleotide sequences of recognition sites have in most cases been determined by analysis of oligonucleotides released from the 3' or 5' labeled termini of cleavage fragments in a manner analogous to that first used for *Hind*II (14). Recently a simple method for palindromic sequences has been devised. It depends on comparing digest fragments of  $\phi$ X174 and SV40 DNA produced by a given restriction enzyme with a table of possible fragments predicted by computer analysis of all tetra, penta, and hexanucleotide palindromes in these DNAs (31). Usually a unique sequence assignment is possible. For enzymes that cleave outside of their recognition site, identification can be made by analysis of the location of mutations that remove the site and by determining the position of bases, which when modified by a specific methylase or by certain chemical agents such as dimethylsulfate, inactivate the site (32).

Sites are classified according to whether they show P-fold rotational symmetry (palindromes) or are asymmetric. Among the symmetric sites are 20 perfect hexanucleotide sites, 6 degenerate hexanucleotide sites, 5 pentanucleotide sites with a central degeneracy, 7 perfect tetranucleotide sites, 1 hexanucleotide site with a central degeneracy, and 1 tetranucleotide site requiring a methylated adenine (see below). Each of the tetranucleotide sites can also be found as a central tetranucleotide in one or more of the hexanucleotide sites. The degenerate hexanucleotide sites, while losing strict structural symmetry at the degenerate position, retain a basic overall symmetry and are probably recognized as symmetrical by the enzymes (see below). When a complete degeneracy exists as in the *Hinf*I sequence, G-A-N-T-C, symmetry is not lost since no discriminating enzyme contacts are made with the degenerate base.

*Dpn*I is unique in that it recognizes the sequence, G-Am-T-C, only when it contains the methylated adenine. It is difficult to rationalize this reversal of the normal role of methylation since other *Diplococcus pneumoniae* strains carry the more conventional restriction enzyme, *Dpn*II, recognizing unmethylated G-A-T-C (33).

Among the 5 restriction enzymes recognizing asymmetric sites, 1 recognizes a tetranucleotide site and 4 recognize pentanucleotides. These enzymes cleave asymmetrically at a distance of 5 to 10 nucleotides 3' to the recognition sequence. *Hga*I deserves special comment since it generates

**\*Restriction endonucleases derive their names from an R-M system nomenclature (30) that uses an italicized three-letter abbreviation for the host organism followed by a fourth letter for strain where necessary and a roman numeral to indicate each R-M system in the organism. For example, *Hind*II is the name of the R-M system from which our original restriction endonuclease comes. Restriction enzymes are indicated as endonuclease R followed by the system name, and similarly, modification enzymes are designated methylase M followed by the system name; endonuclease R·*Hind*II and methylase M·*Hind*II. Most often a shorter form R·*Hind*II or M·*Hind*II is used, and when only restriction enzymes are being considered, they carry just the system name, i. e., *Hind*II, *Hind*III, *Eco*RI, etc.**

DNA cleavage fragments with cohesive termini that have a high probability of specific reunion with the original complementary partner; thus a small genome could be cut into several fragments that would religate only in original order (34).

Three main cleavage modes are observed by enzymes with symmetric recognition sites: even breaks (e.g. *HindII*), staggered breaks generating 3'-single-stranded cohesive termini (e.g. *PstI*), and staggered breaks generating 5'-single-stranded cohesive termini (e.g. *HindIII*). Each of these types of termini has found special uses in recombinant DNA work. So far, all the enzymes examined cleave so as to produce 3'-hydroxyl, 5'-phosphoryl termini.

#### *Mechanism of Nucleotide Sequence Recognition*

Restriction and modification enzymes, because of their variety and relative structural simplicity, provide excellent model systems for study of sequence-specific DNA-protein interactions. We have had an interest in this area for some time, and I should like to present, in a general way, our approach to this problem as well as possible directions for future research in the area.

The majority of R-M system recognition sites possess 2-fold rotational symmetry. Two basic recognition mechanisms are possible for these sites: *symmetric* recognition involving bilateral symmetric contacts in a duplex site and *asymmetric* recognition involving a set of nonsymmetric contacts (Fig. 1). For single-stranded sites, only the asymmetric mechanism can apply. An important consideration then, is whether or not restriction enzymes or their corresponding modification methylases can act on single-stranded sites.

Most restriction endonucleases appear to require duplex sites, as originally demonstrated for *HindII* (14). A few enzymes, for example, *HaeIII*, *HhaI*, *SfaI*, *MboI*, and *HinfI* act slowly on single-stranded DNAs (35, 36, 37), but this is now thought to be due to formation of transient duplexes (38). It is probable that most of the restriction enzymes employ a symmetric recognition mechanism. This is based on several arguments. First, since hemimethylated sites are generated during replication, the recognition process must be responsive to methylation on either strand. This is most easily achieved by bilateral, symmetric protein-DNA contacts at the methylation positions within the duplex site. Second, from the standpoint of genetic economy it is less expensive to specify a protein monomer site recognizing  $n/2$  bases than one recognizing  $n$  bases (Fig. 1). Finally, the *EcoRI* endonuclease exists as dimers and tetramers of a single 28,500 dalton subunit, and under physiological conditions, cleaves both strands of a duplex site in one binding event (18). It seems likely from symmetry considerations that such a dimeric or tetrameric structure will prove to be the rule for other enzymes.

Modification methylases may recognize sites in a fashion quite different from the restriction enzymes. Some of these enzymes appear capable of

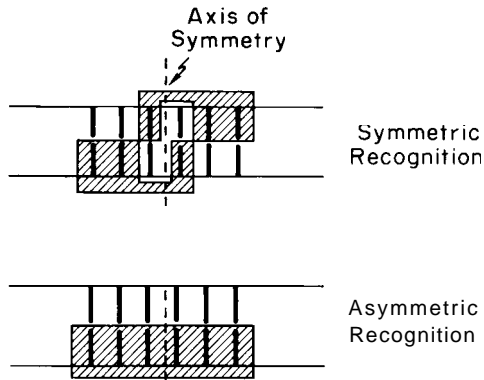


Fig. 1: Two different ways in which restriction and modification enzymes may interact with two-fold rotationally symmetrical nucleotide sequences in order to achieve recognition. In the symmetric recognition model, the protein possesses subunits arranged in two-fold rotational symmetry and each identical half interacts with a minimum of  $n/2$  nucleotides. In the asymmetric model, the protein is assumed to have an asymmetric structure and must interact with at least one nucleotide at each base pair position for a total of  $n$  nucleotides. The simplest case, where all the interactions are on one strand, is shown.

acting on purely single-stranded sites, implying an asymmetric recognition process (Fig.1). Michael B. Mann, in my laboratory (39), has shown that *M·HhaI* methylates C residues in the random copolymer, poly(dN-acetyl G, dC) which is unable to form any Watson-Crick base pairing (based on absence of a thermal melting transition), and that *M·HhaI* and *M·HpaII* methylate poly(dX, dC) which also shows no thermal melting transition. *M·HaeIII* and *M·HpaII* also methylate denatured salmon sperm DNA to the same total extent as native DNA, although at half the rate. A lower rate and extent (30% ) was achieved with *M·HhaI*. These observations support the notion that these methylases can act on single-stranded sites with preservation of specificity. This implies that discriminatory interactions need involve only the bases on one strand. Rubin and Modrich (18) have shown that the *EcoRI* methylase is a functional monomer of molecular weight 39,000 that transfers methyl groups to each strand of the *EcoRI* site in individual catalytic events that are interrupted by dissociation from the site. On theoretical grounds, an asymmetric recognition mechanism is reasonable for the modification methylases because, as pointed out by the above authors, the usual *in vivo* hemimethylated duplex substrate is inherently asymmetric.

Turning again to the restriction endonucleases, it was early proposed that recognition of a symmetric site might depend on some unusual structure of the site. Kelly and I (14) initially suggested the enzymes might interact with open (melted) sites (Fig. 2) because at that time we felt there was insufficient opportunity for base specific interactions in the helical grooves. Meselson *et al.* (41) proposed that symmetric sites might transiently form cruciform structures with special features that would promote enzyme recognition (Fig. 2). Both open and cruciform struc-

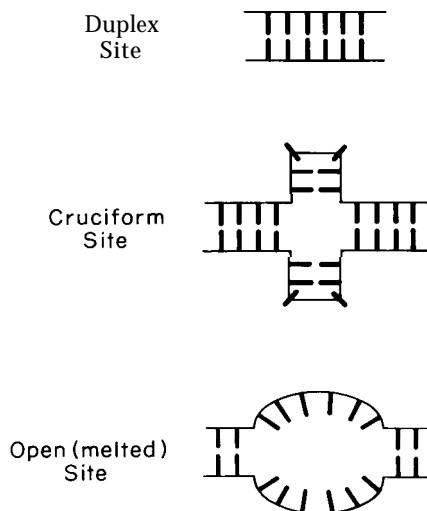


Fig. 2: Representations of three possible structures of a recognition site that might facilitate specific protein-DNA interaction.

tures are energetically unfavorable, and there are no compelling theoretical reasons to favor them. We now accept, as a result of the *lac* repressor-operator studies (42, 43) and recent structural analyses of base pairs (40), that the base groups exposed in the major and minor grooves of helical DNA are sufficient for discriminatory interactions. Therefore, to simplify the discussion, it will be assumed that sites are recognized while in the helical configuration.

The restriction enzyme *Hha*I was chosen by Michael Mann and myself for initial studies because the site, (5')pG-C-GJC is particularly simple and can easily be synthesized in alternating polymer form. We have chosen chemical modification of the bases as an approach to determining those groups in the major or minor grooves that play a role in recognition. Effects on catalytic activity rather than binding are most easily measured and have been used in our studies, although we acknowledge that each may provide somewhat different information.

We have depended heavily on the analysis of Seeman *et al.* (40) for interpretation of our results. These authors compare the various potential sites for discriminatory protein-DNA contacts in the major and minor grooves of the different base pairs. A G·C base pair, the only kind in the *Hha*I site, is shown in Fig. 3. The major and minor grooves may be visualized as divided into outer and central regions. In the central major groove, the O6 atom of guanine is hydrogen bonded to the amino N4 of cytosine. The outer major groove contains the N7 atom of guanine and C5 hydrogen atom of cytosine. The central minor groove contains the 2-amino (N2) group of guanine. The outer minor groove contains the N3 atom of guanine and the O2 atom of cytosine. (The latter is hydrogen bonded to the 2-amino group of guanine). A top view of the major groove

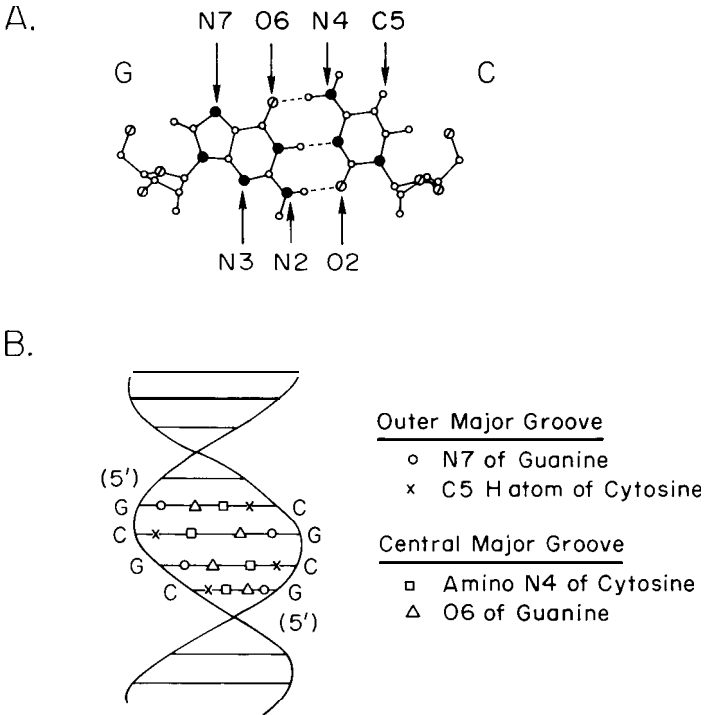


Fig. 3: Potential major and minor groove discriminatory protein-DNA interaction positions in the *HhaI* restriction endonuclease recognition site. A. A stereochemical drawing of a G-C base pair in DNA (adapted from Seeman *et al.* (40)). Potential atoms for interaction are indicated. B. A rough sketch of a section of helical DNA containing an *HhaI* site. Base pairs are indicated by horizontal bars. The view is from above the major groove, and approximate positions of interacting atoms are shown.

in the *HhaI* site is shown in Fig. 3 with potential atoms for interaction diagrammatically represented.

The *HhaI* modification methylase transfers methyl groups from S-adenosylmethionine onto the 5-position of the internal cytosines (situated between the two guanines) in the *HhaI* site and protects against cleavage by the *HhaI* endonuclease (39). Since the methyl groups probably interfere sterically, we infer that contacts between protein and DNA must take place at these two outer major groove positions of the duplex site. Mann and I have also shown that methylation introduced on the 5-position of the external cytosines inhibits cleavages by R-*HhaI*. Therefore, it is likely that these are also closely fitted by the enzyme. In another experiment, we used dimethylsulfate to introduce methyl groups on the N7 positions of guanine in an *HhaI*-site positioned 20 bases from the 5' terminus of a  $\Phi$ X174RF DNA fragment of known sequence. After this treatment, the fragment was digested with *HhaI* endonuclease, and fractionated by gel electrophoresis into cleaved and uncleaved molecules. These were then treated by Maxam and Gilbert (44) sequencing methods so as to cleave at the methylated position and analyzed by electrophoretic gels. Bands representing G's in

the site were greatly increased in intensity in the gel tract representing uncleaved molecules and absent from the gel tract representing cleaved molecules. We concluded that methylation of the N7 position of any G residue in the site conferred protection against cleavage. Again, the effect is likely to be steric, and we infer that the enzyme closely fits these positions in the outer major groove of the site (Fig. 2B). In summary, we have been able to demonstrate that methylation at any of eight positions in the outer portions of the major groove inhibits cleavage. Study of the central major groove groups by this approach is more difficult because modifications often destroy helical pairing.

To examine minor groove interactions, we looked at activity on alternating poly(dI-dC). Inosine contains a H-atom in place of the 2-amino group of guanine in the central minor groove. A dI-dC base pair mimics a dA-dT base pair when viewed from the minor groove. R-*HhaI* cleaves alternating poly(dI-dC) efficiently, thus the 2-amino group is not essential for discrimination, and more specifically, plays no role in discrimination of A·T from G·C base pairs. The central minor groove thus seems not to be occupied. Interaction is still possible in the outer positions of the minor groove where the N3 atom of purines and the O2 atom of pyrimidines are exposed. However, since these are both electron-rich hydrogen bond acceptors and occupy strictly similar positions regardless of base pair type, they are not considered likely for discriminatory interactions (40). We have tentatively concluded then that *HhaI* endonuclease occupies the major groove and derives all discriminatory contacts from groups in the central and outer major groove positions.

It is likely that *HhaI* and other restriction endonucleases also interact with the sugar-phosphate backbone within a site. We suggest that stabilizing interactions of this sort, and also non-discriminatory outer minor groove interactions, may extend to adjacent nucleotides to either side of a site. These interactions could explain two observations. First is the size effect. Greene et al. (45) found that the affinity of *EcoRI* endonuclease for the symmetric octanucleotide (5') pT-G-A-A-T-T-C-A, containing a central *EcoRI* recognition site is 200 times less than for the *EcoRI* site in SV40 DNA. We found similarly that a symmetric decanucleotide containing terminal *HpaII* sites is not detectably cleaved by *HpaII* endonuclease, but addition of nucleotides to the end restores the site (46). Second is the finding that some sites are cut preferentially, depending on external sequence context (47, 48), suggesting that weak contacts are made at neighboring nucleotides outside of the site.

#### Recognition of Degenerate Sites

Degenerate sites are not strictly symmetrical by structural criteria. *AccI* endonuclease recognizes the site (5')G-T-(A/C)-(G/T)-A-C which exists as four combinations of the degenerate nucleotides: G-T-A-G-A-C, G-T-A-T-A-C, G-T-C-G-A-C, and G-T-C-T-A-C. The first and last combinations are asymmetric. Yet it is very appealing to think of the enzyme as interacting

with each of these sequences in a similar way so as to preserve symmetry. The discrimination rules of Seeman et al. (40) allow for this possibility. They describe several potential positions for major and minor groove interaction with each of the Watson-Crick base pairs. A protein-DNA interaction, e.g. a single hydrogen-bond directed to one of the positions, is insufficient to allow discrimination between all the base pairs, although two interactions can be sufficient. Thus, a restriction or modification enzyme making only a single contact at symmetrically placed base pairs could allow a degeneracy, i.e., an ambiguity in recognition.

According to this scheme, four types of degeneracy appear to be possible: (A/G)-(T/C) (Pu-Py type), (A/C)-(G/T), (A/T)-(T/A), and (G/C)-(C/G). The Pu-Py type degeneracy could arise from outer major groove contact directed toward the purine N7 atom. The (A/C)-(G/T) degeneracy could result from a single interaction directed to the central major groove amino N4 of cytosine and amino N6 of adenine position, or to the carbonyl oxygen of thymine or guanine, since these pairs of groups occupy similar positions in the specified degeneracy. The (A/T)-(T/A) degeneracy could result from a central minor groove interaction since the C2 hydrogen atom of adenine occupies a sterically similar position in each A·T orientation. Finally, the (G/C)-(C/G) degeneracy could result from interaction in the central minor groove with the 2-amino group of guanine which is in a sterically similar position for each orientation of the G·C pair. Inspection of the six degenerate hexanucleotide sequences in Table 1 reveals that, of the four predicted degeneracies, only (G/C)-(C/G) has not yet been found.

Among the symmetrical pentanucleotide sites, two are completely degenerate at the middle nucleotide position implying either absence of protein-DNA interaction or possibly non-discriminatory outer minor groove contacts. Three of the pentanucleotide sites contain a middle (A/T) nucleotide degeneracy. This is compatible with a single interaction directed to the C2 hydrogen atom of adenine, which is inherently symmetrical since it falls almost directly on the dyad axis of the site. Other types of degeneracy in the middle nucleotide position of pentanucleotide sites appear less likely.

The general agreement between the above predicted and observed degeneracies further reinforces the notion that restriction enzymes accomplish nucleotide sequence recognition through major and minor groove interactions.

#### *Relaxation of Sequence Specificity*

Several DNA enzymes, e.g., terminal deoxynucleotidyl transferase (49) and pancreatic DNase (50), show changes in specificity according to species of divalent cation and ionic conditions in the reaction mixture. Some restriction endonucleases appear to be similarly affected. *EcoRI* endonuclease cleaves the sequence (5')G-A-A-T-T-C in a reaction mixture containing 100m M Tris-Cl (pH 7.3), 50 mM NaCl, and 5 mM MgCl<sub>2</sub>. When the conditions are changed to 25 mM Tris-Cl (pH 8.5), 2 mM MgCl<sub>2</sub>, the

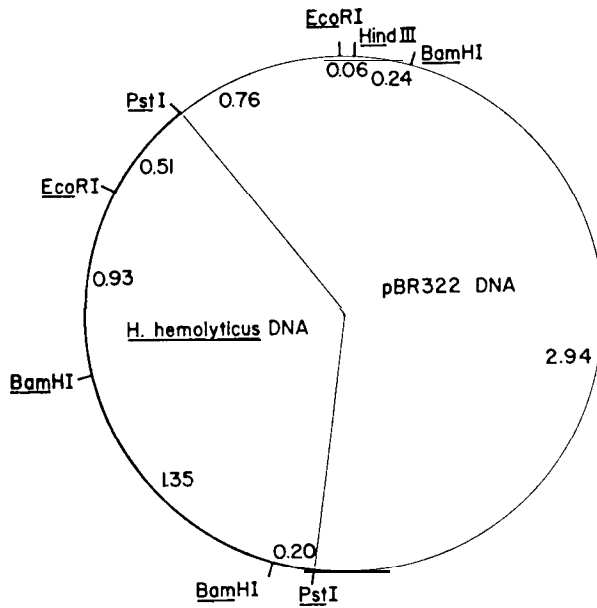


Fig. 4: A restriction enzyme cleavage map of pDI10 DNA. Distances are in kilobases

specificity is lowered to the central tetranucleotide sequence (5')A-A-T-T (51). However, the enzyme retains a strong preference for the canonical site; extensive digestion is required to achieve cleavage at the new tetranucleotide sites, and there is great variability among them in regard to cleavage rate. The latter presumably reflects the degree of relatedness to the canonical site.

Hsu and Berg (52) obtained decreased specificity with *EcoRI* by substituting  $Mn^{2+}$  for  $Mg^{2+}$ . They also noted relaxed specificity with *HindIII*, but not *HpaII*, in the presence of  $Mn^{2+}$ . This appears to be a promising area for more investigation.

#### Cloning R-M System Genes

Detailed studies of restriction and modification enzymes require quantities of pure enzyme. However, enzymes are often obtained in poor yield from source bacteria. Because of this we have begun to explore the possibilities of cloning various R-M system genes as a means to achieve enzyme overproduction. Using this approach, Michael Mann and Nagaraja Rao, in my laboratory, have recently cloned the *HhaII* system from *Haemophilus haemolyticus* in the *E. coli*-pBR322 host-vector system using a "shotgun" approach (53). Total chromosomal DNA was cleaved with *PstI* endonuclease and inserted into the *PstI* cloning site of the plasmid, located in the ampicillin resistance gene, by means of a GC-extension procedure developed by Rougeon *et al* (54). After transfection into an  $r^{-}m^{-}$  *E. coli* host (HB 101), tetracycline-resistant recombinant clones were tested for acquisition of a new restriction phenotype using phage  $\lambda$ . A single such clone was found

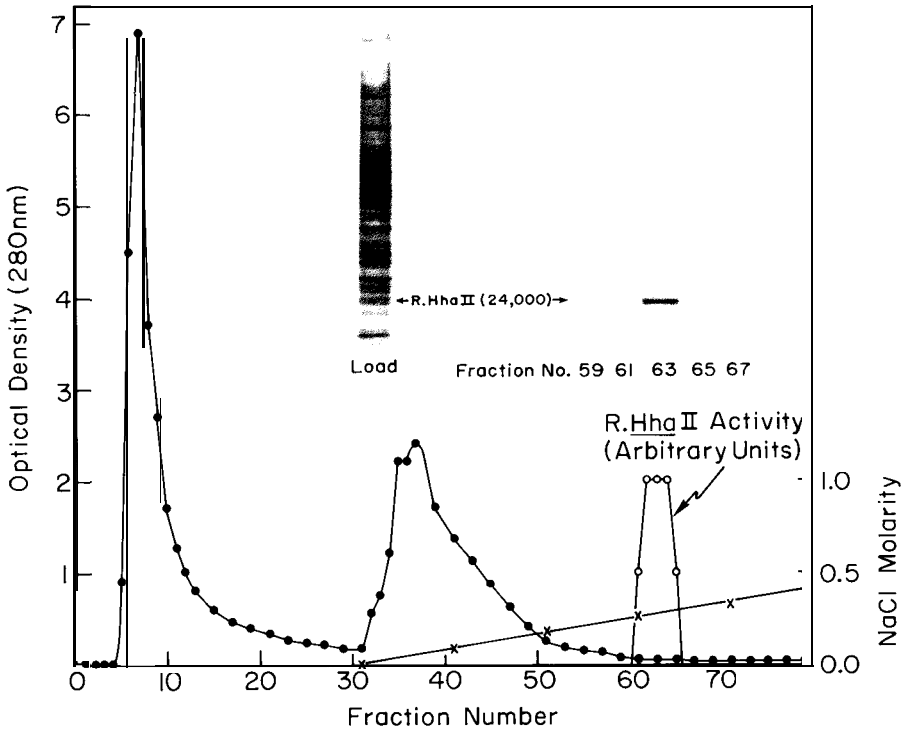


Fig. 5: Single-stranded DNA-agarose affinity chromatography of *HhaII* endonuclease. A crude **extract** from about 10 gm of thermally induced KJ34 cells was processed to remove nucleic acids and then chromatographed on a single-stranded DNA agarose column (2.5X20 cm) using a 1,000 ml gradient essentially as described by Mann et al. (53). Fractions (approximately 12 ml) were collected and assayed for protein by absorbance at 280 nm and for *R.HhaII* activity by gel electrophoresis of lambda DNA digestion products. The endonuclease units represent estimates of the percent of complete digestion. Protein species in load and in peak endonuclease fractions were examined by SDS-polyacrylamide gel electrophoresis.

among 1400 tested. The recombinant plasmid, pDI10, recovered from this clone contained a 3.0 kb DNA insert flanked by *PstI* sites. A cleavage map is shown in Fig. 4. The HB101 clone carrying pDI10 exhibits classical restriction and modification behavior with phage  $\lambda$  (e.o.p.  $\sim 10^7$ ) and several other phages.

The pDI10 DNA efficiently re-transfects new HB101 cells suggesting that methylation is expressed well in advance of restriction. To account for this, we have suggested that the methylase might act as a positive regulator for expression of the restriction gene. By this scheme, methylase would initially be occupied with methylation of host chromosomal sites, becoming free to induce restriction enzyme only after its job was complete. This is only one possible scheme to explain the apparent sequential action of these genes. Study of the regulation should prove interesting.

To increase plasmid copy number and consequent enzyme overproduction, the 3.0 kb DNA fragment was excised from pDI10 DNA and transferred into a second plasmid vector, pKC16, a hybrid of pBR322 and

phage  $\lambda$  containing a thermally-inducible h-replication region (55), to yield a new hybrid plasmid pDI21. Using a clone containing pDI21, a 20 minute 42°C treatment raises plasmid copy number and enzyme yield several fold over that obtainable with pDI10. The restriction endonuclease and modification methylase were purified from crude extracts of this clone by single-stranded DNA-agarose affinity chromatography. Typical results for the endonuclease are shown in Fig. 5. This essentially one-step procedure yielded an active fraction showing a single major protein band of about 24,000 daltons by SDS-polyacrylamide gel electrophoresis (Fig. 5). The purified endonuclease gives a DNA cleavage pattern on  $\Phi$ X 174RF DNA identical to *Hinf*I with the sequence specificity (5')pGA-N-T-C. The DNA methylase from the clone protects against both *Hha*II and *Hinf*I cleavage.

We believe that future developments in the field of restriction and modification enzymes will depend heavily on gene cloning, both for enzyme overproduction and for genetic studies. There is a great advantage to having the *Hha*II genes on a small segment of DNA that can be propagated and expressed in *E. coli*. The genes are easily accessible for genetic studies in the new host, whereas this would be difficult or impossible in the original *Haemophilus* strain. The DNA segment is small enough to be readily sequenced, thus providing direct information on gene arrangement, regulatory sequences, and protein amino acid sequences. The latter will be valuable for future crystallographic studies of enzyme structure, a goal which must be achieved if we are to fully understand the nature of the protein-DNA interactions involved in nucleotide sequence recognition.

**REFERENCES**

1. Nathans, D. and Smith, H.O. (1975) *Ann. Rev. Biochem.* 44, 273-293.
- la. Luria, S.E. and Human, S.L. (1952) *J. Bacterial.* 64, 557-569.
2. Bertani, G. and Weigle, J.J. (1953) *J. Bacterial.* 65, 113-121.
3. Arber, W. and Dussoix, D. (1962) *J. Mol. Biol.* 5, 18-36.
4. Dussoix, D. and Arber, W. (1962) *J. Mol. Biol.* 5, 37-49.
5. Arber, W. (1965) *Ann. Rev. Microbiol.* 19, 365-378.
6. Linn, S. and Arber, W. (1968) *Proc. Nat. Acad. Sci. USA* 39, 1300-1306.
7. Meselson, M. and Yuan, R. (1968) *Nature* 217, 1110 - 1114.
8. Boyer, H.W. (1971) *Ann. Rev. Microbiol.* 25, 153- 176.
9. Horiuchi, K. and Zinder, N.D. (1972) *Proc. Nat. Acad. Sci. USA* 69, 3220-3224.
10. Adler, S.P. and Nathans, D. (1973) *Biochem. Biophys. Acta* 299, 177- 188.
11. Murray, N.E., Batten, P.L., and Murray, K. (1973) *J. Mol. Biol.* 81, 395-407.
12. Smith, H.O. and Wilcox, K.W. (1970) *J. Mol. Biol.* 51, 379-391.
13. Weiss, B. and Richardson, C.C. (1967) *J. Mol. Biol.* 23, 405-417.
14. Kelly, T.J. Jr. and Smith, H.O. (1970) *J. Mol. Biol.* 51, 393-409.
15. Modrich, P. and Zabel, D. (1976) *J. Biol. Chem.* 251, 5866-5874.
16. Greene, P.J., Betlach, M.C., Goodman, H.M., and Boyer, H.W. (1974) *Methods Mol. Biol.* 7, 87-111.
17. Smith, L.A. and Chirlkjian, J.C., personal communication.
18. Rubin, R.A. and Modrich, P. (1977) *J. Biol. Chem.* 252, 7265-7272.
19. Roy, P. H. and Smith, H. O. (1973) *J. Mol. Biol.* 81, 427-444.
20. Roy, P.H. and Smith, H.O. (1973) *J. Mol. Biol.* 81, 445-459.
21. Vanyushin, B.F., Belozersky, A.N., Kokurina, N.A., and Kadirova, D.X. (1968) *Nature* 218, 1066-1067.
22. Arber, W. and Linn, S. (1969) *Ann. Rev. Biochem.* 38, 467-500.
23. Old, R., Murray, K., and Roizes, G. (1975) *J. Mol. Biol.* 92, 331-339.
24. Danna, K. and Nathans, D. (1971) *Proc. Nat. Acad. Sci. USA* 68, 2913-2917.
25. Sharp, P.A., Sugden, B. and Sambrook, J. (1973) *Biochemistry* 12, 3055-3063.
26. Yoshimori, R. (1971) Ph.D. thesis, Univ. of Calif., San Francisco.
27. Middleton, J.H., Edgell, M.H., and Hutchison, C.A.111. (1972) *J. Virol.* 10, 42-50.
28. Gromkova, R. and Goodgal, S.H. (1972) *J. Bacterial.* 109, 987-992.
29. Roberts, R.J. (1978) *Gene*, 4, 183- 193.
30. Smith, H.O. and Nathans, D. (1973) *J. Mol. Biol.* 81, 419-423.
31. Fuchs, C., Rosenvold, E.C., Honigman, A., and Szybalski, W. (1978) *Gene* 4, 1-23.
32. Kleid, D., Humayun, Z., Jeffrey, A., and Ptashne, M. (1976) *Proc. Nat. Acad. Sci. USA* 73, 293-297.
33. Lacks, S. and Greenberg, B. (1975) *J. Biol. Chem.* 250, 4060-4066.
34. Brown, N.L. and Smith, M. (1977) *Proc. Nat. Acad. Sci. USA* 74, 3213-3216.
35. Blakesley, R.W. and Wells, R.D. (1975) *Nature* 257, 421-422.
36. Horiuchi, K. and Zinder, N.D. (1975) *Proc. Nat. Acad. Sci. USA* 72, 2555-2558.
37. Godson, G.N. and Roberts, R.J. (1976) *Virology* 73, 561-567.
38. Blakesley, R.W., Dodson, J.B., Nes, I.F., and Wells, R.D. (1977) *J. Biol. Chem.* 252, 7300-7306.
39. Mann, M.B. and Smith, H.O. (1978) in *Transmethylation*, eds. E. Usdin, R. T. Borchartdt, C. R. Creveling (Elsevier/North Holland) pp 483-493.
40. Seeman, N.C., Rosenberg, J.M., and Rich, A. (1976) *Proc. Nat. Acad. Sci. USA* 73, 804-808.
41. Meselson, M., Yuan, R., and Heywood, J. (1972) *Ann. Rev. Biochem.* 41, 447-466.
42. Adler, K., Beyreuther, K., Fanning, E., Geisler, N., Gronenborn, B., Klemm, A., Müller-Hill, B., Pfahl, M., and Schuitz, A. (1972) *Nature* 237, 322-327.
43. Gilbert, W., Maxam, A., and Mirzabekov, A. (1976) in *Control of Ribosome Synthesis*, Alfred Benson Symposium IX, eds. Kjeldgaard, N.O., Maaloe, O. (Munksgaard, Copenhagen) pp. 139- 148.

44. Maxam, A. and Gilbert, W. (1977) *Proc. Nat. Acad. Sci. USA* 74, 560-564.
45. Greene, P.J., Poonian, M.S., Mussbaum, A.L., Tobias, L., Garfin, D.E., Boyer, H.W., and Goodman, H.M. (1975) *J. Mol. Biol.* 99, 237-261.
46. Mann, M.B. and Smith, H.O. (1977) *Nucl. Acids Res.* 4, 4211-4221.
47. Thomas, M. and Davis, R.W. (1975) *J. Mol. Biol.* 91, 315-328.
48. Smith, H.O. and Birnstiel, M. (1976) *Nucl. Acids Res.* 3, 2387-2398.
49. Roychondhury, R., Jay, E., and Wu, R. (1976) *Nucl. Acids Res.* 3, 863-877.
50. Melgar, E. and Goldthwait, D.A. (1968) *J. Biol. Chem.* 243, 4409-4416.
51. Polisky, B., Greene, P., Garfin, D.E., McCarthy, B.J., Goodman, H.M., and Boyer, H.W. (1975) *Proc. Nat. Acad. Sci. USA* 72, 3310-3314.
52. Hsu, M. and Berg, P. (1978) *Biochemistry* 17, 131-138.
53. Mann, M.B., Rao, N.R., and Smith, H.O. (1978) *Gene* 3, 97-112.
54. Rougeon, F., Kourilsky, P., and Mach, B. (1975) *Nucl. Acids Res.* 2, 2365-2378.
55. Rao, N.R. and Rogers, S.G. (1978) *Gene* 3, 247-263.