Volume 14 Number 13 1986 Nucleic Acids Research

Isolation and computer-aided characterization of *MmeI*, a Type II restriction endonuclease from *Methylophilus methylotrophus*

A.C.Boyd*, I.G.Charles, J.W.Keyte and W.J.Brammar

Department of Biochemistry, University of Leicester, Leicester LE1 7RH, UK

Received 19 March 1986; Revised and Accepted 2 June 1986

ABSTRACT

A Type II restriction endonuclease, MmeI, has been purified from the obligate methylotroph, Methylophilus methylotrophus. The enzyme was shown to have the non-palindromic recognition sequence

$$\begin{array}{l} 5' - T\ C\ C\ Pu\ A\ C\ (N)_{20} - 3' \\ 3' - A\ G\ G\ Py\ T\ G\ (N)_{18} - 5' \end{array}$$

and to cleave (as indicated) on the 3' side, generating a two nucleotide 3' projection.

Determination of the recognition sequence was achieved using two new computer programs; RECOG, which predicts recognition sequences from the pattern of restriction fragments obtained from DNAs of known sequence, and GELSIM, which generates graphical simulations of DNA band patterns obtained by gel electrophoresis of restriction digests of sequenced DNA molecules.

INTRODUCTION

Type II restriction endonucleases that cut DNA molecules at, or near, defined nucleotide sequences, have now been obtained from a wide variety of bacterial species (1) and have become essential tools in molecular biology. The sequence-specificity of an endonuclease can often be determined by analysis of the ends generated by the action of the enzyme. It is also possible to define a target sequence by analysis of the pattern of cleavage of DNA molecules of known nucleotide sequence (2,3,4,5).

Here we describe the isolation of a new restriction enzyme, MmeI, from the Gram-negative bacterium Methylophilus methylotrophus, an obligate methylotroph and source of commercial single cell protein (SCP). A computer program, RECOG, which includes an algorithm not used in similar programs (2, 3,4,5), was developed to determine the recognition sequence of the enzyme using accurate mapping data on a sequenced DNA, pBR322 (6). (The pBR322 sequence revision, adding one base pair to the TcR gene (7), has a negligible effect on the analyses described below, so the original coordinates are used throughout.) MmeI restriction generates a complex

mixture of partial and complete digest products; therefore, another computer procedure, GELSIM, was written to simulate graphically the migration of DNA fragments during agarose gel electrophoresis, simplifying analysis of MmeI gel patterns.

Computer methods described below are generally applicable to the problem of characterizing newly isolated restriction enzymes, especially preparations too crude to be analysed biochemically.

MATERIALS AND METHODS

Methylotroph culture conditions

M.methylotrophus is an obligate methylotroph used commercially as a source of single cell protein (8). Cells were grown aerobically at 37°C in medium containing 0.9% (v/v) methanol, and (per litre), NaH_2PO_4 (1.4g), K_2HPO_4 (1.9g), $(NH_4)_2SO_4$ (1.8g), $MgSO_4$ (0.2g) with trace elements $CuSO_4.5H_2O$ (0.02mg), $MnSO_4.4H_2O$ (0.1mg), $FeCl_3$ (0.98mg), $ZnSO_4.7H_2O$ (0.1mg) and $CaCO_3$ (1.8mg).

Substrate DNAs and enzymes

Plasmid pBR322 and M13mp7 RF DNAs were prepared by a scaled-up alkaline/SDS lysis method followed by purification on a CsCl/ethidium bromide gradient (9, 10). The E.coli strain GM48 (11) (dam-3 dcm-6 gal ara lac thr leu thi tonA tsx) was transformed with pBR322 and was used as a source of "dam⁻" plasmid DNA. Phage Lambda (cI857, Sam7) DNA was prepared from a suitable lysogen by a standard technique (12). SV40, PhiX174 RF and M13 RF DNAs were gifts from (respectively) B.K. Ely, G.E. Blair and J.G.G. Schoenmakers. Restriction enzymes PstI, Sau3A, TaqI, EcoRI and BamHI were from BRL Inc. and used as recommended.

Symbols for degenerate nucleotides

The programs described here were developed before any standard symbols for degenerate nucleotide positions were accepted. Therefore, as well as the commonly used symbols X = purine, Y = pyrimidine, N = A, C, G or T, new ones were devised mnemonically to describe degeneracies found in many recognition sequences. These are: 1 = A or C ("ACe"); 8 = A or T ("ATe") and their "arithmetic complements" 9 = G or T; 2 = G or C (1 + 9 = 2 + 8 = 10). Under this system, for example, AccI recognizes GT19AC and HaeI recognizes 8GGCC8 (1).

Computer programs

Programs were written in CDC Extended Basic 3.0 and run on a CDC Cyber 73 computer.

Purification of MmeI

The purification was based on a general method described previously (13). In a typical experiment, cells (approximately 10g wet weight) were resuspended in extract buffer (EB: 10mM K_2HPO_4 - KH_2PO_4 pH 7.0, 7mM 2-mercaptoethanol, 1mM EDTA, 10% (w/v) glycerol) and disrupted in a French press. The debris was removed by a high speed spin (18,000 rpm; MSE 6 x 250ml rotor) and the supernatant adjusted to 0.1M NaCl. The lysate was applied directly to a phosphocellulose (Whatman P11) column, and 100 1ml fractions collected after elution with a 0.1 - 1.0M gradient of NaCl in EB. MmeI activity eluted between 0.2 and 0.25M salt. Active fractions were dialysed against storage buffer (SB: EB containing 50% (w/v) glycerol) and stored at -20°C. Attempts to refine the crude extract preparation prior to column loading, or to add another column chromatography step, were unsuccessful in increasing the yield or purity of the enzyme. Nevertheless, material of sufficient purity and quantity to determine the recognition sequence of the enzyme was obtained.

The existence of another restriction enzyme produced by M.methylotrophus was discovered fortuitously when pBR322 DNA from a dam$^-$ strain (GM48) was used instead of Lambda DNA in the assay. The methylated adenine residues within GATC sequences found in DNA from normal (dam$^+$) strains of E.coli are absent from this DNA (11). Using this substrate, an endonuclease different from MmeI was seen to elute at approximately the same salt concentration, obscuring MmeI-specific bands in the assay, thus suggesting a higher relative activity. The new enzyme, MmeII ((14); manuscript in preparation), recognizes the same sequence as the Dam methylase (15), and shares some properties with the previously described restriction enzyme, MboI (1).

Enzyme assay

1 - 2 µl of column fractions were incubated in a total volume of 20 µl assay buffer (666: 6mM Tris.HCl pH 7.5, 6mM $MgCl_2$, 6mM 2-mercaptoethanol) for 1-2 hours with 0.5 µg of Lambda DNA at 37°C. After incubation, 5 - 10 µl loading buffer (10mM Tris.HCl pH 7.5, 20mM EDTA, 10% glycerol, 2mg/ml agarose beads, 0.01% bromophenol blue) was added, and the products resolved by electrophoresis on horizontal 1% (w/v) agarose gels in Tris-acetate buffer (20mM Tris-acetate pH 8.2, 10mM sodium acetate, 0.5mM EDTA and 0.5 µg/ml ethidium bromide). Experimentation with assay conditions for the partially purified enzyme preparations revealed that both MmeI and MmeII were inhibited by NaCl concentrations greater than 50mM, 666 buffer giving optimal activity. The addition of ATP or S-adenosyl methionine did not affect

either enzyme activity, showing that they are Type II restriction enzymes.

Determination of site of cleavage by MmeI

The position of cleavage of substrate DNA by MmeI was determined using double-stranded DNAs generated by priming DNA synthesis on M13mp9 template DNA with synthetic oligodeoxynucleotides. The latter were synthesized by the method of Matthes et al. (16), as modified by Sproat and Gait (17), and purified by electrophoresis on a 20% polyacrylamide gel. The oligonucleotides were end-labelled using T4 polynucleotide kinase (Pharmacia) and γ-^{32}P-ATP (Amersham : 3000 Ci/mmol) in 50 µl kinase buffer I (18). The reaction was incubated for 30 min at 37°C and stopped by heating at 65°C for 10 min. A 2 µl sample of the reaction mixture, containing approximately 8 ng of end-labelled oligonucleotide, was annealed with 0.5 µg single-stranded M13mp9 DNA in a total of 10 µl of 10 mM Tris.HCl, pH 7.5, 10mM $MgCl_2$, for 10 min at 65°C, before cooling to room temperature for 10 min. The volume was increased to 20 µl by adding 9 µl of a solution containing all four deoxynucleoside triphosphates to give final concentrations of 25mM for each, plus 1 µl (5 units) of Klenow fragment DNA polymerase (Pharmacia). The polymerase reaction was incubated at room temperature for 20 min and terminated by heating at 65°C for 10 min. A sample of 8 µl of the polymerase reaction mixture was added to 1 µl 10 x 666 buffer and 1 µl MmeI (ca 0.5 units), incubated at 37°C for 30 min, then inactivated at 65°C for 10 min prior to loading onto a 6% polyacrylamide sequencing gel. The radiolabelled DNA bands produced by digestion with MmeI were analysed on a buffer-gradient polyacrylamide gel (19) alongside a conventional dideoxynucleotide sequencing ladder obtained using M13mp9 template DNA and the relevant synthetic oligonucleotide as primer.

RESULTS

Mapping of MmeI sites on pBR322

At first, mapping of MmeI sites on pBR322 (4362bp long; (6)) was complicated by the appearance of partially digested fragments, which persisted even after prolonged or repeated incubation with the enzyme. At least eight bands were visible in a MmeI digest of pBR332, all >1500bp long; most of these must be partial digest products (Fig. 1a). The assumption which led to the successful mapping of MmeI sites was that the enzyme produced all possible complete and partial products in a given digest. Using this assumption, it is clear that pBR322 must have at least four MmeI sites, since three sites on a circular molecule can only generate a maximum of seven different-sized partial and complete digest products.

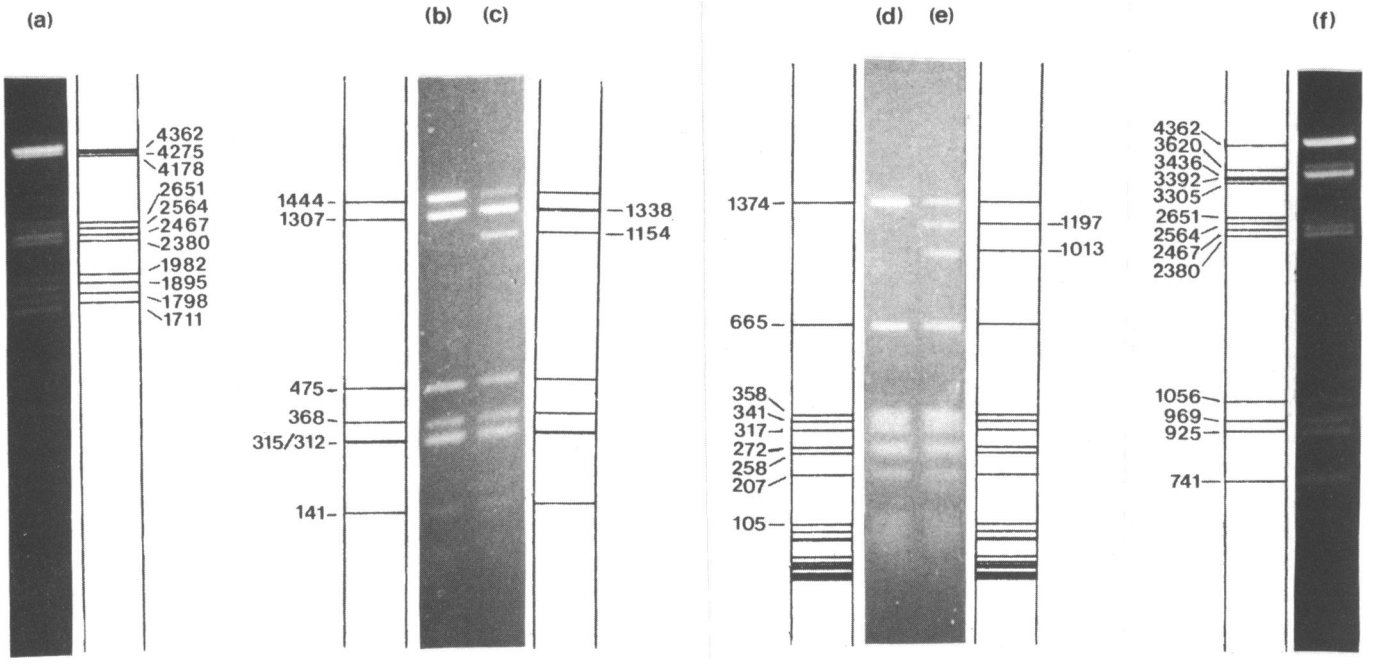

Figure 1. Restriction digests of pBR322: (a) MmeI, (b) TaqI, (c) TaqI/MmeI, (d) Sau3A, (e) Sau3A/MmeI and (f) PstI/MmeI. Each gel photograph is flanked by a computer-generated simulation with fragments sizes (in bps) predicted from the plasmid sequence.

Many double digests were used to complete the map, but only some are discussed here. In every case, fragment sizes were estimated using a least-squares fit to the reciprocal relationship between DNA fragment size and mobility on agarose gels described previously (20,21). The relationship can be represented by a simple size/mobility equation of the form:

$$\text{Size} = Q2 + (Q1/(\text{Mobility} - Q3)) \qquad (I)$$

where Q1, Q2 and Q3 are gel-dependent parameters. This equation was also exploited in the GELSIM procedure described below. A TaqI/MmeI double digest (Fig. 1c) showed that the 1444 bp TaqI "A" fragment (Fig. 1b) of pBR322 (2576 - 4019) was cut twice by MmeI to give new fragments of 1340 and 1150bp approximately. Similarly, the 1374bp Sau3A "A" fragment (Fig. 1d; 1668 - 3041) was cut twice by MmeI in a double digest (Fig. 1e) to give fragments of about 1200 and 1010bp. These data suggest (i) that there are two MmeI sites in the region of overlap between the TaqI "A" and Sau3A "A" fragments and (ii) that they are separated by about 190bp. This leaves at least two more MmeI sites to account for. The MmeI/PstI double digest data (Fig. 1f) are consistent with the location of two sites near coordinate 250: and the MmeI single digest data (Fig. 1a), taken together with all the double digest data, suggest that no more than four sites are present. The crude map was confirmed with other double digest experiments (not shown) and refined using a computer program (22). The MmeI cut points estimated in this way were at coordinates 212, 296, 2697 and 2874 with a probable error of less than 50bp either side. (See Fig. 2 for a restriction map of pBR322 showing the four MmeI sites.)

Determination of MmeI recognition sequence

There are two main methods for determining the recognition sequence of a restriction enzyme. The first and most direct is the biochemical one of sequencing around several cut points and searching by eye for homologous sequences near or around these points of cleavage (23). The second is an indirect, usually computer-based, method which exploits the fact that the complete sequences of several replicons are known; e.g., pBR322 (6), PhiX174 (24), SV40 (25) and M13 (26). Several techniques of the second kind have been described (2, 3, 4, 5). All are heuristic in approach; that is, by assuming that a Type II enzyme's recognition sequence will conform to one of a number of "patterns" (here called "templates"), the search procedure is simplified. The basis for this assumption is the observation that the recognition sequences of most restriction enzymes are palindromic: e.g., EcoRI recognizes the sequence GAATTC, which is the same as its complement

(1). A significant proportion of such enzymes, however, allow for a certain flexibility or degeneracy at some positions of their recognition sequences. An example of this kind is HincII, which recognizes GTYXAC. This is a shorthand way of denoting the four sequence permutations the enzyme will cleave, i.e., GTTAAC, GTCAAC, GTTGAC and GTCGAC. Any procedure for determining the recognition sequences of restriction enzymes must therefore allow for the possibility of degeneracy at some positions. A small number of recognition sequences, however, are entirely non-palindromic: these include GAAGA and its distinct complement TCTTC, the sequence recognized by MboII (27). Of the computer programs so far described, only one (RSITE; (5)), can deal with enzymes like MboII, but it cannot predict a degenerate non-palindromic sequence like that of Tth111II, CAAXCA (28).

A common feature of all these programs is that they work on fragment length data. The two which are restricted to palindromic sequences generate, from a set of template sequences, tables which give the number of predicted sizes of fragments resulting from cleavage of a particular DNA at these sequences. The number and sizes of fragments produced by the enzyme under study are determined empirically and the table is scanned either manually (2) or automatically (3,4) for a template sequence giving similar data. Any such sequence is a good candidate for the recognition sequence. The RSITE program (5) is highly interactive. Since it will find certain non-palindromic sequences as well as palindromic sequences, it does not generate tables of the kind produced by the other programs as they would be very much larger and difficult to interpret. Instead, it works by a stepwise refinement procedure in which the user enters fragment sizes and estimated errors, one by one, until the program can eliminate all but one candidate sequence, which is then output. The non-palindromic recognition sequence of HinGUII (GGATG) was determined in this way (5), after attempts to find a palindromic recognition sequence failed.

The success of the method is entirely dependent on the scope of the templates used by the program. It is clearly desirable to define the set in such a way that all known recognition sequences, and likely extensions, are covered. To this end, a new program (RECOG) was developed to determine the recognition sequence of MmeI. Like the previously described programs (2,3, 4), RECOG incorporates a table-generation section for testing palindromic sequences, but uses a new algorithm for testing non-palindromes. In each case, a more comprehensive set of templates, which attempts both to cover all known sequence types (1) and to anticipate some new ones, was included.

Table 1. Palindromic templates used by RECOG. Example enzymes are included if known (1). (The symbols "p" and "q" represent any bases which conserve palindromicity.)

1. Non-degenerate tetrameric palindromes, AATT - TTAA:

ppqq
AluI

2. Pentameric palindromes with central degeneracy:

ppNqq	pp8qq	pp2qq
HinfI	EcoRII	CauII

3. Non-degenerate hexameric palindromes, AAATTT - TTTAAA:

pppqqq
EcoRI

4. Degenerate hexameric palindromes:

ppXYqq	ppYXqq	pp88qq	pp22qq	pp19qq	pp91qq
AflIII	HincII	------	------	AccI	------
pXpqYq	pYpqXq	p8pq8q	p2pq2q	p1pq9q	p9pq1q
AcyI	AvaI	HqiAI	------	NspBII	------
XppqqY	YppqqX	8ppqq8	2ppqq2	1ppqq9	9ppqq1
HaeII	CfrI	HaeI	------	------	------

5. Non-degenerate tetrameric and hexameric palindromes with one to six unspecified central bases: (ppNqq covered in 2. above)

ppNNqq	ppNNNqq	ppNNNNqq	ppNNNNNqq	ppNNNNNNqq
NlaIV	-------	---------	---------	----------
pppNqqq	pppNNqqq	pppNNNqqq	pppNNNNqqq	pppNNNNNqqq
SauI	--------	Tth111I	XmnI	BglI

pppNNNNNNqqq
HqiEII

6. Degenerate hexameric palindromes with one to six unspecified central bases:

(This is an obvious extension of section 4., comprising templates ranging from ppXNYqq to 9ppNNNNNNqq1. Accordingly, there are 6 x 18 = 108 templates altogether. Only one known enzyme (DraII, recognizing XGGNCCY) fits any of these templates.)

Table 1 shows the set of palindromic templates incorporated within RECOG. Virtually all known restriction enzymes specificities (1) fit one or more of these templates. Two notable exceptions are GCGGCCGC (NotI) and

Table 2. Non-palindromic templates used by RECOG. Example enzymes are included if known (1). (The symbol "p" here represents any base.)

1. Non-degenerate tetrameric sequences, AAAA - TTTT:

pppp
MnlI

2. Non-degenerate pentameric sequences, AAAAA - TTTTT:

ppppp
MboII

3. Degenerate pentameric sequences:

Xpppp	pXppp	ppXpp	Ypppp	pYppp	ppYpp
-----	-----	-----	-----	-----	-----

4. Non-degenerate hexameric sequences, AAAAAA - TTTTTT:

pppppp
BbvII

5. Degenerate hexameric sequences:

Xppppp	pXpppp	ppXppp	Ypppp	pYpppp	ppYppp
------	------	------	GdiII	------	Tth111II

GGCCNNNNNGGCC (SfiI) (1), both essentially octameric sequences. These sequences occur very infrequently, so RECOG is ill-adapted to investigate such enzymes. This argument also applies to the enzyme RsrII, recognizing CGG8CCG (1). Non-palindromic templates are shown in Table 2. Clearly, the number of templates needed to cover most known specificities is low: but, obviously, almost all seqences fit one or more of them (in contrast to the palindromic templates), hence the requirement for a better algorithm to identify potential recognition sequences.

A search for a palindromic recognition sequence for MmeI was conducted first. The mapping of the four MmeI sites in pBR322 described above showed that the largest complete digest fragment produced was 2350 - 2450bp long. The table of palindromes generated by RECOG was scanned automatically for all sequences which (i) occurred at least four times and (ii) would produce a fragment of this size. 44 palindromes of this kind were found (Table 3). All could be rejected by noting that the second largest fragment predicted for each sequence was outside the size range (1665 - 1735bp) determined for the MmeI "B" complete digest fragment of pBR322. Thus, none of the large number of palindromic sequences generated from the templates within RECOG

Table 3. Palindromic sequences which are candidates for the recognition sequence of MmeI. Each sequence is preceded by the number of occurences in pBR322, and followed (in brackets) by the sizes of the two largest predicted fragments. Thus, the sequence 1GGCC9 occurs four times, with the two largest predicted fragments being 2440 and 1578bp long. (Modified from program RECOG output.)

4	1GGCC9	(2440, 1578)	4	A9TNNNNA1T	(2374, 1005)
4	XCATGY	(2449, 1254)	4	9GGNNNNCC1	(2407, 1437)
9	TC88GA	(2352, 1008)	5	TA9NNNN1TA	(2381, 855)
			6	C2CNNNNG2G	(2352, 1005)
4	1GTNAC9	(2351, 1077)	7	GA9NNNN1TC	(2424, 561)
4	8CTNAG8	(2356, 1115)	9	1TGNNNNCA9	(2363, 818)
4	8TGNCA8	(2410, 1242)	10	2CGNNNNCG2	(2397, 711)
5	9ATNAT1	(2378, 803)			
7	AGXNYCT	(2447, 767)	5	AYCNNNNNGXT	(2367, 1009)
			5	GXANNNNNTYC	(2367, 614)
4	2TCNNGA2	(2352, 1008)	6	T8ANNNNNT8A	(2375, 900)
4	AT9NN1AT	(2401, 1817)	9	TA8NNNNN8TA	(2384, 463)
5	CG9NN1CG	(2359, 1404)	10	2CGNNNNNCG2	(2381, 657)
5	C1GNNC9G	(2385, 883)			
5	XAGNNCTY	(2442, 908)	4	XGGNNNNNNCCY	(2350, 1339)
7	2ACNNGT2	(2390, 696)	4	GA9NNNNNN1TC	(2352, 1283)
8	2GCNNGC2	(2432, 866)	4	1ACNNNNNNGT9	(2373, 1437)
			4	AG1NNNNNN9CT	(2384, 962)
4	CG8NNN8CG	(2356, 850)	4	AC2NNNNNN2GT	(2444, 1082)
4	CT9NNN1AG	(2375, 1044)	5	2CCNNNNNNGG2	(2368, 882)
5	TT2NNN2AA	(2370, 1125)	7	TYANNNNNNTXA	(2357, 1056)
5	1TTNNNAA9	(2383, 1039)	7	9CANNNNNNTG1	(2359, 1351)
5	TC9NNN1GA	(2418, 914)	8	2GCNNNNNNGC2	(2439, 682)
5	C2CNNNG2G	(2435, 1470)	9	GC2NNNNNN2GC	(2438, 760)
6	XTTNNNAAY	(2390, 786)			

was the recognition sequence for MmeI.

The algorithm which searches for non-palindromic sequences does not use fragment length data at all: instead, it uses the mapped cut sites on a sequenced DNA as input. The program extracts subsequences of user-specified length centred around each mapped site. From these subsequences, firstly, all posssible hexamers are extracted and converted to potential recognition sequences using the non-palindromic templates (Table 2). For example, the sequence CGGTAC is converted to CGGTAC, YGGTAC, CXTGAC, CGXTAC and the complements GTACCG, XTACCG, GYACCG and GTXCGG. When this is completed for all subsequences, RECOG simply searches for any sequence common to each. If none occurs, the process is repeated for pentamers and tetramers in turn until a common sequence is found. (Note that the algorithm does not require that all cut sites are input, or that those input should be adjacent. This would be useful, for example, when investigating an enzyme which produces a

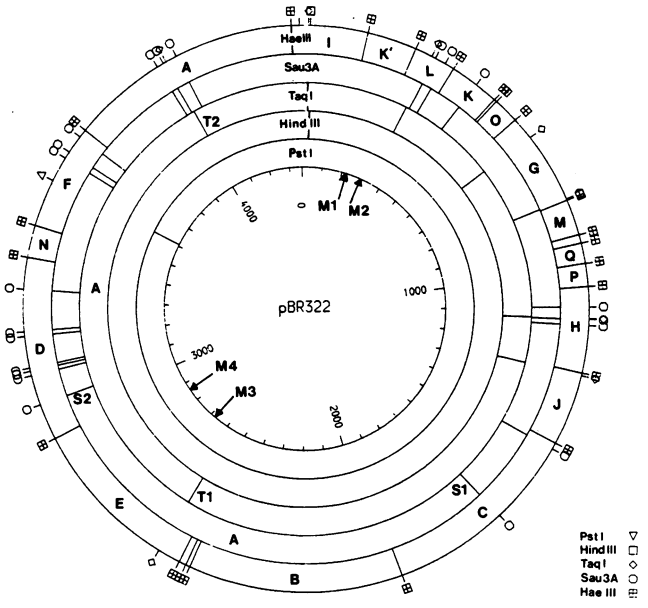

Figure 2. Restriction map of pBR322 showing the MmeI sites M1, M2, M3 and M4 in relation to other enzyme sites used in mapping.

complex pattern of fragments allowing only a partial restriction map to be compiled.) The algorithm was applied to the four mapped MmeI cut sites at coordinates 212, 296, 2697 and 2874 of pBR322. Only one hexameric sequence, GTYGGA, was found by the program. This occurs (in the same orientation) at coordinates 197-202, 284-289, 2664-2669 and 2848-2853. Moreover, the sequence occurs nowhere else on the plasmid, in agreement with the mapping data, thus making it a strong candidate for the recognition sequence of MmeI. Since the sequence GTYGGA is on the negative strand of the pBR322 sequence as it is normally written (6), it will be referred to henceforth by its complement TCCXAC. Both variants of this sequence occur in pBR322: TCCGAC at M1, M2 and M3 and TCCAAC at M4 (Fig. 2). In each case, the mapped cut sites are located slightly 3' of this sequence in pBR322, suggesting that MmeI cleaves outside its recognition sequence. This is a property shared by most enzymes with non-palindromic recognition sequences (1). The output from GELSIM (described below) was generated on the essentially arbitrary assumption that the cleavage site of MmeI was 12bp 3' of its recognition sequence.

Confirmation of the recognition sequence

The sequence TCCXAC was confirmed as the recognition sequence of MmeI

by restriction analysis of another sequenced DNA, PhiX174 (24). There are five occurrences of TCCXAC within PhiX174 RF DNA at co- ordinates -225, 2691, -3237, 5197 and 5376. All except the second have an adenine residue at the degenerate purine position. (A minus sign preceding a coordinate here, and below, denotes the occurrence of the sequence on the complementary strand.) It would have been possible, therefore, to confirm the recognition sequence by mapping the MmeI sites in PhiX174 and comparing them to the predicted ones. But, given the difficulties caused by persistent incomplete digestion encountered in mapping pBR322, it was decided to adopt a novel approach. A computer procedure (GELSIM) linking a small set of programs was devised to produce graphical simulations of agarose gel electrophoretograms. It uses as input (i) the size/mobility equation (I) deduced from mobilities of marker fragments of known size on the gel (21) and (ii) the sizes of fragments predicted from the putative recognition sequence. A comparable program has been described (29) which simulates polyacrylamide gel electrophoretograms, using the semi-logarithmic relationship between size and mobility. The mobilities of fragments are deduced from the equation and plotted on a graph as lines. Another program within GELSIM can generate the predicted fragment sizes from a query sequence such as TCCXAC with the option of including all possible partial products as well. It uses two formulae, derived by simple induction, which give the maximum number (T) of different-sized fragments obtainable in a partial digest of linear or circular molecules containing N sites. They are:

$$T = (N + 1)(N + 2)/2 \quad \text{(for linear molecules);} \quad \text{(II)}$$

$$T = N^2 - N + 1 \quad \text{(for circular molecules)} \quad \text{(III)}$$

Double digests are simulated in a similar way. The net effect is a graph which, with appropriate linear scaling, can be compared band by band with a photograph of the gel: the need to map restriction sites is thus eliminated. Of course, it is possible to compare the predicted with actual fragment size data as numbers, but, without recourse to statistics, real discrepancies would be less obvious to the eye than in the graphical representation.

Before analysing a PhiX174 digest, GELSIM was tested by simulating some of the gels used in the mapping of MmeI sites in pBR322 (Fig. 1). The GELSIM procedure was provided with the sequence of pBR322, the recognition sequences of TaqI, Sau3A and PstI, the query sequence TCCXAC and suitable

size/mobility equations deduced from gels. The correspondence between simulations and actual gel photographs is striking, but not conclusive, of course, since the same data were used to deduce the sequence TCCXAC in the first place. (In the TaqI/MmeI and Sau3A/MmeI double digests, only the bands actually seen on the gels are shown in the simulations. The reason for some bands being missing is probably the fact that MmeI does not cut DNA stoichometrically. Therefore, any band produced by MmeI will be less intense than a similar-sized one produced by the other enzyme. Small or co-migrating fragments would thus tend to be invisible. The effects of cleavage at all four MmeI sites, however, are visible in the MmeI and PstI/MmeI digests - see Figs. 1a and 1f.)

PhiX174 RF DNA, previously digested with PstI, was restricted with MmeI. If TCCXAC is the MmeI recognition sequence, the sequence at 5376 can be discounted since it is less than 10bp from the PstI site. The PstI-restricted DNA can then be considered as a linear molecule containing four TCCXAC sequences. From formula (II) above, up to 15 distinct fragment sizes should be produced. The sequence of PstI-linearized PhiX174 DNA was input to GELSIM, together with the query sequence TCCXAC. When the size/mobility equation deduced from a nearby marker track was input, the simulated pattern was seen to resemble closely that of the actual PstI/MmeI digest. Due to "smiling" of the gel, however, the band-to-band correspondence was slightly distorted. To remedy this, some of the bands in the PhiX174 digest itself were used to deduce a size/mobility equation. When this was done, the overall pattern was as before, but the band-to-band correspondence was improved. The result is seen in Fig. 3a, and confirms TCCXAC as the recognition sequence of MmeI. (In addition, a MmeI digest of SV40 DNA was analysed. The SV40 sequence contains two MmeI sites at 1020 and -4564 (25), and the resulting gel pattern of two complete digest bands with a partially-digested full-length linear band was observed (not shown), supporting the above result.)

Another sequenced molecule, M13mp7 RF DNA, was analysed in an exactly similar way. The sequence (obtained from B.K. Ely at N.I.M.R., Mill Hill, London) was deduced from that of M13 (26) and part of the E.coli lac operon as modified to contain a polylinker (30,31). The coordinate system is close to that of M13 itself (26), with the ca. 830bp lac sequence inserted near coordinate 5870, giving a total length for M13mp7 of ca. 7236bp. TCCXAC occurs four times in the deduced sequence at coordinates -300, 5441, -5762 and 6613. Thus the sequence at 6613 is within the lac region. Formula (III) predicts that 13 different-sized fragments should be produced in a

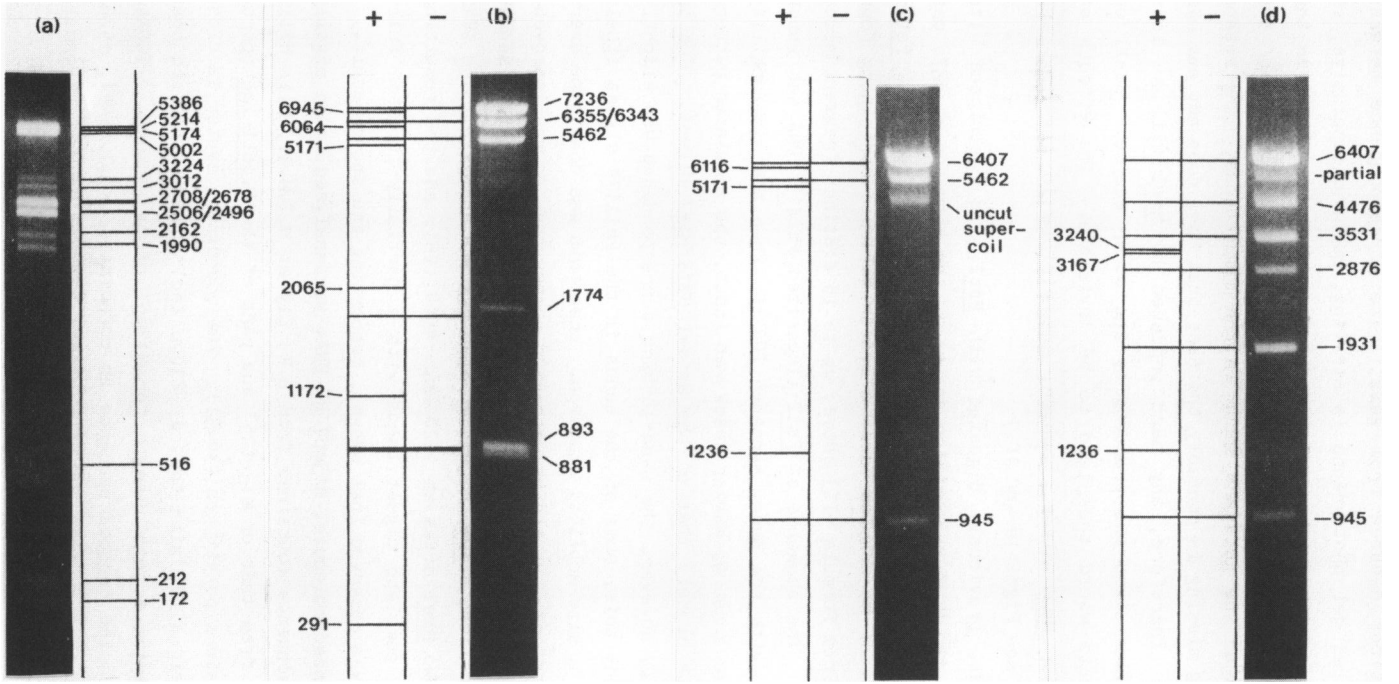

Figure 3. Confirmatory restriction digests: (a) PhiX174, PstI/MmeI; (b) M13mp7, MmeI; (c) M13, MmeI; (d) M13, BamHI/MmeI. Tracks are flanked by simulations and fragment sizes are labelled as before (Fig. 1). The "+" simulations of (b), (c) and (d) are produced by GELSIM if the sequence at 5441 is assumed to be cut by MmeI: the "-" simulations are obtained if that sequence is assumed not to be a MmeI site. (The band labelled "partial" in (d) is due to incomplete BamHI cleavage.)

MmeI digest of this molecule. Comparison of simulations with real digests showed that some bands were missing from the gel. By carrying out trial simulations it was found that the empirical pattern could be simply obtained if it was assumed that the TCCAAC site at 5441 was not cut by MmeI (Fig. 3b). This assumption was supported by a simulation of a double digest of M13mp7 by MmeI and EcoRI (not shown). Clearly, bands of 2065, 1172 and 291bp (Fig. 3b) predicted by the presence of the 5441 TCCAAC sequence are not visible. It could be, of course, that the non-stoichiometry of MmeI activity noted above is here manifested to a very high degree; i.e., that there is a recognition sequence for the enzyme at 5441, but no detectable cleavage. If there is no cleavage at all, however, three possibilities suggest themselves: (i) sequences flanking TCCAAC at 5441 somehow prevent cutting; (ii) methylation in the sequence prevents site recognition or, more mundanely, (iii) there has been a mutation or sequencing error. To test (iii), a sample of M13 RF DNA prepared from the same phage stock used in its sequencing (26) was obtained from the laboratory of Prof. J.G.G. Schoenmakers. As stated above, the M13 sequence (6407bp long) is entirely contained within that of M13mp7, with minor differences (31), and thus contains the three TCCXAC sequences at coordinates -300, 5441 and -5762. Formula (III) predicts that up to seven distinct fragments would be produced in a MmeI digest of M13 RF. Comparison of MmeI and MmeI/BamHI digests of this DNA with the simulations showed that again there was no detectable cleavage at the TCCAAC sequence at 5441 (Figs. 3c and 3d), i.e., that some bands are missing. This shows that there has not been a site-destroying mutation at 5441 during the construction of M13mp7 from M13, but still allows the possibility of an error in the sequencing of M13 itself.

To test this possibility, the appropriate region of M13, within the vector M13mp9, was sequenced by the dideoxynucleotide method (32), using a synthetic oligodeoxyribonucleotide (5'-GCGAAAGGAGCGGGCGC-3'), which hybridises to nucleotides 5568 to 5584, as primer. The data showed the sequence from 5441 to 5446 to be TCTAAC, which is not a recognition site for MmeI. This result, together with the demonstration that the corresponding region of the parental M13 DNA is not cleaved by MmeI [Fig. 3(c) and 3(d)], is sufficient proof that the published sequence (2b) was in error at this position. This finding is not too surprising, since the corresponding sequences in the DNAs of two other phages, fd (33) and f1 (34), that are closely related to and highly homologous with M13, are both TCTAAC.

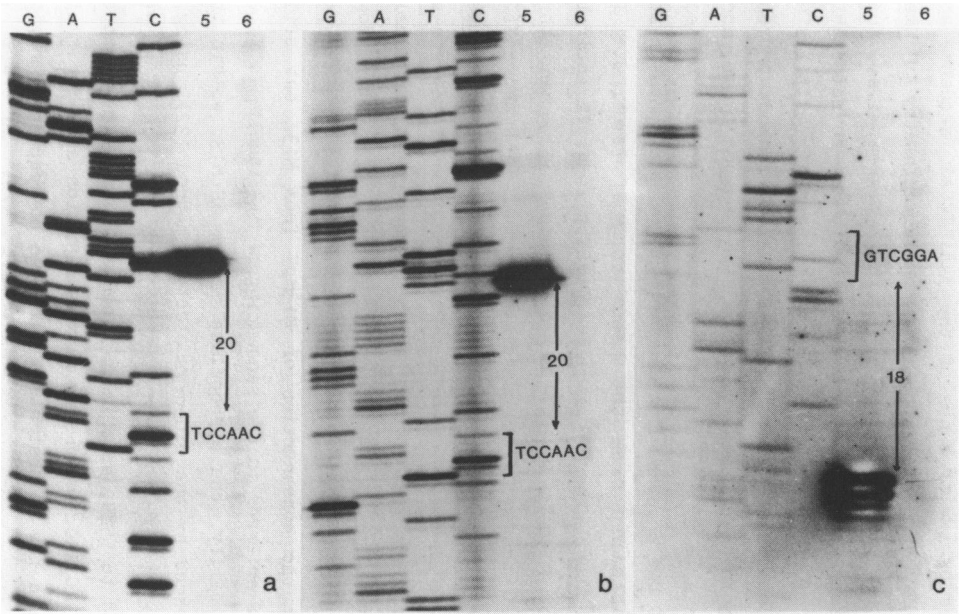

Figure 4. The site of cleavage of M13mp9 DNA by MmeI endonuclease. Three synthetic oligonucleotides, labelled at their 5' ends, were used as primers for Klenow fragment DNA polymerase on single-stranded M13mp9 template DNA. The products of these reactions were cleaved with MmeI and analysed on sequencing gels (track 5 in each case) alongside conventional dideoxy-sequencing reactions carried out with the same primers (tracks G,A,T,C in each case). The tracks labelled 6 in a), b) and c) are controls in which the extended primers are not treated with MmeI. (The multiple bands in track 6 of c) probably arise from the presence of shorter oligonucleotides within the primer preparation.)

a) The primer was 5'-GCAAAGCGGATTGCATC-3', (complementary to nucleotides 404-388 of M13mp9 DNA);

b) The primer was 5'-AAAATCCCTTATAAATC-3', (complementary to 5857-5841);

c) The primer was 5'-GCCATCAAAAATAATTC-3', (complementary to 7057-7041 of M13mp9).

Determination of the MmeI cleavage-site

The relationship between the sites of cleavage and the recognition sequence for MmeI were determined directly on double-stranded DNA substrates generated by the extension of specific deoxyoligonucleotide primers on M13mp9 template DNA. The method is a modification of a general procedure previously described (23). Three oligomers were designed to hybridise to sequences about 60 nucleotides distal to the three MmeI recognition sites in M13mp9. The 5'-^{32}P-labelled primers were extended on the single-stranded template by Klenow DNA polymerase and the reaction products were cleaved with MmeI. The end-labelled DNA fragments produced were analysed alongside dideoxynucleotide sequencing ladders produced using the same synthetic oligonucleotide primers on the M13mp9 template.

At MmeI sites 1 and 2, each having the recognition-sequence GTTGGA on the template strand, the enzyme cuts the substrate DNA strand 20 nucleotides 3' to the end of the recognition sequence (Fig. 4a and b). At site 3, where the recognition sequence is TCCGAC on the template strand, the cut-site is 18 nucleotides 5' to the recognition sequence (Fig. 4c). These data indicate that the enzyme generates a staggered break by cutting 20 nucleotides 3' to the recognition sequence TCCXAC and 18 nucleotides 5' to its complement.

DISCUSSION

The recognition sequence of MmeI, TCCXAC, is only the fourth so far described which is a non-palindrome containing a degenerate base position. The other three are Tth111II (CAAXCA), GdiII (YGGCCG) and TaqII (2ACC2A) (28, 1, 35). It seems likely that many more enzymes of this kind exist. Since, in general, such enzymes cleave the DNA some distance away from their recognition sequences (1), biochemical analysis of the generated ends can fail to identify these sequences. The programs described above, therefore, may well be more successful in these cases. All that is required for the method to work is the mapping of several (at least three) cut-sites on a sequenced DNA. One potential problem arises if the substrate DNA sequence is incorrect at a mapped site: this might result in the program failing to find a consensus. In this event, providing various subsets of the mapping data as input to RECOG could prove successful. If all such efforts fail, the final option is to extend the repertoire of templates within RECOG. The template sets shown above (Tables 1 and 2) were not supplemented after the MmeI recognition sequence was found, but there are obvious extensions. This is especially true for degeneracies witin non-palindromic sequences. Such

an extension would be necessary, for example, to identify recognition sequences like 2ACC2A (TaqII) (35). As shown above, when a candidate sequence is identified, it can be rapidly confirmed by generating simulations of predicted digests and comparing these with empirical data. Of course, GELSIM has a more general application: it can be used to generate simulations of digests using any known restriction enzyme(s), providing the sequences of the substrate DNA is known. This has many uses in the design of cloning experiments, with the rapidly increasing availability of plasmid and phage vector sequences.

The tendency of MmeI to produce all possible partial and complete products in a digest is unexplained, but at least one other enzyme, HphI (36), has a similar property. One possibility is that the enzyme can either cleave one or both strands at any site, but that sites at which one strand has been cleaved are resistant to subsequence complete cleavage. This, however, remains to be tested using a highly-purified preparation of the enzyme.

The method used here to determine the position of cleavage of the DNA by MmeI, using readily obtainable synthetic oligonucleotide primers in combination with a general protocol for determining the nature of the 3' ends generated (23), should be widely applicable. It is likely to be most useful for those enzymes that act outside the recognition sequence itself and therefore produce termini with random nucleotide sequences. It is fortunate that the three MmeI recognition sites in M13mp9 DNA provided examples of both possible recognition sequences (TCCAAC and TCCGAC) and of both orientations with respect to the primers.

The MmeI endonuclease was shown to cleave its substrate twenty nucleotides, or two full turns of the DNA helix, away from its recognition sequence. No previously described Type II enzyme acts at such a distance from its recognition site, though the Type III enzymes, HinfIII and HineI, cleave about 25 base-pairs 3' of their recognition sequences (1). The non-palindromic recognition sequence of MmeI, along with its apparent inability to give complete cleavage of substrate DNA, are other properties in common with Type III enzymes. The lack of requirement of MmeI for ATP contrasts with the Type III enzymes, however, and suggests that MmeI should be classified as a Type II restriction endonuclease.

The description "restriction enzyme" is now applied to any sequence-specific endonuclease, although this class of enzymes was originally defined genetically in E.coli K12 by the observation that hsdR mutants (lacking the

Type I enzyme Endo R.EcoK) did not restrict (destroy) incoming, unmodified phage Lambda DNA (37). A mutant of M.methylotrophus, partially deficient in restriction of incoming plasmid DNA from heterologous donors such as E.coli, has been isolated. Crude extracts of this mutant reveald no detectable MmeI activity, showing that the enzyme acts as a restriction enzyme in vivo. Similar mutants lacking MmeII activity have also been isolated and characterized ((14); manuscript in preparation).

ACKNOWLEDGEMENTS

This research was funded by an SERC/CASE award in conjunction with Imperial Chemical Industries and by an SERC Project Grant (GR/C/94711). We wish to thank colleagues who provided bacterial strains and DNA; John Windass, who first detected endonuclease activity in extracts of M.methylotrophus; Julie Horton for help with DNA sequencing; Dave Byrom and the rest of the ICI Joint Lab. for helpful discussions; the staff of the Computer Lab. for system software support and Anne Bates, for typing the manuscript.

*Present address: Department of Genetics, Glasgow University, Church Street, Glasgow G11 5JS, UK

REFERENCES

1. Kessler, C., Neumaier, P.S. and Wolf, W. (1985) Gene 33, 1-102.
2. Fuchs, C., Rosenvold, E.C., Honigman, A. and Szybalski, W. (1978) Gene 4, 1-23.
3. Gingeras, T.R., Milazzo, J.P. and Roberts, R.J. (1978) Nucl. Acid Res. 5, 4105-4127.
4. Keller, C., Corcoran, M. and Roberts, R.J. (1984) Nucl. Acid Res. 12, 379-384.
5. Tolstoshev, C.M. and Blakesley, R.W. (1982) Nucl. Acid Res. 10, 1-17.
6. Sutcliffe, J.G. (1978) Cold Spring Harbor Symp. Quant. Biol. 43, 77-90.
7. Peden, K.W.C. (1983) Gene 22, 277-280.
8. Windass, J.D., Worsey, M.J., Pioli, E.M., Pioli, D., Barth, P.T., Atherton, K.T., Dart, E.C., Byrom, D., Powell, K. and Senior, P.J. (1980) Nature 287, 396-401.
9. Birnboim, H.C. and Doly, J. (1979) Nucl. Acid Res. 7, 1513-1523.
10. Ish-Horowicz, D. and Burke, J.F. (1981) Nucl. Acid Res. 9, 2989-2998.
11. Marinus, M.G. (1973) Molec. Gen. Genet. 127 47-55.
12. Miller, J.H. (1972) Experiments in Molecular Genetics, Cold Spring Harbor Laboratory, New York.
13. Greene, P.J., Heyneker, H.L., Bolivar, F., Rodriguez, R.L., Betlach, M.C., Covarrubias, A.A., Backman, F., Russel, D.J., Tait, R. and Boyer, H.W. (1978) Nucl. Acid Res. 5, 2373-2380.
14. Boyd, A.C. (1983) Ph.D. thesis, Leicester University.
15. Modrich, P. and Geier, G.E. (1979) J. Biol. Chem. 254, 1408-1413.
16. Matthes, M.W.D., Zenke, W.M., Grundstrom, T., Staub, A., Wintzerith, M. and Chambon, P. (1984) EMBO. J. 3, 801-805.
17. Sproat, B.S. and Gait, M.J. (1985) Nucl. Acid Res. 13, 2959-2978.
Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982). In Molecular

Cloning, a laboratory munual. Cold Spring Harbor Laboratory.
18. Maxam, A.M. and Gilbert, W. (1980) In Methods in Enzymology, Grossman, L. and Moldave, K., Eds., Vol. 65, pp. 499-560, Academic Press, New York.
19. Biggin, M.D., Gibson, T.S. and Hong, C.F. (1983) Proc. Nat. Acad. Sci. USA. 80, 3963-3965.
20. Southern, E.M. (1979) Anal. Biochem. 100, 319-323.
21. Schaffer, H.E. and Sederoff, R.R. (1981) Anal. Biochem. 115, 113-122.
22. Schroeder, J.L. and Blattner, F.R. (1978) Gene 4, 167-174.
23. Brown, N.L. and Smith, M. (1980). In Methods in Enzymology, Grossman, L. and Moldave, K. Eds., Vol. 65, pp. 391-404, Academic Press, New York.
24. Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M. and Smith, M. (1978) J. Mol. Biol. 125, 225-246.
25. Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., van de Voorde, A., van Heuverswyn, H., van Herreweghe, J., Volckaert, G. and Ysebaert, M. (1978) Nature 273, 113-120.
26. van Wezenbeek, P.M.G.F., Hulsebos, T.J.M. and Schoenmakers, J.G.G. (1980) Gene 11, 129-148.
27. Hutchison, C.A. and Smith, M. (1980) J. Mol. Biol. 140, 143-148.
28. Shinomiya, T., Kobayashi, M. and Sato, S. (1980) Nucl. Acid Res. 8, 3275-3285.
29. Lilley, D.M.J. (1982) Nucl. Acid Res. 10, 19-26.
30. Messing, J., Crea, R. and Seeburg, P.H. (1981) Nucl. Acid Res. 9, 309-321.
31. Yanisch-Perron, C., Vieira, J. and Messing, J. (1985) Gene 33, 103-119.
32. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Nat. Acad. Sci. USA. 80, 3963-3965.
33. Sugimoto, K., Sugisaki, H., Okamoto, T. and Takanami, M. (1978) Nucl. Acid Res. 5, 4495-4503.
34. Beck, E. and Zink, B. (1981) Gene 16, 35-38.
35. Barker, D., Hoff, M., Oliphant, A. and White, R. (1984) Nucl. Acid Res. 12, 5567-5581.
36. Kleid, D.G. (1980). In Methods in Enzymology, Grossman, L. and Moldave, K. Eds., Vol. 65, pp. 163-166, Academic Press, New York.
37. Wood, W.B. (1966) J. Mol. Biol. 16, 118-133.