Volume 5 Number 11 November 1978 Nucleic Acids Research

A computer assisted method for the determination of restriction enzyme recognition sites

T.R.Gingeras, J.P.Milazzo* and R.J.Roberts

Cold Spring Harbor Laboratory, Cold Spring Habor, NY 11724, and *Computing Center, State University of New York, Stony Brook, NY 11794, USA

Received 7 August 1978

ABSTRACT

A computer program has been developed which aids in the determination of restriction enzyme recognition sequences. This is achieved by cleaving DNAs of known sequence with a restriction endonuclease and comparing the fragmentation pattern with a computer-generated set of patterns. The feasibility of this approach has been tested using fragmentation patterns of ØX174 DNA produced by enzymes of both known and unknown specificity. Recognition sequences are predicted for two restriction endonucleases (BbvI and SfaNI) using this method. In addition, recognition sequences are predicted for two other new enzymes (PvuI and MstI) using another computer-assisted method.

INTRODUCTION

The Type II restriction endonucleases are invaluable to the molecular biologist, for they allow the dissection of large DNA molecules into discrete fragments. This is possible because they recognize a specific sequence of bases within the DNA molecule and cleave at or close to that sequence (1,2). Deduction of the recognition sequence thus becomes an essential element in the characterization of a new restriction endonuclease. When the site of cleavage lies within the recognition sequence, relatively straightforward biochemical procedures are available for its determination (2). However, some enzymes exist whose recognition sequence lies some nucleotides away from the site of cleavage, a situation which renders its identification more difficult. The first such example was HphI which recognizes a unique pentanucleotide but cleaves 7 or 8 nucleotides away from this sequence (3). The nature of the recognition sequence was deduced by the laborious process of deriving sequence information from regions surrounding many HphI sites and comparing those sequences until a common feature emerged. Similar strategies were necessary for the enzymes MboII (4,5) and HgaI (6) which also recognize unique

pentanucleotides and cleave away from these sequences.

Since the complete sequences of several small DNA genomes are now available, an alternative method for the determination of recognition sequences seemed appropriate. A recognition sequence may be viewed as a pattern which is repeated at various positions along a DNA molecule. The distance between successive occurrences of the pattern is reflected in the lengths of the fragments generated by the restriction enzyme. Were it possible to measure these fragment lengths exactly, it would seem reasonable to believe that the pattern could be unambiguously deduced by searching the original DNA sequence for the occurrence of patterns which repeated themselves at exact intervals corresponding to the fragment lengths. Unfortunately, present biochemical procedures give only approximate values for these fragment lengths. Nevertheless, it is still conceivable that unique solutions will exist which should correspond to the restriction enzyme recognition sites. In this paper, we explore the feasibility of using the computer to perform this task of pattern recognition and thus to predict restriction enzyme recognition sites.

MATERIALS AND METHODS

Replicative form ØX174 DNA (am3) was a gift from G.N. Godson and R.W. Chambers. The following restriction endonucleases were prepared and used according to published procedures: HhaI (7), HpaI and HpaII (8), MboII (9), and TaqI (10). BbvI from Bacillus brevis (ATCC 9999) and PvuI from Proteus vulgaris (ATCC 13315) were prepared by a standard procedure (TRG and RJR, to be published) and SfaNI from Streptococcus faecalis was a gift from D. Sciaky. MstI from a Microcoleus strain was a gift from New England Biolabs. Digestion with these enzymes (5-15 µl) was achieved using 2 µg ØX174 RFI in a reaction mixture (50 µl) containing 6 mM Tris-HCl pH 7.9, 6 mM $MgCl_2$ and 6 mM $SHCH_2CH_2OH$ at 37°C (except for TaqI, which was incubated at 65°C). Electrophoresis was carried out on 1.4% or 2.0% agarose gels as previously described (11). The fragments resulting from digestion of ØX174 RF with TaqI were used as length standards, based upon their known sequence, and used to calibrate a semi-log plot of mobility against molecular weight. All other fragment lengths were estimated from their mobilities in agarose gels by comparison with these standards.

Programs were written in ASC II Fortran and executed on a UNIVAC 1110 computer. The nucleotide sequence of the viral strand of the ØX174

genome (12) was stored in the computer under the file name of PHIXSEQ.

Program Descriptions

BB

This program generates a list of all possible unique tetranucleotides (NNNN)*, pentanucleotides (NNNNN) and hexanucleotides (NNNNNN). Bearing in mind that only one strand of the ØX174 sequence was stored in the computer, it was necessary to account for the fact that a unique non-palindromic sequence such as AAAC would be equivalent to its inverse complement GTTT (i.e., the second strand sequence) when computing the length of fragments produced by cleavage at this site. In all, 136 different tetranucleotides and their complements must be considered, 512 pentanucleotides, and 2080 hexanucleotides. The subsequent listing of these 2728 sequences was stored under the file name BASES.

FTAB

This program was used to scan PHIXSEQ and provide the distance between successive occurrences of any given sequence of nucleotides. It was driven by the 2728 entries of BASES and produced a list of the predicted fragment lengths for any restriction endonuclease which recognized one of these sequences. A sample output from this program is shown below.

Sequence	Code No.	Fragment Lengths (in nucleotide pairs)
ACGAC	230	20 63 90 107 165 192 389 390 399 814 946 1588.

This program takes into account the circular nature of the ØX174 genome and the complete table is referred to as the PHIXIBUF table.

An additional program, SEARCHFOR, is also available which can produce both these fragment lengths and the location of their endpoints. This was used to cross-check many of the entries in the PHIXIBUF table to ensure its accuracy.

* All sequences are written 5'→3' and N represents any one of the four deoxyribonucleotides.

FRAGLEN

This program rearranged the PHIXIBUF table into a format such that it could be easily searched, for all sequences (from BASES) able to generate a fragment of given length. The resulting table, termed the "Master Table", contains a listing of all fragment sizes found in the PHIXIBUF table, in increasing order of size. For each fragment length entry, all possible sequences able to generate such a fragment are listed. An example of the contents of this table is shown below:

ØX Fragment Length	No. of Occurrences	Sequences which Produce Fragments of this Size
100	26	8, 10, 14, 21, 37, 46, 51, 57, 58, 61, 61, 65, 101, 104, 127, 216, 340, 467, 531, 536, 594, 617, 701, 968, 1136.

Sequences are written in codified form in order to save space within the computer. A utility written into this program translates these codes into actual sequences. The codes are the same as those seen in column 2 of the PHIXIBUF table.

EXECJCL

This program allows the comparison of an experimental set of fragment lengths with those present in the Master Table. It is an interactive program, as can be seen from the example of its use shown in Figure 1. In this example the fragment lengths are those of the HpaII fragments of ØX174 Rf DNA. The largest fragment is 2748 base pairs in length and an error of ±10% was estimated. All sequences generating fragments between 2473 and 3023 base pairs (bp) in length are retrieved from the Master Table and stored (281 possibilities). The second largest fragment (1690 bp with an estimated error of ±5%) is then entered and the process repeated, giving 234 possibilities. By comparison of these two sets, the sequences common to both sets (31 possiblities) are selected and form a new set (the first intersect). The third fragment length (374 bp) and

FIGURE 1

```
PLEASE KEY IN LENGTH AND PERCENT SEPARATED BY A BLANK
>2748 0.1
SET TO BE SEARCHED RANGES FROM 2473 TO 3023
IF YOU WISH TO OMIT THIS SET OF LIMITS TYPE IN OMIT
>NO
THE NUMBER OF OCCURRENCES IS   281
DO YOU WISH TO SEE THE BASES CHOSEN..ANS YES OR NO
>NO
PLEASE KEY IN LENGTH AND PERCENT SEPARATED BY A BLANK
>1690 0.05
SET TO BE SEARCHED RANGES FROM 1605 TO 1774
IF YOU WISH TO OMIT THIS SET OF LIMITS TYPE IN OMIT
>NO
THE NUMBER OF OCCURRENCES IS   234
DO YOU WISH TO SEE THE BASES CHOSEN..ANS YES OR NO
>NO
THE NUMBER OF BASES IN THE INTERSECTION IS   31
DO YOU WISH TO SEE THE BASES CHOSEN..ANS YES OR NO
>NO
DO YOU WISH TO STOP THIS ITERATION..ANS YES OR NO
>NO
PLEASE KEY IN LENGTH AND PERCENT SEPARATED BY A BLANK
>374 0.05
SET TO BE SEARCHED RANGES FROM  355 TO  393
IF YOU WISH TO OMIT THIS SET OF LIMITS TYPE IN OMIT
>NO
THE NUMBER OF OCCURRENCES IS   341
DO YOU WISH TO SEE THE BASES CHOSEN..ANS YES OR NO
>NO
THE NUMBER OF BASES IN THE INTERSECTION IS    5
DO YOU WISH TO SEE THE BASES CHOSEN..ANS YES OR NO
>NO
DO YOU WISH TO STOP THIS ITERATION..ANS YES OR NO
>NO
PLEASE KEY IN LENGTH AND PERCENT SEPARATED BY A BLANK
>348 0.05
SET TO BE SEARCHED RANGES FROM  330 TO  365
IF YOU WISH TO OMIT THIS SET OF LIMITS TYPE IN OMIT
>NO
THE NUMBER OF BASES IN THE INTERSECTION IS    1
DO YOU WISH TO SEE THE BASES CHOSEN..ANS YES OR NO
>NO
DO YOU WISH TO STOP THIS ITERATION..ANS YES OR NO
>NO
PLEASE KEY IN LENGTH AND PERCENT SEPARATED BY A BLANK
>218 0.05
SET TO BE SEARCHED RANGES FROM  207 TO  229
IF YOU WISH TO OMIT THIS SET OF LIMITS TYPE IN OMIT
>NO
THE NUMBER OF OCCURRENCES IS   298
DO YOU WISH TO SEE THE BASES CHOSEN..ANS YES OR NO
>NO
THE NUMBER OF BASES IN THE INTERSECTION IS    1
DO YOU WISH TO SEE THE BASES CHOSEN..ANS YES OR NO
>NO
DO YOU WISH TO STOP THIS ITERATION..ANS YES OR NO
>YES
  CCGG
 DO YOU WISH TO RESTART ENTIRE PROCEDURE YES OR NO
>NO
```

Step #1

1st fragment length as input

Step #2

2nd fragment length as input

Step #3

3rd fragment length as input

Step #4

4th fragment length as input

Step #5

5th fragment length as input

PREDICTED RECOGNITION SITE

Figure 1: An example of the use of the program EXECJCL. The theoretical lengths of the HpaII fragments ØX174Rf DNA were used as input. Errors of ±10% were assumed for the largest fragment and ±5% for the remaining fragments.

error (±5%) are now entered and the possible sequences which could have generated such a fragment are retrieved (361 possibilities). Comparison of these possibilities with the sequences present in the first intersect is

performed and the common sequences (5 possibilities) retained to form a new set (the second intersect). The process is repeated until all input fragments have been used. From the example shown in Figure 1, it can be seen that a unique answer is found at step 4 and is retained until the end of the program.

One additional feature of this program appears when a digest contains two fragments of similar length, such that each lies within the error limits of the other. In this case, the program responds by retaining only those sequences which occur twice within that portion of the Master Table defined by the maximum length of the larger fragment and the minimum length of the smaller fragment.

MONITOR

This program allows access to the various programs described in this paper and also allows access to the various DNA sequences upon which these programs can operate. In addition, it allows the construction of new Master Tables for any sequences stored in the computer.

RESULTS

Strategy

As can be seen from Table 1, the majority of the known restriction enzymes recognize sites which consist of a linear array of four, five, or six nucleotides. Consequently, our initial goal was to develop a strategy whereby a table (the Master Table) could be constructed by the computer which would contain a list of all possible fragment lengths produced from ØX174 DNA by an enzyme recognizing any one of the 2728 possible unique combinations of 4, 5, or 6 nucleotides arranged as a linear string. For any given fragment length the computer also listed all possible sequences able to generate such a fragment. The program EXECJCL then searched this list, using a fragment length as input data, and retrieved all sequences able to generate such a fragment. This process was repeated when a second fragment length was introduced, but now only those sequences common to both steps were saved. Upon entering more fragment lengths, the number of possible sequences able to generate the complete set diminishes until either (1) a unique answer is obtained, (2) no possibility remains, or (3) no more input data is available.

Restriction Endonucleases with Known Recognition Sites: Practical Aspects.

Fragments generated by restriction enzyme digestion of ØX174 DNA

Table 1 Sequence Patterns Recognized by Restriction Endonucleases

Pattern	Example	Sequence	Total No. of Examples from Ref. 2.
NNNN	HpaII	CCGG	8
NNNNN	HphI	GGTGA	3
NNNNNN	EcoRI	GAATTC	16
NNXNN	HinfI	GAXTC	4
NNPyPuNN	HindII	GTPyPuAC	1
NPyNNPuNN	AvaI	CPyCGPuG	1
PuNNNNPy	HaeII	PuGCGCPy	1
NNAcGtNN	AccI	GTAcGtAC	1
AtNNNNAt	HaeI	AtGGCCAt	1

In this table the following abbreviations are used:

N = any one of the four deoxyribonucleotides, but with a specific value assigned for any given restriction enzyme.

X = any one of the four deoxyribonucleotides and no specific value is necessary.

Pu = A or G can be present at this point in the sequence.

Py = C or T can be present at this point in the sequence.

Ac = A or C can be present at this point in the sequence.

Gt = G or T can be present at this point in the sequence.

At = A or T can be present at this point in the sequence.

were resolved by electrophoresis on agarose gels and the length of each fragment was determined from its mobility by comparison with a set of standards. The TaqI fragments of ØX174 DNA (Figure 2) were arbitrarily used for this purpose.

The estimated fragment sizes from each restriction enzyme digest were used to drive the EXECJCL program. Since determination of fragment lengths by gel electrophoresis is fraught with error, a range of uniform error limits were applied to each of the experimentally-determined lengths. Four enzymes (HhaI, HpaI, HpaII, and MboII) with known recognition sequences provided the initial test and the results generated by successive steps of the program for each of the four digests are presented in Table 2.

The program was able to arrive at a unique recognition site for three

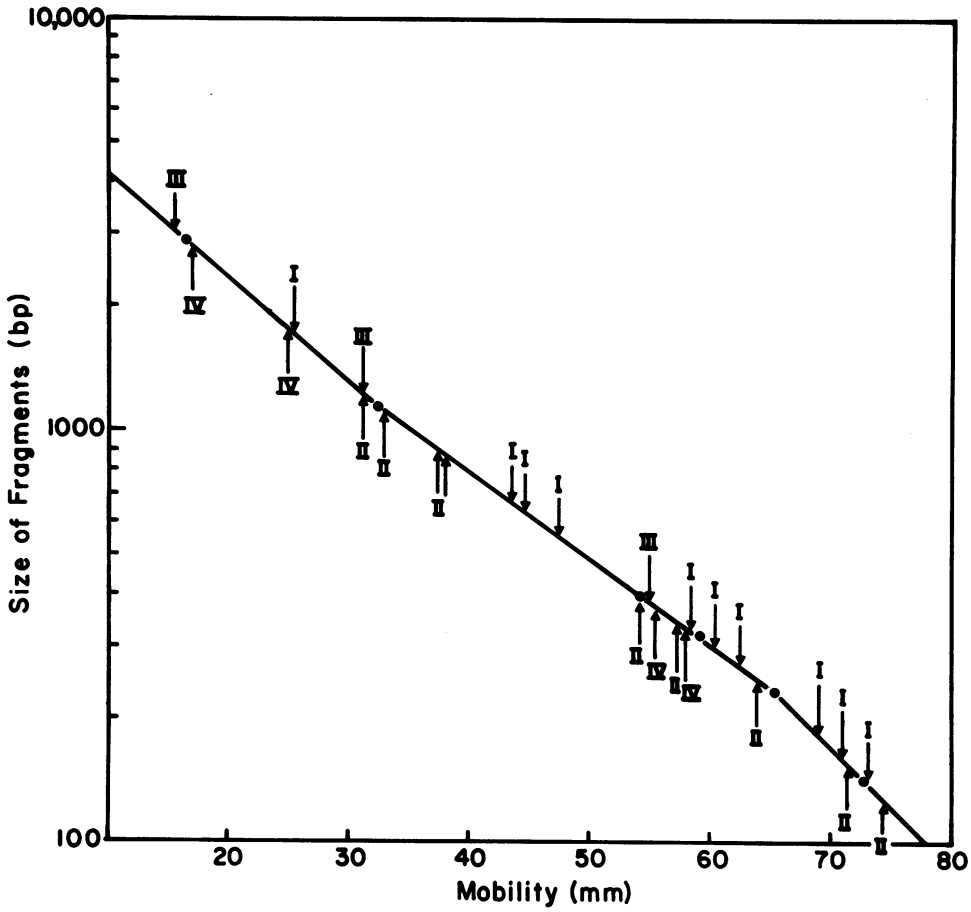

Figure 2: Calibration curve based upon the TaqI fragments of ØX174 Rf DNA fractionated on a 1.4% agarose gel. The sizes of all the fragments from HhaI (I), MboII (II), HpaI (III), and HpaII (IV) digests of ØX174 DNA were derived from this curve.

(HpaI, HpaII, and MboII) of the four enzymes used. For the fourth enzyme, HpaI, the error (21.1%) in determining the size of the largest fragment length prevented the correct sequence from appearing among the possibilities at step 1, even with the largest error limit chosen. Several important principles emerged from this study. The first, and most obvious point, is that when the actual length of the fragment lies outside of the range defined by the experimentally-determined length plus its associated

Table 2: Operation of the EXECJCL Program Using Uniform Error Limits

		Experimental Lengths	Lengths from sequence	Number of Sequence Possibilities 1% error	2%	4%	8%	10%	20%
HhaI (GCGC)									
	1	1700	1553	61	102	199	363	458	809
	2	670	640	2	5	18	62	103	335
	3	620	614	0	0	3	17	31	153
	4	530	532			0	12	24	79
	5	325	308				1	6	43
	6	295	289				0	2	26
	7	270	269				0	1*	15
	8	185	201					1*	12
	9	160	192					1*	8
	10	140	145					1*	5*
HpaII (CCGG)									
	1	2800	2748	31	61	119	238	278	583
	2	1775	1690	1	2	7	44	55	246
	3	375	374	0	0	0	9	11	66
	4	335	348	0	0	0	3	4	19
	5	240	218	0	0	0	0	1*	16*
HpaI (GTTAAC)									
	1	3055	3722	25	64	133	249	306	579
	2	1250	1264	0	4	11	28	41	171
	3	385	392	0	0	1	4	6	52
MboII (GAAGA)									
	1	1225	1103	57	103	196	391	487	859
	2	1125	1064	0	1	14	72	102	299
	3	900	860		0	2	14	28	141
	4	890	801			0	1	3	45
	5	400	396(2x)			0	0	2	25
	6	345	324				1	1*	10
	7	245	224				0	1*	4*
	8	150	118					0	2
	9	125	89					0	0

*This number represents or contains within its members the correct recognition sequence.

errors (as was the case for the HpaI digest), the program must necessarily fail since at this entry the correct answer will be discarded. Consequently, when very large or very small fragments are present in the digest, some precautionary steps must be taken. The simplest is to use a very large error limit at this step in order to ensure that the correct recognition sequence falls within the range chosen. Since the number of possible sequences which can generate a particular length of fragment increases dramatically as the range of possible fragment length increases, it is advisable to enter such fragments towards the end of the input data, by which time the set of possible sequences has already decreased to some manageable value. It can be seen from the calibration curve of Figure 2 that fragments larger than 2000 nucleotides or smaller than 100 nucleotides in length are most prone to experimental error. Table 2 also illustrates the three possible outcomes of the program:

(1) When the actual fragment length lies within the error limits assigned to each experimentally determined fragment length, the program can yield a unique answer. This may be either at the last step of the program (e.g., HpaII, 10% error), or at a previous step (e.g., HhaI, 10% error). In the latter case, the retention of this unique answer through subsequent steps serves to verify its accuracy.

(2) A unique answer may be generated during the running of the program but then disappear as further fragment lengths are entered. This arises when the actual length of one of the fragments lies outside of the range determined experimentally (MboII, 10% error). This is best dealt with by increasing the error limits on fragments entered after a unique solution has been found. (3) The program may run out of experimental fragment lengths without generating a unique solution (e.g., HpaII, 20% error). This situation occurs most frequently when relatively few fragments are produced and further experiments may be necessary to distinguish the possibilities. Occasionally some of the possibilities can be discarded by consideration of the total number of fragments produced by each of the sequences present in the final set. If this exceeds the number observed experimentally, then that solution must be incorrect. In general, a more useful strategy is to use the program SEARCHFOR to locate the map positions of the predicted sequences. A simple mapping experiment can then be designed to distinguish the possibilities.

Based upon these initial results, an improved strategy for the operation of the EXECJCL program was derived. Moderate error limits

(i.e., between 5-10%) are used for fragments in the size range 100 to 2000 nucleotides long and these are entered into the program first. The large and small fragments may then be assigned high errors and used to complete the input. The effectiveness of this approach is illustrated in Table 3 using the same raw data as was employed for Table 2.

The ability of the EXECJCL program to distinguish among the 2728 possible sequences present in the Master Table is illustrated in Figure 3. As expected, the rate with which a unique sequence is derived decreases as the error limits increase (i.e., as the accuracy of the experimental lengths diminishes). However, even with uniform errors as large as 10% or 20%, unique solutions can still be reached. In particular, with a 10% size error, the number of possible solutions reaches a manageable point very quickly, while a unique solution requires only 7 of the 18 fragments. It should be noted that, in contrast to the experimental data used to compile Table 3, Figure 3 used the known fragment lengths. The number of possible solutions at each stage of the program is different in the two cases reflecting the altered range being searched in the each case. This leads to the interesting observation that although it is experimentally desirable to estimate the lengths as accurately as possible, it is by no means a prerequisite for the correct functioning of this program.

Restriction Endonucleases with known Recognition Sites: Theoretical Aspects.

Although many of the restriction enzymes available at the moment give patterns from which the total number of fragments and their lengths can be determined with some accuracy, this is not always the case. Contaminating non-specific nucleases can sometimes lead to the degradation of fragments and cause the loss of bands from a digest, while low enzyme concentrations can lead to partial digestion. Thus, for many enzymes it is difficult to obtain a complete digest from which unambiguous assignment of fragment number and length can be determined. Many of the uncharacterized enzymes have remained so for precisely these reasons and it was of some interest to ask whether putative recognition sequences could be predicted from incomplete digests by the use of the EXECJCL program. We have addressed this problem by using known fragment lengths derived from an HpaII digest of ØX174.

The first experiment showed the effect of varying the order with which fragment lengths are provided to the program. The results are shown in Table 4. Using a 5% error limit it can be seen that no matter in

Table 3

Predictions from EXECJCL using Variable Error Limits[1]

		Experimental Lengths	Lengths from Sequence	Number of Sequence Possibilities (%) Error Limits: 25	20	10	5
HpaII							
	1	2800	2748		583		
	2	1775	1690		-	129	
	3	375	374		-	28	
	4	335	348		-	6	
	5	240	218		-	1 = CCGG	
HpaI							
	1	3055	3722	712	-		
	2	1250	1264		-	-	62
	3	385	392		-	-	8 contains GTTAAC
MboII							
	1	1225	1103		-	487	
	2	1125	1064		-	102	
	3	900	860		-	28	
	4	890	801		-	3	
	5	400	396(2x)		-	2	
	6	345	324		-	1	
	7	245	224		-	1	
	8	150	118		1		
	9	125	89		1 = GAAGA		

[1]The HhaI digest is not shown because there was no difficulty in arriving at a distinct prediction.

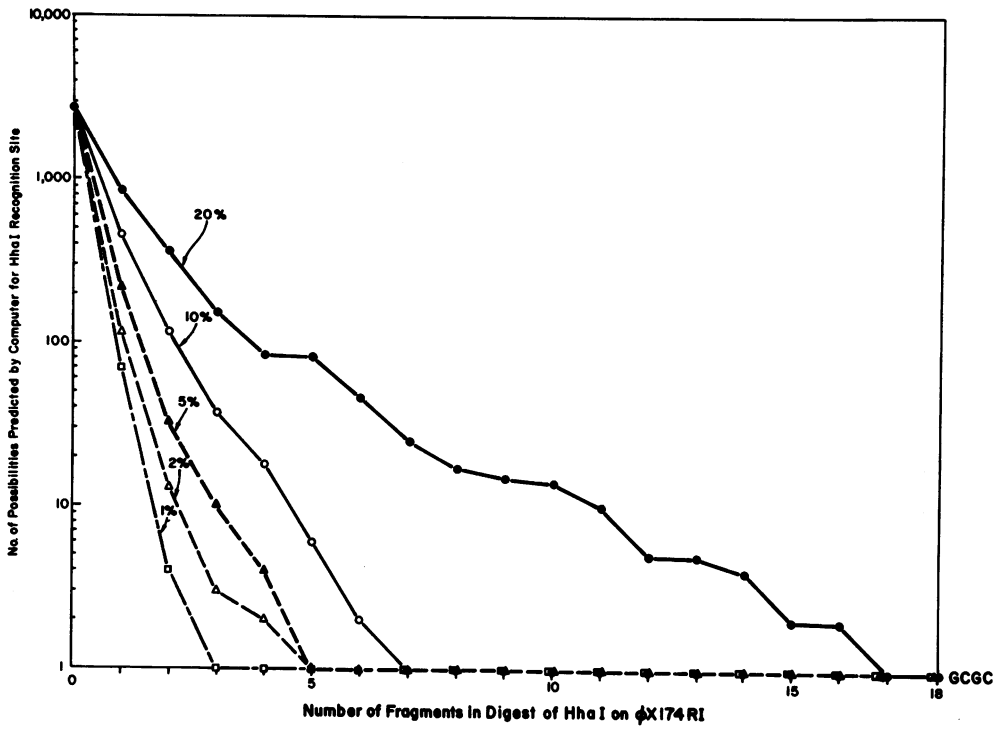

Figure 3: The effectiveness of the EXECJCL program. These plots reflect the ability of the program to retrieve the recognition site for the enzyme HhaI from among the 2728 possible sequences in BASES. The known lengths of the 18 HhaI fragments of ØX174 DNA were used as input data with error limits as indicated above each curve. In each case the fragment lengths were entered in decreasing order of size.

what order the fragments are entered into the program, a unique solution can always be generated with 4 of the 5 fragments and that in case 2, a unique solution is generated after only 3 fragments are entered. It is also clear from this example that the largest fragment present in the digest has the most dramatic effect in reducing the number of possibilities. Unfortunately, a fragment of this length is also subject to the greatest possibility of error in its length determination, since it lies outside the linear range of the mobility curve. As expected, the order with which the fragments are provided to the program has no effect upon the total number of possibilities finally predicted; however, the rate with which those

Table 4

The Effect of Varying the Order of Fragment Lengths

Provided to EXECJCL

5% error limit on all length values[1]

Case 1		Case 2		Case 3	
Fragment Length	Possibilities	Fragment Length	Possibilities	Fragment Length	Possibilities
2748	131	218	410	378	436
1690	16	2748	8	1690	39
378	2	378	1*	2748	2
348	1*	1690	1*	348	1*
218	1*	348	1*	218	1*

10% error limit on all length values

Case 1		Case 2		Case 3	
Fragment Length	Possibilities	Fragment Length	Possibilities	Fragment Length	Possibilities
2748	326	278	779	218	804
1690	54	218	229	2748	30
378	12	1690	37	378	4
348	7	348	27	1690	2
218	2*	2748	2*	348	2*

[1]These fragment lengths are from HpaII digest of ØX174 Rf DNA.

*This number represents or contains within its members the correct recognition sequence.

possibilities are reduced can be affected significantly. Using a 10% error limit, the effect of the largest fragment is most clearly seen in case 3 where its introduction at step 2 reduces the possibilities to 30, while the third fragment further reduces the possibilities to 4. Such a small number could be distinguished fairly rapidly by experimental mapping.

A second experiment consisted of systematically omitting one or more fragments from the input data, and the results are shown in Table 5. The omission of a single fragment from the HpaII digest gives a unique answer in all cases except that in which the largest fragment is omitted. In that case, 4 possibilities remained with sequences CCGG, ACTCA, ATGTC, and AATGTC. Examination of the total number of fragments in digests corresponding to each of these recognition sequences showed that only CCGG could be the recognition site because in the three other cases, more fragments were predicted than occurred in the digest but, in particular, a large fragment of length around 2700 bp was missing. The experiment again showed the relative value of a large fragment in reducing the number of possibilities. By omitting more fragments from the digest, it was still possible to obtain a unique and correct answer in certain cases, although the most useful information to emerge from this experiment was that a manageable number of possibilities can be generated with relatively little information available and from which the correct recognition sequence might be deduced by further experimentation.

One further experiment was carried out to determine the accuracy with which fragment lengths should be known in order that correct recognition sites be deduced. Using computer-determined fragment lengths, the error limits were systematically increased until a point was reached such that a unique solution was still generated by the computer but that by increasing the error, more than one possibility remained. The results for several enzymes of known sequence which cleave ØX174 DNA are shown in Table 6. It can be seen that with the exception of EcoRII, errors in the range from 15% to 34% still lead to the correct deduction of the recognition sequence. The rather low error needed for EcoRII fragments reflects the fact that for this enzyme only two cleavage sites exist in ØX174 DNA and 413 sequences occur only twice in the ØX174 genome.

Restriction Endonucleases with Unknown Recognition Sites.

Biochemical analysis of the recognition sequences for the endonucleases BbvI from Bacillus brevis and SfaNI from Streptococcus

Table 5 The Effect of Omitting One or More Fragments from Digest Patterns

	Fragment Length	Omission of One Fragment					
		0	1	2	3	4	5
1.	2748	131	-	131	131	131	131
2.	1690	16	274	-	16	16	16
3.	378	2	39	10	-	2	2
4.	348	1	12	5	1	-	1
5.	218	1	4†	1	1	1	-

		Omission of Two Fragments			
		1,2	2,3	2,4	2,5
1.	2748	-	131	131	131
2.	1690	-	-	-	-
3.	378	350	-	10	10
4.	348	124	7	-	5
5.	218	46	1	1	-

		Omission of Three Fragments			
		2,3,4	3,4,5	2,4,5	2,3,5
1.	2748	131	131	131	131
2.	1690	-	16	-	-
3.	378	-	-	10	-
4.	348	-	-	-	7
5.	218	8	-	-	-

For this experiment the known fragment lengths generated by HpaII were used as input and a uniform error of 5% applied. The numbers above the columns indicate the fragment(s) omitted.

†Sites selected were CCGG, ACTCA, ATGTC, and AATGTC. Of these, only CCGG could be the recogniton site based on total number of fragments in digest.

°CCGG is the predicted site.

Table 6 Tolerable errors in fragment length determination.

	Enzyme	Site	Number of Fragments	Maximum % Error Allowing A Unique Prediction
1.	AluI	AGCT	24	23%
2.	BbvI	GC(A/T)GC	14	21%
3.	EcoRII	CC(A/T)GG	2	0.2%
4.	HaeIII	GGCC	11	16%
5.	HgaI	GACGC	14	15%
6.	HpaII	CCGG	5	17%
7.	HhaI	GCGC	18	>25%
8.	Hin1056I	CGCG	14	18%
9.	HphI	GGTGA	9	18%
10.	MboII	GAAGA	11	19%
11.	MnlI	CCTC	35	>30%
12.	TaqI	TCGA	10	34%

faecalis has proved somewhat difficult because of the persistent presence of non-specific nucleases. These enzymes were therefore chosen as an appropriate test of the effectiveness of EXECJCL program. Figure 4 shows the digestion profile of BbvI on ØX174 Rf DNA and the fragment lengths listed next to each gel band were provided to the EXECJCL program. Error limits of 5, 10 or 15% were chosen and the results arising from the computer are shown in Table 7. A unique sequence, 5' GCAGC 3' is predicted for the recognition site.

A similar experiment using the fragments resulting from the SfaNI digest of ØX174 Rf DNA led to a predicted recognition sequence 5' GATGC 3' for SfaNI. Initial mapping of some of these sites within ØX174 DNA gave locations consistent with this prediction (Sciaky and Roberts, 1978, to be published). It is of some interest to note that this pentanucleotide sequence had also been derived manually by searching for the ØX174 sequence for similarities in the region of SfaNI sites and required many hours of laborious effort in order to reach it.

Two other restriction enzymes of previously unknown recognition sequence have also been studied using computer methods. The first of these is PvuI from Proteus vulgaris (Gingeras and Roberts, to be published) which fails to cleave ØX174 DNA, SV40, G4, or fd DNAs. It

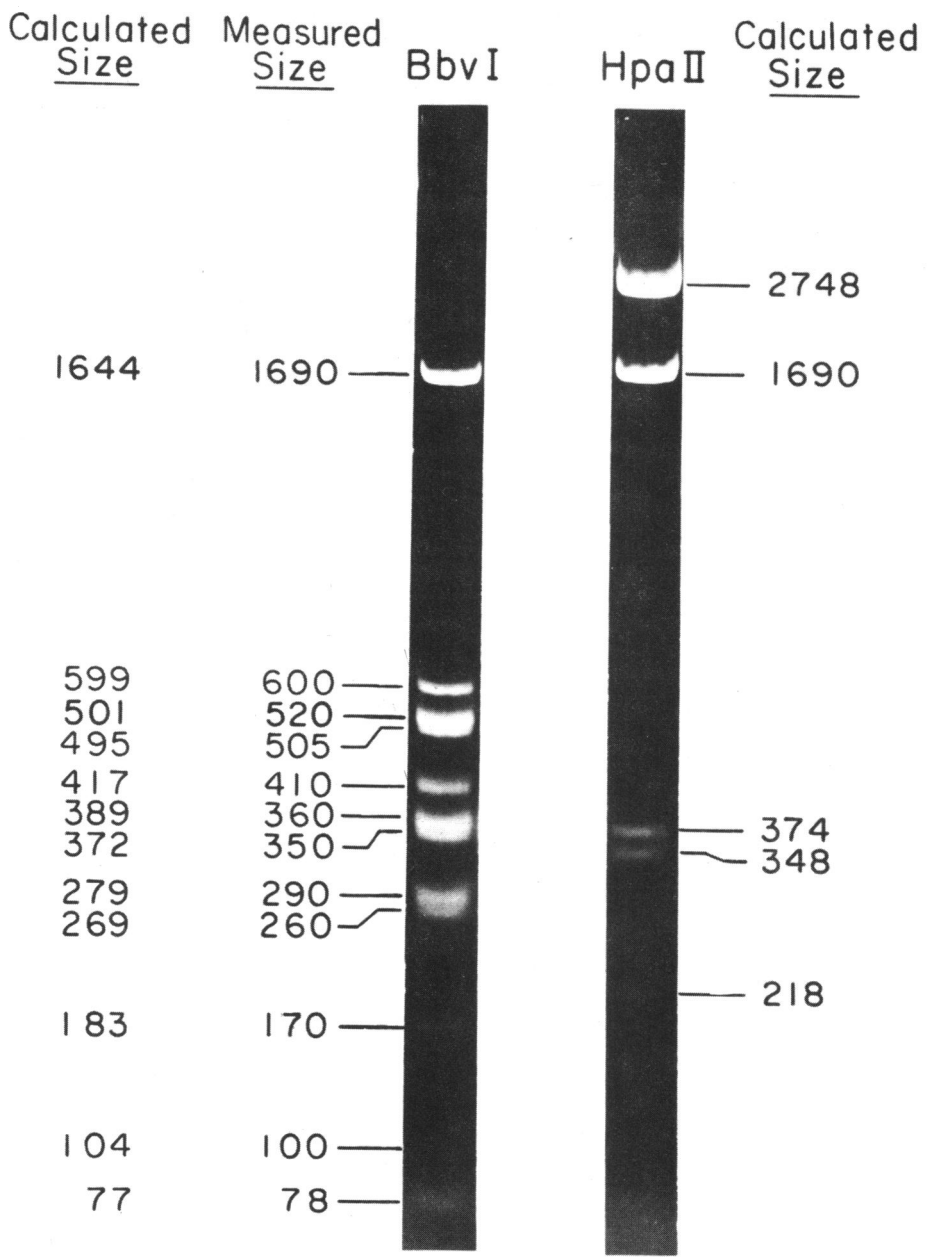

Figure 4: BbvI on ØX174 Rf DNA fractionated on a 2% agarose gel. ØX174 are also displayed and were used as size markers. The experimentally determined fragment lengths are listed alongside the BbvI digest. The theoretical lengths for both digests are also shown.

Table 7

The use of EXECJCL Program to Determine the Recognition Site of BbvI.

Experimental Lengths (bp)	Number of Sequence Possibilities		
	5% error	10%	15%
1690	234	449	640
600	27	204	207
520	4	37	99
505	4	14	34
410	2	8	19
360	1*	4	19
350	0	1*	4
290		1*	3
260		1*	2
170		1*	2
100		1*	2
78		1*	1*

* is the sequence 5' GCAGC 3'

does, however, cleave the plasmid pBR322 at one site. Clearly, this enzyme is not a candidate for the EXECJCL program. However, an alternative approach to seek a possible recognition site was available. From the low frequency of cleavage, it seemed likely that the site should be a hexanucleotide and furthermore was most probably a palindrome. By searching the sequences of these 5 DNAs, a unique hexanucleotide palindrome, 5' CGATCG 3', was found which occurred only once within the pBR322 sequence and failed to occur in the sequences of the other 4 DNAs. We therefore predict that this will indeed prove to be the recognition site for PvuI. Additional support for this conclusion is derived from the finding that a PvuI site occurs within a segment of

Xenopus laevis DNA whose sequence is known (B. Sollner-Webb, personal communication). The region of this DNA which contains the PvuI site has been mappped and does indeed contain the sequence 5' CGATCG 3'.

A similar approach has been used for the enzyme MstI from a strain of Microcoleus (I. Schildkraut and D. Comb, personal communication). In this case, digestion showed that no site existed in SV40 DNA--one site occurred in ØX174 DNA and two sites occurred in the G4 genome. Again, from the few number of sites observed on these substrates, a hexanucleotide palindrome is the most likely recognition sequence and the computer search for such palindromes within these DNAs gave only one sequence, 5' TGCGCA 3', as a likely candidate. Based upon this prediction, pBR322 DNA should contain 4 sites. Subsequent digestion showed that this was indeed the case. In addition, the location of the single site of ØX174 DNA has been shown to occur extremely close to the single XhoI recognition site precisely at the point at which the sequence 5' TGCGCA 3' is located (I. Schildkraut and D. Comb, personal communication).

DISCUSSION

The program EXECJCL described in this paper is designed to predict restriction enzyme recognition sites by considering only the length of the fragments produced upon digestion of a DNA of known sequence with a restriction endonuclease. Its scope is presently limited to only those sequences which are linear arrays of 4, 5, or 6 nucleotides. We are presently extending this to cover the other families of sequences already shown to be recognized by restriction endonucleases (see Table 1). Using this program to study both enzymes of known and unknown specificity, its predictive power seems good. Accurate fragment lengths are not required, and experimental digests with poor resolution of bands or missing bands can still lead to useful predictions. Several practical points have emerged regarding the most effective utilization of the program. (1) Because the number of sequences which generate fragments of sizes greater than 2000 nucleotides are relatively small, an accurate estimate of these larger fragment sizes considerably reduces the number of possible sequences which may be recognized. This is of limited practical value because such fragments have the greatest risk of error in size determination and so fairly wide error limits must be applied when entering these fragments into the computer. (2) The order with which fragment lengths are provided to the program, although not affecting the total number of final

possibilities, has a marked effect upon the rate at which a unique prediction is recovered (3). Restriction enzyme digests which contain partial products or missing fragments can still be used by the program because, in general, not all fragments are needed in order to arrive at either a unique answer or a small set of possibilities.

For two restriction endonucleases whose specificity was previously unknown, the program has led to predictions for their recognition sequences. In the case of BbvI from Bacillus brevis, that sequence is 5' GCAGC 3' and is likely correct as a DNA methylase which recognizes this same sequence has been previously isolated from another strain of Bacillus brevis (13). The enzyme SfaNI from Streptococcus faecalis is predicted to recognize the sequence 5' GATGC 3' and, again, by mapping some of these sites within the ØX174 genome, the map positions are consistent with this prediction. Clearly, this particular program is ideally suited for enzymes that cleave ØX174 DNA at many sites. However, it is less suitable for enzymes which cleave ØX174 DNA at only 1 or 2 sites since 435 sequences occur only once upon the ØX174 genome and 413 sequences occur twice within the ØX174 genome. Nevertheless, it is still possible to use a computer approach for the determination of such sequences by taking advantage of the fact that complete sequences for SV40 (14, 15), G4 (16), fd (17), and pBR322 (18) are now available. In fact, the programs outlined above are now available to search these sequences through the use of the MONITOR program. Therefore, a restriction enzyme of unknown specificity need cut only one of these 5 substrates in order for the recognition sequence to become accessible to computer oriented methods.

By using the SEARCHFOR program, recognition sequences have also been predicted for the enzymes PvuI from and MstI. These sequences were deduced by a combination of manual and computer-assisted methods and a future goal of our work will be to fully automate this procedure.

We would not wish to suggest that the use of this program will supercede biochemical methods for restriction enzyme recognition site determination, but rather look upon it as a means of providing a hypothetical recognition site, which can be tested by suitable biochemical experiments. These may take the form of sequence determination by standard procedures (2) or by the newer methods used to determine the recognition sequence of PstI (19). Even when the program is unable to generate a unique solution, but rather can only predict a small number of

possibilities, it is often possible to distinguish among them by stochastic means. From the propensity of restriction enzymes to recognize palindromic sequences, the presence of a palindrome within the final list of possibilities makes it an extremely likely candidate. However, other candidates with unusual patterns should not be disregarded since it seems unlikely that the number of possible recognition patterns for these enzymes has been exhausted.

Programs which search for particular sequence features have been described by Staden (20, 21), and programs able to predict secondary structure have been described by Korn (22). In addition, we have recently learned of a program similar in essence to the one described here, which may also be used to predict restriction enzyme recognition sites (23). It is clear that the use of the computer for analyzing nucleic acid sequences is still in its infancy. As more information becomes available, it will become an essential element in data analysis. Already, complete sequences for SV40, ØX174, fd, and G4 are available and are surely only the beginning of a wave of sequence data that threatens to dwarf the most resourceful memory.

Copies of the programs are available from the first author.

ACKNOWLEDGEMENTS

We thank D. Comb and I. Schildkrant for a gift of MstI. This research was supported by a grant to RJR (PCM76-82448) from the National Science Foundation. TRG was supported by a postdoctoral fellowship from the National Institutes of Health.

References

1. Roberts, R.J. (1976). CRC Critical Reviews in Biochemistry 4: 123-164.
2. Zabeau, M. and Roberts, R.J. (1978). Mol. Genetics, in press.
3. Kleid, D., Humayun, Z., Jeffrey, A. and Ptashne, A. (1976). Proc. Nat. Acad. Sci. USA 73: 293-297.
4. Endow, S.A. (1977). J. Mol. Biol. 114: 441-450.
5. Brown, N.L., Hutchison, C.A. III, and Smith, M. (1978). J. Mol. Biol., in press.
6. Brown, N.L. and Smith, M. (1977). Proc. Nat. Acad. Sci. USA 74: 3213-3216.
7. Roberts, R.J., Myers, P.A., Morrison, A., and Murray, K. (1976). J. Mol. Biol. 103: 199-208.
8. Sharp, P.A., Sugden, B., and Sambrook, J. (1973). Biochemistry 12: 3055-3063.
9. Gelinas, R.E., Myers, P.A., and Roberts, R.J. (1977). J. Mol. Biol. 114: 169-180.

10. Sato, S., Hutchison, C.A. III, and Harris, J.I. (1977). Proc. Nat. Acad. Sci. USA 74: 542-546.
11. Sugden, B., DeTroy, B., Roberts, R.J., and Sambrook, J. (1975). Anal. Biochem. 68: 36-46.
12. Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A. III, Slocombe, P.M., and Smith, M. (1977). Nature 265: 687-695.
13. Vanyushin, B.F. and Dobritsa, A.P. 1975). Biochem. Biophys. Acta 407: 61-72.
14. Reddy, V.B., Thimmappaya, R., Dhar, K.N., Subramanian, B., Zain, S., Pan, J., Ghosh, P.K., Celma, M.L., and Weissman, S.M. (1978). Science 200: 494-502.
15. Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G., and Ysaebaert, M. (1978). Nature 273: 113-120.
16. Godson, G.N., Barrell, B.G., Staden, R., and Fiddes, J.C. (1978). submitted to Nature.
17. Schaller, H., Beck, E., and Takanami, M. (1978) in Single-Stranded DNA Phages (D.T. Denhart, D.H. Dressler, D.S. Ray, eds.) Cold Spring Harbor Press, New York (in press).
18. Sutcliffe, G. (1978), unpublished results.
19. Brown, N.L. and Smith, M. (1976) FEBS Letters 65, 284-287.
20. Staden, R. (1977). Nuc. Acids Res. 4: 4037-4051.
21. Staden, R. (1977). Nuc. Acids Res. 5: 1013-1015.
22. Korn, L.J., Queen, C.L. and Wegman, M.W. (1977). Proc. Nat. Acad. Sci. USA 74: 4401-4405.
23. Fuchs, C., Rosenvold, E.C., Honigman, A. and Szybalski, W. (1978) Gene (in press).