Volume 15 Number 15 1987 Nucleic Acids Research

Restriction endonucleases for pulsed field mapping of bacterial genomes

Michael McClelland, Robert Jones*, Yogesh Patel and Michael Nelson

Department of Biochemistry and Molecular Biology, University of Chicago, 920 E. 58th St., Chicago, IL 60637, USA

Received May 5, 1987; Revised and Accepted July 6, 1987

ABSTRACT

Fundamental to many bacterial genome mapping strategies currently under development is the need to cleave the genome into a few large DNA fragments that can be resolved by pulsed field gel electrophoresis. Identification of endonucleases that infrequently cut a genome is of key importance in this process. We show that the tetranucleotide CTAG is extremely rare in most bacterial genomes with G+C contents above 45%. As a consequence, most of the sixteen bacterial genomes we have tested are cleaved less than once every 100,000 base pairs by one or more endonucleases that have CTAG in their recognition sequences: Xba I (TCTAGA), Spe I (ACTAGT), Avr II (CCTAGG) and Nhe I (GCTAGC). Similarly, CCG and CGG are the rarest trinucleotides in many genomes with G+C content of less than 45%. Thus, Sma I (CCCGGG), Rsr II (CGGWCCG), Nae I (GCCGGC) and Sac II (CCGCGG) are often suitable endonucleases for producing fragments that average over 100,000 base pairs from such genomes. Pulsed field gel electrophoresis of the fragments that result from cleavage with endonucleases that cleave only a few times per genome should assist in the physical mapping of many prokaryotic genomes.

INTRODUCTION

The genomes of bacterial species vary widely in their base composition. As a consequence, the distribution of fragment sizes produced by restriction endonuclease digestion of these genomes can be expected to vary considerably. Furthermore, higher order asymmetry in DNA base composition (di- and trinucleotide frequencies, for example) are useful for predicting cleavage frequencies in bacterial genomes. We have examined mono-, di- and trinucleotide frequencies in available prokaryotic DNA sequences and used them to predict *rare* restriction sites in these genomes that occur less than once every 100,000 base pairs. Given that such genomes are about one to six million bases in length, less than sixty fragments should result.

To test these predictions we performed restriction digestion on DNA

prepared from eleven bacterial species across a range of genomic G+C contents from 28% to 75% G+C. The digest patterns were examined using conventional electrophoresis in 1% agarose gels. Selected rare cutters were then re-examined by pulsed field gel electrophoresis, using the vertical high resolution separation apparatus described by Gardiner et al. (1,2).

We demonstrate that restriction enzyme suitable for genomic mapping can be selected on the basis of their recognition sequences in the context of the DNA sequence of the prokaryote genome to be mapped.

METHODS

Genomic DNA Preparation

(a) DNAs from *Fusobacterium nucleatum* 4H (Michael Smith), *Staphylococcus aureus* PS 96 (J.S. Sussenbach), *Haemophilus influenzae* Rd, *Streptococcus pneumoniae* 641 (S. Lacks), *Bacillus amyloliquefaciens* H (F.E. Young), *Xanthomonas holcicola* ATCC 13461 , *Xanthomonas malvacaerum* ATCC 9924, *Rhodobacter sphaeroides* (S.Kaplan), *Streptomyces albus* G (J.M. Ghuysen) and *Arthrobacter luteus* ATCC 21606 were kindly supplied by Geoffrey Wilson, Keith Lunnen, and Mary Ellen Looney of New England Biolabs. These DNAs were shown to be protected from digestion by restriction endonucleases derived from their own restriction systems: Fnu 4H I, Sau 3A I, Hind III, Mbo I+Sau 3A I, Bam H I, Xho I, Xma I, Rsa I, Sal I and Alu I, respectively. *Salmonella typhimurium* LT2 DNA was kindly provided by David Hillyard (University of Utah). All these DNAs were prepared by modifications of the Marmur procedure (3) and by cesium chloride-ethidium bromide centrifugation.

(b) High molecular weight chromosomal DNAs from *Moraxella bovis* ATCC 10900, *Staphylococcus aureus* NCTC 8325, *Staphylococcus aureus* PS 96, *Bacillus subtilis* JAS7, *Klebsiella pneumoniae* ATCC 13883, *Enterobacter aerogenes* ATCC 13048 and *Rhodobacter capsulatus* SB1003 were prepared by lysis *in situ* in low melting temperature agarose (4). Cells were harvested in late log phase or soon after reaching stationary phase. Generally, a 1.0 ml pellet was resuspended in 500 ul of 100 mM EDTA, 10 mM EGTA, 10 mM Tris pH 8.0 (EET Buffer) and placed at 37°C. 1.5 ml of 2% Low Melting Point agarose (SeaKem FMC) was added at 40°C. This was then dispensed into a number of 8 x 3 x 2 mm wells in a mold plate and allowed to solidify. The resulting blocks of agarose are referred to in this paper as 'inserts'. *Staphylococcus* and *Bacillus* cells were treated with 1 ug/ml penicillin for one hour at mid-log phase before harvesting. *Staphylococcus* cells were lysed *in situ* with 1 mg/ml lysostaphin and 1 mg/ml lysozyme at 30°C for 4 hours in EET. All 'inserts' of other cell types were treated with 200 ug/ml of

lysozyme and 0.05% sarkosyl for two hours, except *Rhodobacter* inserts, which did not require this treatment. Inserts were then placed in EET Buffer with 1 mg/ml Proteinase K and 1% SDS, (Lysis Buffer), which was preincubated at 65°C for 30 minutes. Proteinase digestion was performed overnight at 65°C. The inserts were dialyzed in 50 ml of 10 mM Tris pH 8.0, 1 mM EDTA (TE), 100 uM PMSF for one hour. The TE dialysis was repeated at least 2 times without PMSF and inserts were stored in TE.

We frequently found a considerable improvement in results if the inserts were electroeluted using pulsed field gel electrophoresis for 2 hours at a 4 second pulse time prior to endonuclease digestion. This removed any degraded DNA.

Restriction Digestions

(a) Conventional 1% agarose-ethidium bromide electrophoresis. 2 ug of bacterial DNA was digested in a total volume of 30 ul with a ten-fold excess of endonuclease under conditions recommended by the manufacturer (New England Biolabs, Beverly, MA). In some cases, reactions were monitored by the inclusion of 1 ug bacteriophage T7 or lambda DNA to ensure that total digestion had occurred. After digestion, reactions were terminated with 10 ul of stop solution (50 mM EDTA-50% glycerol-0.1% bromophenol blue/xylene cyanol), and samples were loaded onto a horizontal 1% agarose slab gel. Electrophoresis was carried out at 2-3 volts/cm for 10-16 hours.

(b) Pulsed field gel separations. Inserts containing 1 to 10 ug of DNA were placed in 1 ml of the appropriate restriction buffer. After 10 minutes 900 ul of buffer was removed and 10 to 100 units (1 to 10 ul) of restriction endonuclease added. Restriction digestion was performed with gentle shaking at the appropriate temperature for 8 to 16 hours. After one to three hours, an additional aliquot of 10 to 100 units of restriction enzyme was added. Short reaction times did not permit complete cleavage of DNA near the center of the insert. Digestion could be improved by first preparing DNA at high concentration in the inserts and slicing these into four pieces (3 x 2 x 2 mm) before digestion.

Pulsed Field Gel Electrophoresis

This electrophoresis system was originally developed by Schwartz and Cantor (2). We employed the configuration used by Gardiner et al. (1) and available from Beckman Instruments under the trade name GeneLine. This apparatus performed well in the 50,000 to 1,000,000 base pair range and produced straight lanes rather than the curved lanes of the original PFG method. Electrophoresis was performed for 15 to 20 hours in 1/4 X TAE. Pulse times were 4 seconds at 250V for 1 hour and then 4, 10, 12, 15, 30 or 60 seconds, depending on the need for separation in the 100, 200, 300, 400,

600 or 1,000 kilobase pair range, respectively. Gels could be stained with ethidium bromide, examined, and then run further with no apparent loss of quality.

As molecular weight markers, we employed polymerized bacteriophage lambda DNA ladders and *Saccharomyces cerevisae* chromosomes. Lambda ladders were produced by lysing bacteriophage lambda with Lysis Buffer in 1% agarose inserts at 40 ug/ml final DNA concentration. The DNA monomers anneal spontaneously at their 12 base cohesive ends within a few days at 4°C. We thank Mike Lane, (SUNY Syracuse) for this improved protocol and supplying us with inserts from his laboratory (5). Yeast DNA was prepared by the method of Schwartz and Cantor (2).

Southern Blots

A piece of Zetaprobe membrane (Biorad) was cut to cover the gel to within 5 mm of the edges. The membrane and the gel were both wetted in 1/4 X TAE. The gel was replaced in the gel maker. The damp membrane was applied to the surface of the damp gel with care taken to remove any air bubbles between the gel and the membrane and to remove any pools of buffer. 1% Agarose in 1/4 X TAE at 45°C was poured over the membrane to cover the whole gel. This sandwich was trimmed of agarose to within 2 mm from the edges and replaced in the electrophoresis tank (1). Electroblotting was performed for one hour using the electrode assembly supplied with this apparatus. The membrane was removed from the sandwich by peeling away the top layer of agarose (4).

The protocol of Reed and Mann was followed to attach the DNA to the membrane (6). The DNA was fixed and hybridized using ^{32}P labeled probes (7).

Statistical Methods

The expected frequency of restriction endonuclease recognition sequences in the DNA of bacterial species was determined using three predictors: mononucleotide, dinucleotide and trinucleotide frequencies, variously obtained from Ehrlich et al (8), Setlow (9), Bergey's Manual of Systematic Bacteriology (Eds. Krieg N.R. and Holt J.G., vol. 1 and 2, Williams and Wilkins, Baltimore/London, 1984), and from sequenced genes in GENBANK through the use of the NAQ program obtained from the National Biomedical Research Foundation.

Expected frequencies were calculated as Markov chains (10,11). The frequency of the sequence $N_1N_2N_3N_4N_5N_6$, (where N_1 thru N_6 are the bases in the recognition sequence), can be calculated from mononucleotide frequencies as: $p(N_1)\cdot p(N_2)\cdot p(N_3)\cdot p(N_4)\cdot p(N_5)\cdot p(N_6)$, where p(N) is the frequency of N in the genome.

From dinucleotide frequencies as:
$p(N_1N_2)\cdot p(N_2N_3)\cdot p(N_3N_4)\cdot p(N_4N_5)\cdot p(N_5N_6) / p(N_3)\cdot p(N_4)\cdot p(N_5)$.
From trinucleotide frequencies as:
$p(N_1N_2N_3)\cdot p(N_2N_3N_4)\cdot p(N_3N_4N_5)\cdot p(N_4N_5N_6)/p(N_2N_3)\cdot p(N_3N_4)\cdot p(N_4N_5)$.

A detailed application of this method to *E. coli* DNA has recently been published (12).

By adding together oligonucleotide frequencies from coding and noncoding strands in protein coding genes and from highly conserved tRNA and rRNA sequences, differences between individual functional regions have been obscured. Thus, it should be recognized that our method can systematically *overestimate* the calculated frequency of sequences containing these di- and trinucleotides. In particular, CTA is more likely to occur in the coding strand of a protein coding region than is TAG which is excluded from one frame. The overestimate of CTAG abundance using CTA and TAG frequencies as predictors is partly a consequence of this strand bias. The bias was partly obscured when we pooled trinucleotide frequency data from genes in the data base, of which a few were calculated for the noncoding orientation. We are presently partitioning the data set to allow a more precise prediction of restriction endonuclease cleavage frequencies.

RESULTS

The Effect of Base Composition

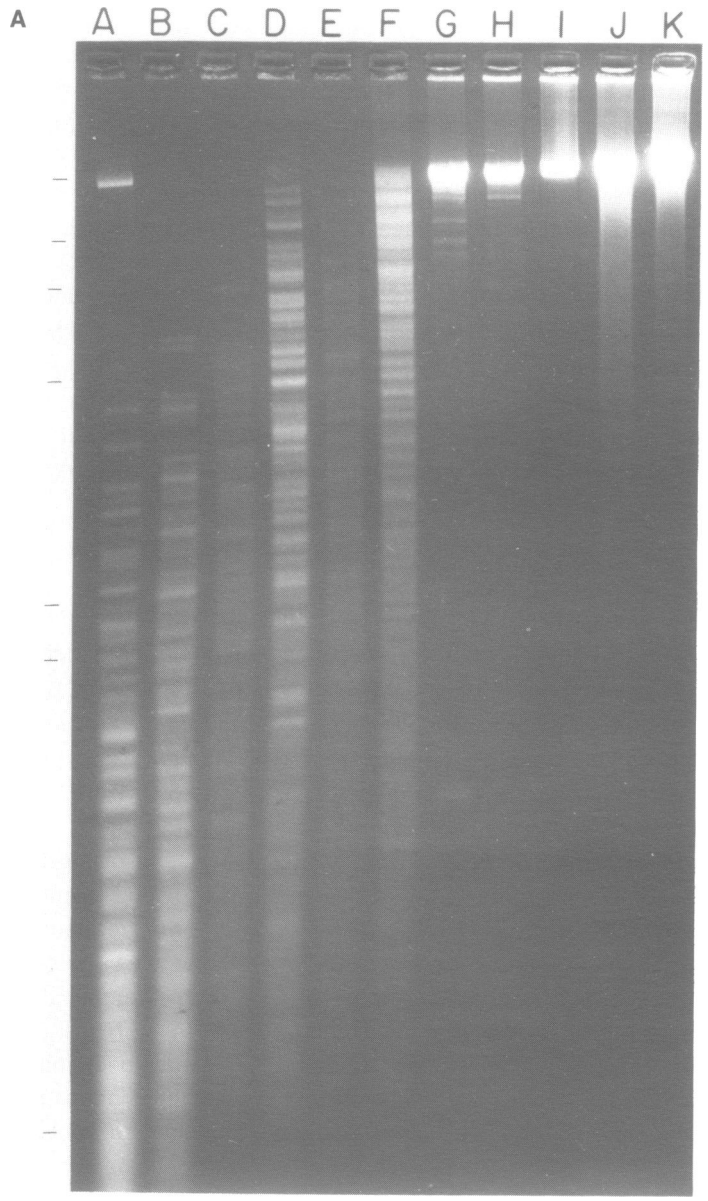
To determine the effect of base composition on the frequency of restriction endonuclease recognition sequences in bacterial genomes, we digested eleven DNAs with four endonucleases: Dra I (TTTAAA), Ssp I (AATATT), Not I (GCGGCCGC) and Sfi I (GGCCN5GGCC). The bacterial species were selected on the basis of their different G+C contents and phylogenetic diversity. The results of this experiment for Dra I and Not I are shown in Figure 1.

The observed number of fragments produced by each of these endonucleases was compared to the number expected from calculations using base composition as a predictor (Table I). In general, the frequency of cleavage calculated from the mononucleotide predictor was quite accurate. However, in some species one endonuclease gave *more* fragments than expected; for example, in the digestion of *Bacillus* DNA by Not I. In other species, there was *less* cutting than expected, for example in the digestion of *Rhodobacter* DNA by Dra I. Clearly, factors other than mononucleotide frequency influence the number of restriction endonuclease recognition

sequences within a bacterial genome. Furthermore, it is clear that certain endonucleases can be used to generate fragments larger than 10,000 base pairs in DNAs from appropriate species.

The Effects of Trinucleotide Frequency and Codon Usage.

Bacterial genomes contain a high proportion of DNA that encodes



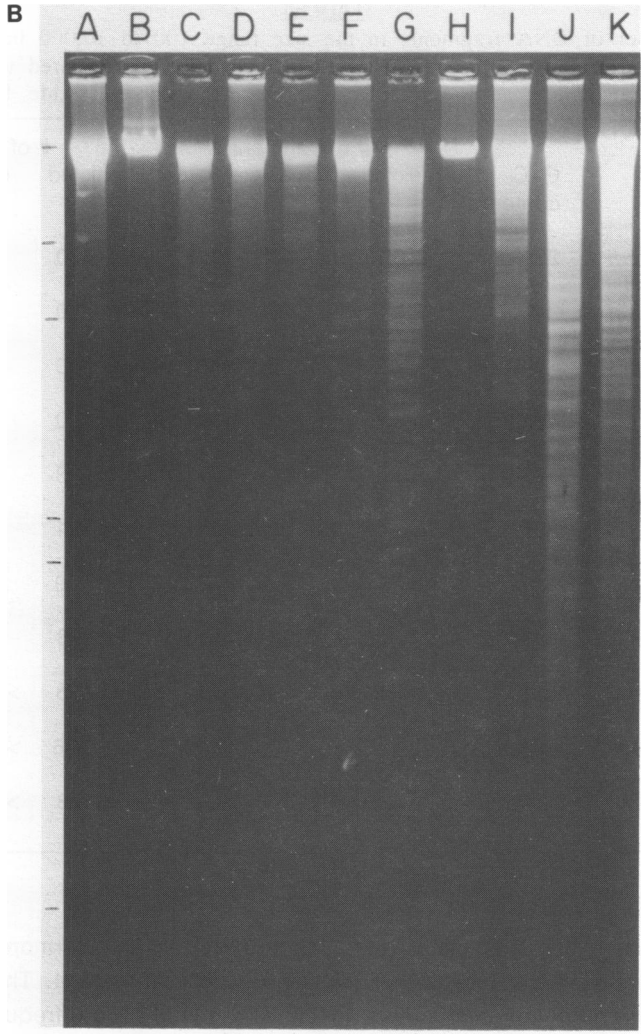


FIGURE 1 Effect of Mononucleotide Frequency on Restriction Cleavage.

1% Agarose TBE gel electrophoresis. 2 V/cm for 16 hours, 2 ug of genomic DNA per lane. DNAs are arranged left to right in ascending order of G+C content as in Table 1: Fnu 4H (28% G+C) (lane A), Sau 96 (34% G+C) (lane B), Hin Rd (39% G+C) (lane C), Dpn (42% G+C) (lane D), Bam H (47% G+C) (lane E), Sty (53% G+C) (lane F), Xho (58% G+C) (lane G), Xma (64% G+C) (lane H), Rsa (66% G+C) (lane I), Sal (70% G+C) (lane J), and Alu (75% G+C) (lane K). Hin dIII digests of lambda DNA were used for size markers. The position of these fragments are shown to the right of each gel.
1A; Dra I digests. 1B; Not I digests. (in lane H M.Xma III methylation blocks Not I digestion

TABLE1

The number of DNA fragments in the size range 100 to 10,000 base pairs *observed* after Dra I, Ssp I, Not I and Sfi I cleavage is compared to the number that would be predicted from genomic mononucleotide frequency.

Species	G+C content	# of fragments Predicted	*Observed* Dra I	Ssp I	# of fragments Predicted	*Observed* Not I	Sfi I
Fusobacterium nucleatum 4H	28%	8,690	*>1000*	*>1000*	0	*0*	*0*
Staphylococcus aureus	34%	5,170	*>1000*	*>1000*	0	*0*	*0*
Haemophilus influenza Rd	39%	3,220	*>500*	*>500*	0	*0*	*0*
Streptococcus pneumoniae	42%	2,373	*>500*	*>500*	0	*0*	*0*
Bacillus amyloliquifaciens	47%	1,340	*>400*	*>400*	3	*2*	*0*
Salmonella typhimurium	53%	548	*>200*	*>200*	21	*6*	*15*
Xanthomonas holcicola	58%	198	*50*	*20*	79	*100*	*>50*
Xanthomonas malvacaerum	64%	39	*25*	*15*	293	*0**	*>100*
Rhodobacter sphaeroides	66%	21	*0*	*20*	425	*>300*	*>300*
Streptomyces albus	70%	5	*2*	*0*	806	*>500*	*>500*
Arthrobacter luteus	75%	1	*0*	*0*	1533	*>500*	*>500*

*Xma DNA is modified against Not I cleavage

protein, but not all codons occur at equal frequencies. Furthermore, there may be other systematic biases in DNA sequence arrangement. Therefore, it is not surprising to find inequalities in di- and trinucleotide frequency over and above that due to mononucleotide frequencies (10,11,12). We determined the di- and trinucleotide frequencies from a number of sequenced bacterial DNAs. This data is presented in Table 2.

Anomalies in trinucleotide frequency can be utilized to predict rare endonuclease recognition sites within a given bacterial genome. For instance, the expected frequency of Sma I (CCCGGG) sites in the *Staphylococcus aureus* genome (34% G+C) is $[0.17]^6$ = 1/43,000 base pairs, when calculated from mononucleotide frequencies. However, the predicted frequency of Sma I sites is two-fold greater when trinucleotide frequencies are taken into account (1/85,457 base pairs). Therefore, not only are G+C sequences rare in the A+T

TABLE 2 Trinucleotide Frequencies in the Genomes of Bacteria

Staphylococcus 34% G+C	*Streptococcus* 40% G+C	*Anabaena* 44% G+C	*Bacillus subtilis* 45% G+C	*Bacillus amyloliquifacien.* 45% G+C
CCG/CGG 0.57	CCG/CGG 0.86	GCC/GGC 1.65	**TAG/CTA**1.81	**TAG/CTA**1.03
CGC/GCG 0.75	CGC/GCG 0.97	CGG/CCG 1.66	CCC/GGG 1.88	CCC/GGG 1.81
CCC/GGG 0.87	GCC/GGC1.24	GCG/CGC 1.80	CGT/ACG 2.10	AGT/ACT 1.91
GCC/GGC 1.03	CCC/GGG 1.30	GGG/CCC 1.82	TAC/GTA 2.13	AGG/CCT 2.00
TCG/CGA 1.42	CGA/TCG 1.58	TCG/CGA 2.00	GTC/GAC 2.19	GTG/CAC 2.04
CGT/ACG 1.46	CAC/GTG 1.77	GAG/CTC 2.24	CAC/GTG 2.20	GAG/CTC 2.10
CTC/GAG 1.49	CGT/ACG1.88	GTC/GAC 2.27	ACT/AGT 2.26	CCA/TGG 2.11
TCC/GGA 1.71	GTC/GAC 2.15	ACG/CGT 2.33	CGC/GCG 2.31	ACC/GGT 2.12
CAC/GTG 1.77	TCC/GGA 2.25		ACC/GGT 2.36	CGA/TCG 2.16
ACC/GGT 1.97	CTC/GAG 2.31		GCC/GGC 2.49	GTC/GAC 2.32
Escherichia coli 50% G+C	*Shigella* 51% G+C	*Salmonella* 53% G+C	*Anacystis* 55% G+C	*Rhizobium* 56% G+C
TAG/CTA1.20	**TAG/CTA**1.24	**TAG/CTA**1.42	TAT/ATA 1.33	*TAA/TTA 0.99*
CTC/GAG 1.94	AGG/CCT 2.11	AGT/ACT 1.99	TTA/TAA 1.61	**TAG/CTA**1.28
CCC/GGG 1.98	GAG/CTC 2.11	GAG/CTC 2.00	ATT/AAT 2.02	GTA/TAC 1.39
AGT/ACT 2.24	ATA/TAT 2.22	GTA/TAC 2.25	**TAG/CTA**2.21	ATA/TAT 1.65
CCT/AGG 2.26	AGA/TCT 2.35	CCC/GGG 2.28	TAC/GTA 2.11	ACT/AGT 2.01
TAC/GTA 2.40	CCC/GGG 2.42	TGT/ACA 2.45	ACA/TGT 2.22	AAT/ATT 2.21
TCC/GGA 2.45	AGT/ACT 2.48	AGA/TCT 2.45	CAT/ATG 2.42	
GTC/GAC 2.48		CCT/AGG 2.46		
Pseudomonas 57% G+C	*Halobacterium halobium* 60% G+C	*Rhodobacter* 65% G+C	*Streptomyces* 70% G+C	
TAG/CTA 1.44	ATA/TAT 1.05	TAA/TTA 0.28	ATA/TAT 0.24	
TAA/TTA 1.70	TTA/TAA 1.33	**TAG/CTA**0.70	TAA/TTA 0.35	
ATA/TAT 1.89	**TAG/CTA**1.45	ATA/TAT 0.73	TTT/AAA 0.67	
GTA/TAC 2.23	TTT/AAA 1.62	ATT/AAT 0.84	**TAG/CTA**0.79	
TGT/ACA 2.27	ATT/AAT 1.62	TAC/GTA 0.92	CAT/ATG 1.18	
TCT/AGA 2.32	CAT/ATG 2.14	ACT/AGT 0.95	CAA/TTG 1.30	
	ACT/AGT 2.14	AAA/TTT 1.24	AGT/ACT 1.34	
		ACA/GTG 1.72	TGT/AGA 1.48	
		CAA/TTG 1.80	TCT/AGA 1.50	
		AAC/GTT 2.08	TGA/TCA 1.52	

Calculated from sequenced DNA available in GENBANK. Frequencies are given in percent. The margin of error is less than ± 0.5% in all samples. Average trinucleotide frequency is 3.125% in DNA. Frequencies above 2.5% are not listed.

rich *S.aureus* genome, but those hexamers which contain CCG or CGG are expected to be twice as rare as other G+C rich hexamers which lack the CCG/CGG trinucleotide. Consistent with this prediction, Sma I showed only one detectable fragment below 20 kilobases on *S.aureus* DNA, using conventional 1% agarose gel electrophoresis (data not shown). Thus, in principle, endonucleases that produce only a few fragments from an entire genome can be selected as those with recognition sequences that contain one

or more of the rarest trinucleotides from the genome of interest.

TAG is the Rarest Trinucleotide in Many Prokaryote Genomes

One striking feature of the data presented in Table 2 is that the trinucleotide TAG/CTA is exceptionally rare in the sequenced DNA of most bacteria. Generally, TAG is the rarest (amber) stop codon and its complement, CTA, is a relatively rare leucine codon. However, codon usage should only affect the frequency of TAG and CTA in one of the six reading frames in protein coding regions, which represent perhaps 50% of the genome. Nevertheless, in *E. coli* TAG+CTA is two fold rarer than predicted by mononucleotide frequency. Furthermore, 5' CTAG 3'/3' GATC 5' which contains CTA and TAG in both strands, is 2.5 fold rarer than predicted by TAG and CTA, or 12.5-fold rarer than predicted by mononucleotides. This substantial bias against CTAG is also true for many other bacterial species. Therefore, endonuclease recognition sequences containing CTAG, such as TCTAGA (Xba I), ACTAGT (Spe I), CCTAGG (Avr II) and GCTAGC (Nhe I) would be expected to be rare in many bacterial genomes. For example, the predicted frequency of Xba I sites in *Rhodobacter capsulatus* is only 1/20,000 base pairs, based upon trinucleotide frequency calculations and even rarer (1/100,000 base pairs) when the frequency of CTAG is used.

To determine whether restriction endonuclease recognition sequences containing CTAG were indeed rare in bacterial genomes, we cleaved the eleven DNAs listed in Table I with four of the endonucleases that contain the sequence CTAG (Xba I, Spe I, Avr II and Nhe I). These digests were examined by conventional electrophoresis in 1% agarose gels. From mononucleotide frequency calculations, the number of fragments that were expected to fall below 10,000 base pairs varied between 200 and 1000. The results can be summarized as follows: few, if any restriction fragments were resolved in most digests. We found that for most DNAs, (apart from species with extremely A+T rich genomes), both Xba I and Spe I sites were exceptionally rare. Avr II sites were rare in almost all species regardless of G+C content; whereas Nhe I sites, while rarer than predicted, were nevertheless more common than Avr II, Xba I or Avr II sites in most species.

We expect that, in general, Xba I sites may be rarer than Spe I sites, and that Avr II sites may be rarer than Nhe I sites, because the most common codon are a subset of RNY (R=A or G, N=A,G,C or T, Y=T or C) (13). TCTAGA and CCTAGG will not follow this RNY motif in either *open* reading frame (TCT|AGA, T|CTA|GA, CCT|AGG and C|CTA|GG), whereas Spe I and Nhe I sites will do so in one frame (ACT|AGT and GCT|AGC). Furthermore, in the GENBANK bacterial DNA database, the CC dinucleotide is almost always rarer than the GC dinucleotide, regardless of the genomic base composition. This relative

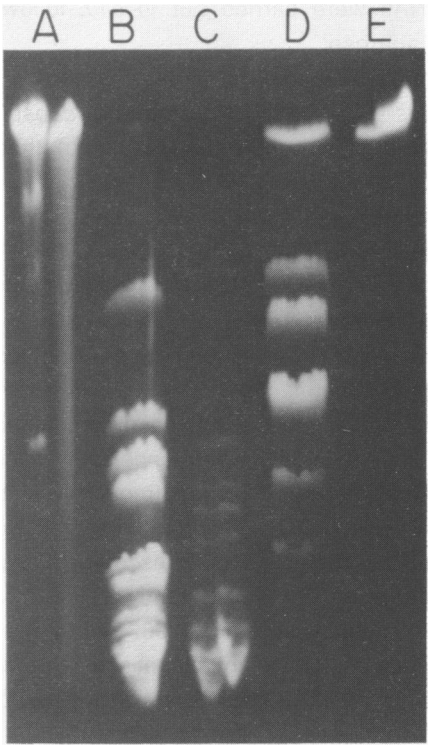


FIGURE 2 PFGE Separation of *S. aureus* Genomic DNA

Staphylococcus aureus DNA cleaved with Rsr II (Lane A), Sac II (Lane B) Sac II + Sma I (Lane C), Sma I (lane D), Not I (lane E). Fragments were separated by PFG using a 20 second pulse time for 20 hours. Molecular weights at the top of the gel are greater than 400,000 base pairs. The smallest visible fragments are 40,000 base pairs.

abundance of GC over CC/GG in bacterial DNAs indicates that Avr II sites (CCTAGG) may usually be rarer than Nhe I sites (GCTAGC).

We are presently determining the phylogenetic extent of the exceptional rarity of CTAG, which should help in understanding the biological significance of this phenomenon.

Pulsed Field Gel Electrophoresis

We sought to confirm that the endonucleases which we have predicted to be exceptionally rare cutters really do produce large (>100 kilobases) fragments from bacterial genomes. Therefore, we prepared high molecular weight DNA in agarose plugs from a series of phylogenetically diverse bacterial species which have different G+C contents. Restriction digests of

these chromosomal DNAs were carried out *in situ*, followed by pulsed field electrophoresis in 1% agarose.

In Figures 2 through 6 we show digests of *Staphylococcus aureus*, *Moraxella bovis*, *Bacillus subtilis*, and *Rhodobacter capsulatus* genomes separated by pulsed field electrophoresis. Results of these experiments are summarized below.

Staphylococcus aureus (34% G+C): We showed by a series of digests (two of which are presented in Figure 2 and 3) that there are at most one Not I (GCGGCCGC) and one Sfi I (GGCCN$_5$GGCC) sites in this genome. There are 16 Sma I (CCCGGG) sites, approximately 20 Sac II (CCGCGG) sites, and about seven Rsr II (CCGWCGG) sites. The largest Sma I fragment is about 600 kilobases. From the Sma I fragments we are able to estimate that the *S. aureus* genome is about 2.2 megabases. In contrast, other endonucleases with six G and C bases in their recognition sequence, but which do not contain the rarest CCG or CGG trinucleotides, cut more frequently in the *Staphylococcus* genome (data not shown).

We observed that the restriction patterns of *S.aureus* NCTC 8325 are similar to digestion patterns from a distantly related, antigenically distinct member of the same species, *S. aureus* PS 96. Finally, we also observed a large (>100 kilobases) molecular weight plasmid in *S.aureus* PS 96 which was not found in *S.aureus* NCTC 8325 and which would not have been detected by conventional electrophoresis.

Moraxella bovis (45% G+C) is closely related to *Neisseria* species. This genome has six Not I sites, as expected from it G+C content. However, we detected only about ten Sma I (CCCGGG) sites, much less than the over 100 sites expected from the G+C content of the species (Figure 3; lanes A, B and C).

Bacillus subtilis JAS7 (45% G+C) has about 25 Not I sites, 20 Sfi I and 20 Nhe I (GCTAGC) sites and between 40 and 100 Apa I (GGGCCC), Avr II (CCTAGG), Bgl I (GCCN5GGC), Sma I (CCCGGG) and Xba I (TCTAGA) sites. Sac II (CCGCGG), Nar I (GGCGCC) and Nae I GCCGGC) fragments are more abundant (Figure 4).

Among the Enterobacteria *E. coli* (50% G+C) *Enterobacter aerogenes* (53% G+C) and *Klebsiella pneumoniae* (58% G+C) pulsed field separations indicate that, while Not I and Sfi I sites occur more than 20 times in these genomes, Xba I and Spe I sites occur as few as 15 times (data not shown). Data presented by Hillyard et al., (submitted) indicates that there are approximately 16 Xba I and 18 Spe I sites in the related *Salmonella typhimurium* LT2 genome (53% G+C).

It should be noted here that cross-protection by endogenous methylation can result in a lower number of cleavage sites for an

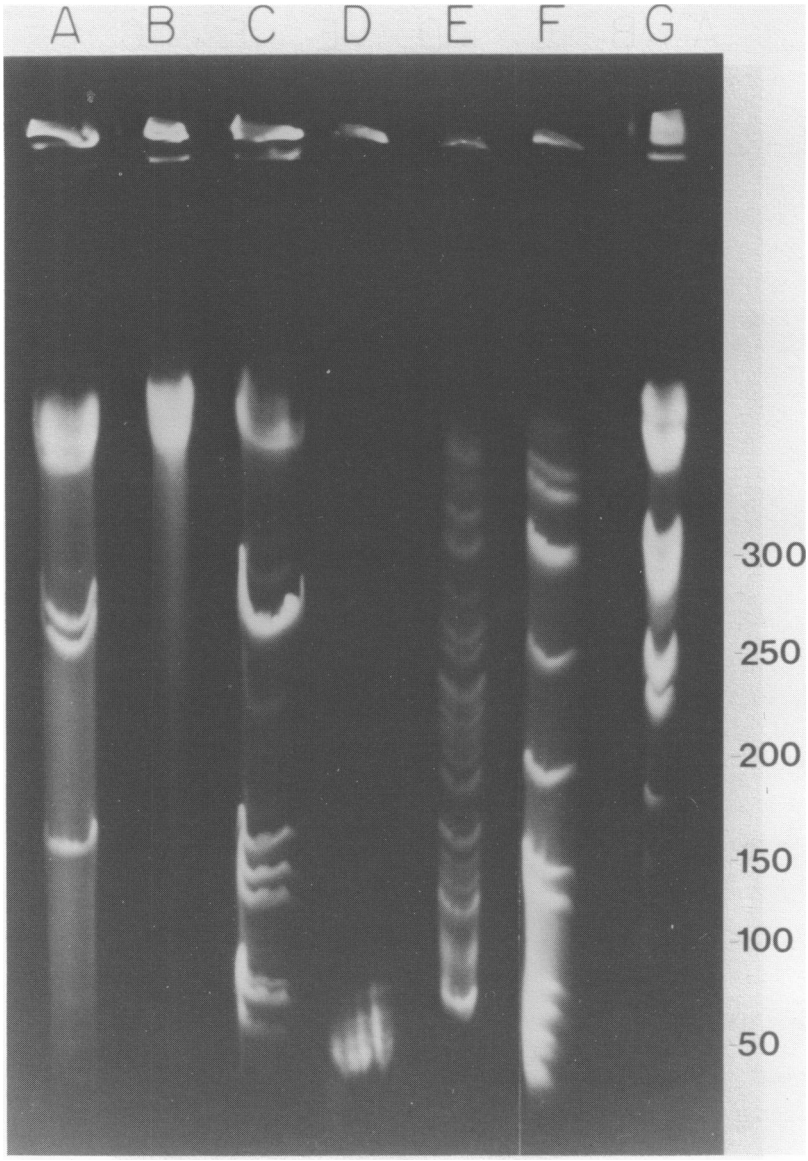


FIGURE 3 PFGE Separation of *Moraxella bovis* and S. *aureus* Genomic DNA

Moraxella bovis DNA cleaved with Not I (lane A), Sfi I (lane B), Sma I (Lane C) and Xba I (Lane D). *Staphylococcus aureus* DNA cleaved with Bgl I (Lane E), Sma I (Lane F) and *E. coli* Not I (Lane G). Fragments were separated by PFG (1) using a 10 second pulse time for 20 hours. Molecular weights, determined from bacteriophage lambda concatemers, are given in kilobase pairs.

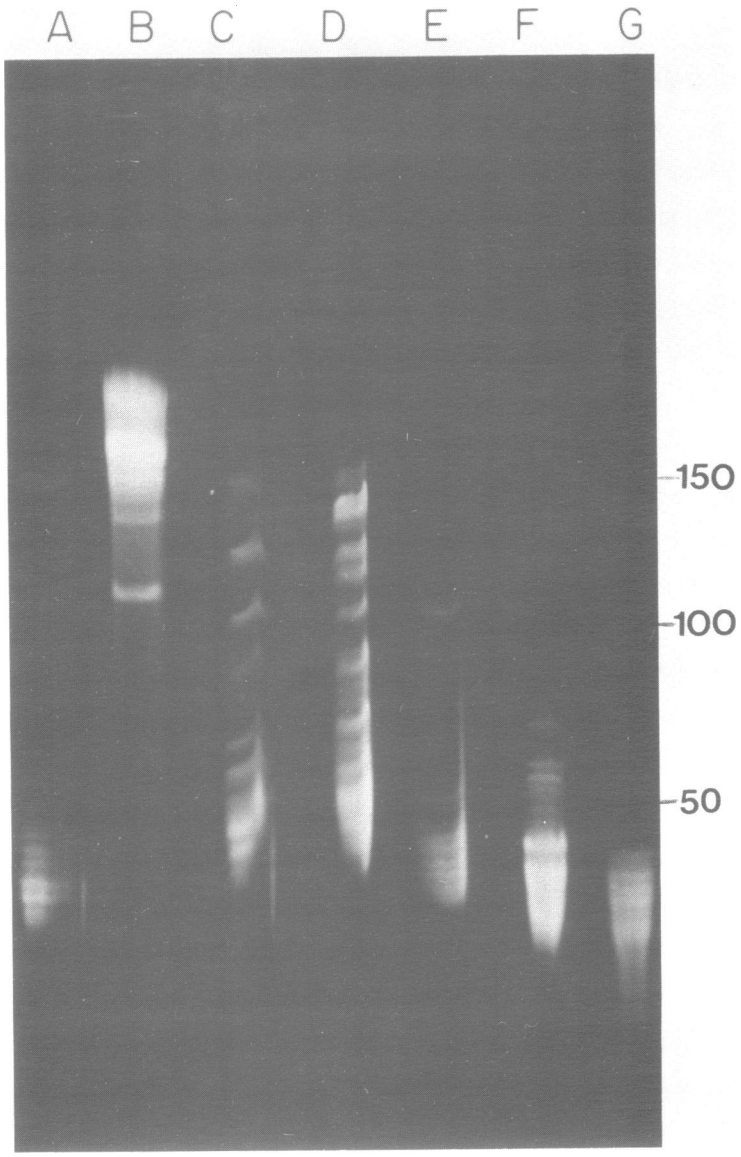


FIGURE 4 PFGE Separation of *Bacillus subtilis* Genomic DNA

Bacillus subtilis DNA cleaved with Bgl I (lane A), Sfi I (lane B), Sma I (lane C), Apa I (lane D), Sac II (lane E), Nar I (lane F), Nae I (lane G). DNAs were separated by PFG using a 5 second pulse time for 20 hours. Molecular weights, determined from bacteriophage lambda concatemers, are given in kilobase pairs.

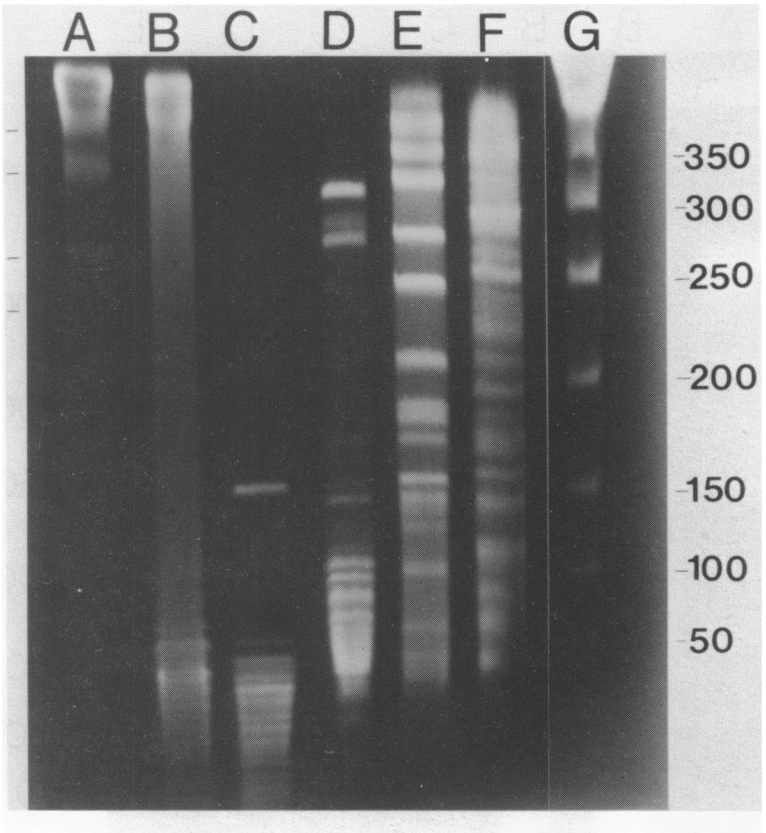


FIGURE 5 PFGE Separation of *Bacillus subtilis* Genomic DNA

Yeast chromosomes (Lane A). *Bacillus subtilis* DNA cleaved with Nar I (lane B), Bgl I (lane C), Nhe I (lane D), Sfi I (lane E), Avr II (lane F), bacteriophage lambda DNA concatemer (lane G). DNAs were separated by PFGE using a 12 second pulse time for 20 hours. Molecular weights, determined from bacteriophage lambda concatemers, are given in kilobase pairs.

endonuclease. For instance, in enterobacteria $G^{6m}ATC$ (dam) methylation will cross-protect against the Xba I sites which overlap at $TCTAG^{6m}ATC$ (14). This effect further reduces the observed number of Xba I sites.

Rhodobacter capsulatus (66% G+C) has abundant Sfi I and Not I sites. In contrast, Dra I (TTTAAA) gives a mixture of high and low molecular weight fragments after pulsed field electrophoretic separation (Figure 6). Ssp I (AATATT) sites are apparently more common, perhaps because TAT is more common than TAA (see Table 2). About fourteen fragments were observed

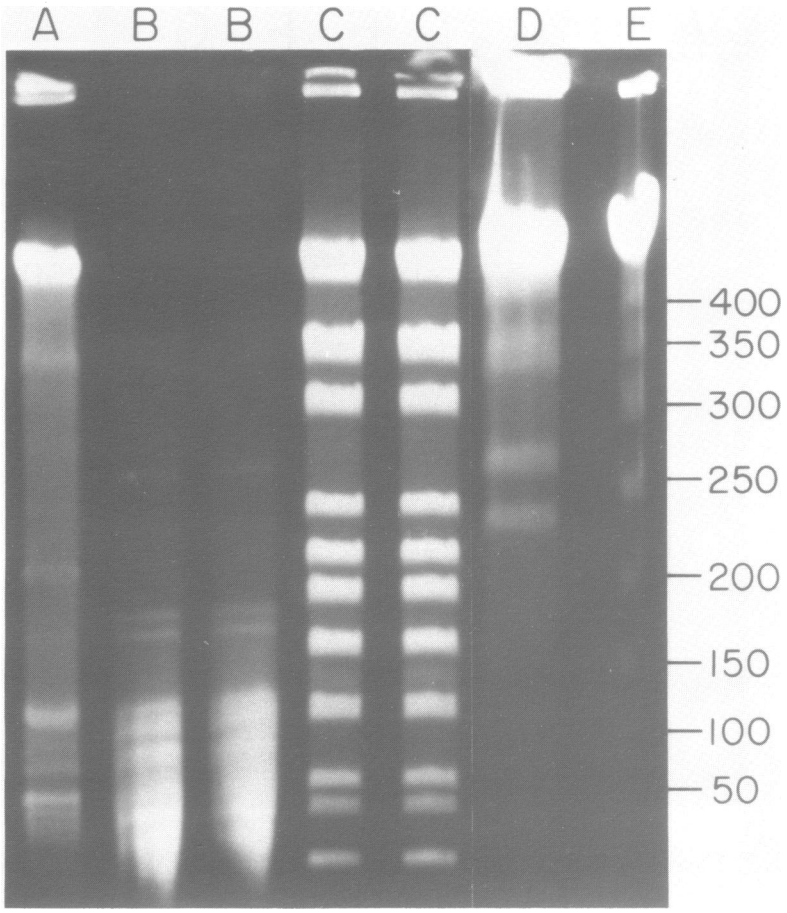


FIGURE 6 PFGE Separation of *Rhodobacter capsulatus* Genomic DNA

Rhodobacter capsulatus DNA cleaved with Dra I (lane A), Hpa I (lane B), Xba I (lane C), yeast chromosomes (lane D). Lambda DNA ladder (lane E). DNAs were separated by PFG using a 15 second pulse time for 20 hours. Molecular weights, determined from bacteriophage lambda concatemers, are given in kilobase pairs.

when *Rhodobacter capsulatus* DNA was cut with Xba I followed by pulsed field electrophoretic separation.

Mapping of Genes to Xba I Fragments of *Rhodobacter capsulatus*

Pulsed field mapping of the *Rhodobacter* genome was further exploited by electroblotting Xba I and Dra I cleaved DNA fragments onto Zetaprobe membranes and hybridizing to ^{32}P radiolabelled DNA fragments containing

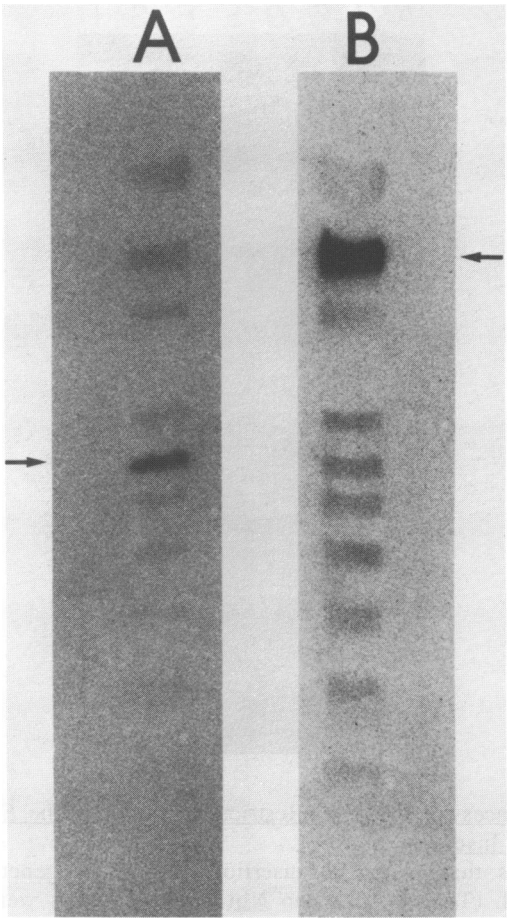


FIGURE 7 Assignment of Genes to Xba I Fragments of the *R. capsulatus* Genome

Rhodobacter capsulatus DNA cleaved with Xba I, and separated by PFG using a 15 second pulse time for 20 hours. The DNA was transferred by electroblotting to Zetaprobe and hybridized with (A) the *Rhodobacter capsulatus* rpo C gene and (B) a fragment of pRCN102 (15). The probes hybridized to a 220,000 base pair and a 375,000 base pair fragment, respectively.

cloned *Rhodobacter capsulatus* genes. DNA was transferred to Zetaprobe membrane embedded in agarose by electroblotting (see methods) and probed with two cloned DNA fragments from the *Rhodobacter capsulatus* genome. Figure 7 shows autoradiographs of an Xba I digest of *Rhodobacter capsulatus*

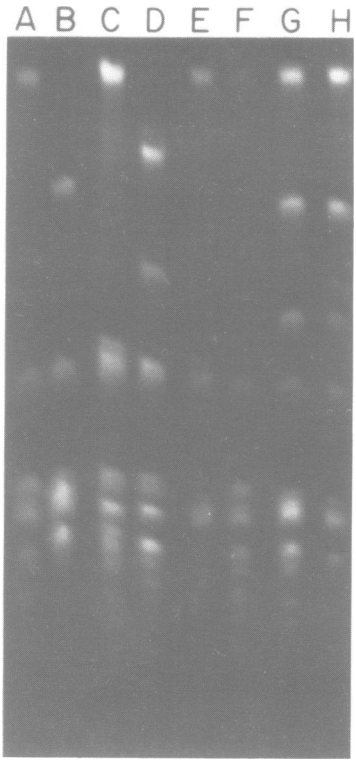


FIGURE 8 Differences in the Not I Restriction Pattern of the *E.coli* Genome Produced by Tn5 Insertion

E.coli strains that have Tn5 insertions at various genetic markers were cleaved with Not I. (Tn5 contains two Not I sites). DNAs were separated by PFG using a 15 second pulse time for 20 hours. The sizes resolved in this gel range from 40,000 to 400,000 base pairs

probed with the rpoC gene (R. Jones and B. Abella, unpublished results) and a fragment from pRCN102 (15) for which no gene has been assigned. Gene rpo C and the sequences present in pRCN102 occur on different Xba I fragments of over 100,000 base pairs. Similar experiments assigned these probes to Dra I fragments. This experiment indicates the feasibility of Zetaprobe electrotransfer Southern blotting as a method to assign genes to large restriction fragments which have been separated by pulsed field gel electrophoresis.

Transposons with rare restriction sites

Rare restriction sites within transposons (16) could help in producing ordered physical and genetic maps by insertion of the transposon into loci

that could then be mapped with PFGE. Such Xba I, Dra I, Spe I, Sma I, Sac II or Not I sites sometimes occur naturally in a transposon or can be engineered. For example, Tn5 has a naturally occurring Not I site. Figure 8 shows strains of *E.coli* with Tn5 inserted at various positions in the *E.coli* chromosome which also contains about 22 other Not I sites. Single insertions into identical genetic backgrounds should allow the genetic map to be correlated with a Not I restriction map.

Other examples of broad host range transposons that already have the appropriate restriction sites include a a naturally occurring Xba I site in Tn10 and a derivative of Tn917 into which a Sma I site has been engineered (17,18). The Tn917 derivative has be inserted into the *S. aureus* genome to produce auxotrophs (Pattee, personal communication).

DISCUSSION

Restriction Endonucleases for Pulsed Field Mapping of Bacterial Genomes

We have demonstrated that the rarity of certain DNA sequences in the genomes of bacteria have a profound effect on the frequency of certain restriction endonuclease recognition sequences. An immediate practical conclusion of these studies is that a few available restriction endonucleases are suitable for pulsed field gel electrophoretic mapping of whole bacterial chromosomes, (in particular, Avr II, Eag I, Nae I, Nhe I, Not I, Sac II, Sfi I, and Sma I, for A+T rich genomes and Dra I, Spe I, Ssp I and Xba I for G+C rich genomes). Although these endonucleases possess cleavage specificities of six base pairs in length, the non-random sequence arrangement of bacterial DNA allows for highly selective chromosomal cutting by these enzymes. These endonucleases can often be used in situations where the two endonucleases with eight base pair specificities, Sfi I and Not I, either cleave too frequently, (such as in G+C rich DNA), or do not cleave at all, (such as very A+T rich DNA).

In our cutting strategies, we take advantage of two- to three-fold deviations from expected di- and trinucleotide frequencies of bacterial DNA. However, these deviations from expected frequencies are amplified when they occur more than once within palindromic six base restriction endonuclease recognition sequences.

For those bacterial species that we have investigated there are windows of G+C content inside of which different endonucleases are likely to produce one to thirty cleavages per bacterial genome, Figure 9. Bacterial genome size is not generally a major consideration, since most prokaryotic chromosomes range from 2,000 kilobases to 5,000 kilobases (19,20,21).

Assuming that a few cuts per genome are optimal for genomic

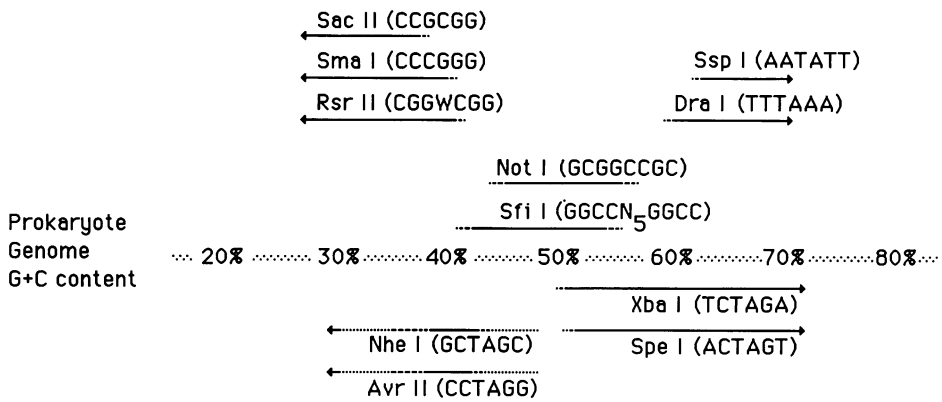


FIGURE 9 Endonucleases that are Expected to Produce Fragments that Average over 100,000 Base Pairs on Prokaryote Genomes of the Specified G+C Content

restriction mapping, then the following strategies for selecting pulsed field mapping endonucleases can be used.

[<40% genomic G+C]: In general, G+C rich hexamer endonucleases, particularly Sma I (CCCGGG) and Sac II (CCGCGG) are useful PFGE mapping tools in this range. Nae I (GCCGGC), Nar I (GGCGCC) and Eag I (CGGCCG) may also be useful. Not I (GCGGCCGC) and Sfi I (GGCCN5GGCC) are **too rare** to occur in most prokaryote genomes with G+C contents below 40%. This fact may be particularly useful since unique Not I or Sfi I sites may be introduced into the genome (Eg. on transposons) to assist in the production of a physical map or to correlate this map to the genetic map.

[40-50% genomic G+C]: In general, Avr II (CCTAGG) and Nhe I (GCTAGC) are of use as PFGE mapping tools in this range of G+C content. Not I and Sfi I will produce one to forty cuts per bacterial genome. In addition, Not I sites, which contain the relatively rare CCG/CGG sequences should generally be rarer than Sfi I sites.

[50-65% genomic G+C]: In general, Xba I and Spe I give 10 to 30 cuts per genome. Not I and Sfi I sites are **too common** to be used for genomic mapping of DNAs with G+C contents over 55%.

[>65% genomic G+C]: For prokaryotes with G+C contents above 65%, Xba I, Spe I, Dra I (TTTAAA) and Ssp I (AATATT) are useful megabase mapping enzymes.

In summary, endonucleases that contain either CTAG or CCG/CGG are exceptionally rare in every bacterial genome we have investigated. Such endonucleases can be used to produce genomic fragments in the size range

appropriate for pulsed field gel separation. We are presently developing strategies to rapidly produce complete physical maps and correlate these with genetic maps.

ACKNOWLEDGEMENTS

We thank Geoff Wilson and co-workers for supplying us with many bacterial DNAs, New England Biolabs for support, Claire Berg for supplying many of the *E.coli* with Tn5 insertions and Melanie Ehrlich for unpublished data. We thank Peter Pattee for helpful discussions. M.M is a Lucille P. Markey Biomedical Scholar. M.N. is a Schipol Fellow. This research was funded by The Louis Block Fund, The Greenblatt Cancer Research Foundation and The Lucille P. Markey Charitable Trust.

*Present address: Department of Biological Sciences, University of Southern California, University Park, Los Angeles, CA 90089, USA

REFERENCES

(1) Gardiner K., Laas W. and Patterson D. (1986) Somatic Cell Mol. Genet. 12:185-195
(2) Schwartz D. and Cantor C.R. (1984) Cell 37:67-75
(3) Marmur J. (1962) Methods in Enzymology 6:726-738
(4) McClelland M. (1987) In: Methods in Enzymology, "Recombinant DNA," Parts D and E, Wu R., ed., Academic Press (in press).
(5) Waterbury P.G. and Lane M. J. (1987) Nucleic Acids Res. 15:3930
(6) Reed K.C. and Mann D.A. (1985) Nucleic Acids Res. 13:7207-7221
(7) Feinberg A.P. and Vogelstein B. (1983) Anal. Biochem. 132:6-13
(8) Ehrlich M., Gama-Sosa M.A., Carreira L.H., Ljundahl L.G., Kuo K.C. and Gehrke C.W. (1985) Nucleic Acids Res. 13:1399-1412
(9) Setlow P. (1974) CRC Handbook of Mol. Biol. and Biochem. G.D. Fasman Ed., CRC Press.
(10) McClelland M. (1985) J. Mol. Evol. 21:317-322
(11) McClelland M. and Nelson M. (1987) In: Gene Analysis and Amplification, Chirikjian J.G., ed., Chapter 10, Volume V (Elsevier-North Holland, New York), in press.
(12) Phillips G.J., Arnold J. and Ivarie R. (1987) Nucleic Acids Res. 15:2627-2638
(13) Shepherd J.C.W. (1981) Proc. Natl. Acad. Sci. USA 78:1596-1600.
(14) Nelson M. and McClelland M. (1987) Nucleic Acids Res. 15: (in press)
(15) Kranz R.G. and Haselkorn R. (1985) 40:203-215
(16) Kleckner N. (1981) Ann. Rev. Genetics 15:341-404
(17) Youngman P., Perkins J.B. and Losick R. (1984) Plasmid 12:1-9
(18) Luchansky J.B. and Pattee P. (1984) J. Bacteriol. 159:894-899
(19) Kingsbury D.T. (1969) J. Bacteriol. 98:1400-1401
(20) Gillis M., de Ley J. and de Cleene M. (1970) Eur. J. Biochem. 12:143-153
(21) Wallace D.C. and Morowitz H.J. (1973) Chromosoma 40:121-126