# Recognition sequences of Type II restriction systems are constrained by the G+C content of host genomes

Michael McClelland

Department of Biochemistry and Molecular Biology, University of Chicago, 920 East 58th Street, Chicago, IL 60637, USA

Received October 27, 1987; Revised and Accepted January 7, 1988

## ABSTRACT

I show that the recognition sequences of Type II restriction systems are correlated with the G+C content of the host bacterial DNA. Almost all restriction systems with G+C rich tetranucleotide recognition sequences are found in species with A+T rich genomes, whereas G+C rich hexanucleotide and octanucleotide recognition sequences are found almost exclusively in species with G+C rich genomes. Most hexanucleotide recognition sequences found in species with A+T rich genomes are A+T rich. This distribution eliminates a substantial proportion of the potential variance in the frequency of restriction recognition sequences in the host genomes. As a consequence, almost all restriction recognition sequences, including those eight base pairs in length (Not I and Sfi I), are predicted to occur with a frequency ranging from once every 300 to once every 5,000 base pairs in the host genome. Since the G+C content of bacteriophage DNA and of the host genome are also correlated, the data presented is evidence that most Type II "restriction systems" are indeed involved in phage restriction.

## INTRODUCTION

Type II restriction systems employ a DNA methylase and an endonuclease of the same recognition sequence specificity. The sequence-specific methylase protects the DNA of the prokaryotic host from cleavage by the endonuclease, which will cleave any incoming bacteriophage DNA that is not appropriately methylated. Studies with restriction systems have led to the hypothesis that the primary role of these enzymes is to restrict infection by bacteriophage [1]. Other activities, such as the involvement of restriction systems in recombination, have also been proposed [2].

At least 580 described endonucleases have the characteristics of Type II restriction enzymes [3]. They represent 120 different recognition sequence specificities. The vast majority of Type II restriction recognition sequences are from four to six base pairs long, for example Hae III (GGCC) and Eco RI (GAATTC). Therefore, in DNA that consists of equal proportions of the four bases at random, restriction recognition sequences should occur, on average, once every $4^4$ to $4^6$ base pairs (256 to 4,096 base pairs). However, bacterial DNAs vary widely in their G+C content from about 25% G+C to 75% G+C [4]. The frequency of occurrence of a restriction recognition sequence in DNA is profoundly influenced by the G+C content of the DNA. Thus, if a species with a genomic base composition of 75% G+C carried a GGCC specific restriction system, then that recognition sequence should occur, on average, once every 50 ($1/0.375^4$) base pairs whereas sequences such as GAATTC would be expected to occur approximately once every 29,000 base pairs ($1/0.375^2 \times 0.125^4$). In a genome with a 50% G+C base composition GGCC would occur once every 256 base pairs and GAATTC once every 4,096 base pairs.

The G+C contents of most bacteriophages are similar to the G+C content of their bacterial host [(5)] (McClelland, unpublished). The rarity of GAATTC in the G+C rich DNA of phage that infect species with G+C rich genomes might limit the utility of such an endonuclease in phage restriction. Therefore, if the role of most Type II restriction systems is in restriction of phage infection, then the G+C content of the restriction recognition sequences must reflect the G+C content of the bacterial genome: restriction recognition sequences will occur in a range of frequencies appropriate to their function. I have found a dependence of restriction recognition sequence-specificity on the G+C content of bacterial genomes that is consistent with this hypothesis.

## METHODS

All double-stranded DNA sequence-specific endonucleases which are isolated from prokaryotes, and which have a $Mg^{2+}$ requirement but are ATP and S-adenosyl-methionine independent, are considered Type II restriction systems. The G+C contents of 455 species that carry characterized Type II restriction systems were obtained from the literature [(4,6,7)]. The data were divided into six groups based on the G+C content of the bacterial host genome, (20-29%, 30-39%, 40-49%, 50-59%, 60-69% and 70-79% G+C) and into six groups based on the recognition sequence length and G+C content of the restriction system:

(i) tetranucleotides with four G:C base pairs, including those with split sequences (such as GGNCC);

(ii) tetranucleotides with two G:C base pairs and two A:T base pairs, including those with split sequences (such as GANTC);

(iii) hexanucleotides with two G:C base pairs and four A:T base pairs;

(iv) hexanucleotides with four G:C base pairs and two A:T base pairs;

(v) hexanucleotides with six G:C base pairs;

(vi) recognition sequences between four and six base pairs in length, including those with redundant sets of specificities (such as GTYRAC and CCWGG).

### Sample Bias

Many isoschizomer restriction endonucleases have been found in closely related species and strains. To avoid this sample bias only one isoschizomer was used from each genus to generate the data presented in this paper. This reduced the sample size from 455 to 308 species in 65 genera.

### Calculating Expected Recognition Sequence Frequency

The "expected frequency" of restriction recognition sequences in the bacterial genome is used throughout this paper. An expected recognition sequence frequency is calculated from the

G+C contents of the host bacterial genome and of the recognition sequence. For example, if a bacterial genome has a G+C content of 60%, then G and C have an abundance of 30% each and A and T have an abundance of 20% each in the genome. In the recognition sequence GGATCC there are four occurrences of G or C and two occurrences of A or T. The calculated recognition sequence frequency is once every $1/(0.3)^4 \times (0.2)^2$ or 3,086 base pairs, on average. Note that these calculations do not take into account di- and trinucleotide frequencies in the genome because such data is available for very few species.

The Random Model

The recognition sequence frequencies were calculated for **each** of the 308 palindromic restriction systems in the sample in **every** bacterial species in the sample. Thus, a total of 308 x 308 (94,864) calculations were performed. The distribution of the palindromic restriction sequence frequencies in this random model was compared to the distribution of frequencies in the 308 genomes of the bacterial species in which the recognition sequences actually occur.

## RESULTS

The genomic G+C content was determined for 308 species that contain Type II restriction systems (3). The data was divided into six groups depending on the G+C content and length of the restriction recognition sequences (see Methods). The median bacterial genomic G+C content and 25th and 75th percentile of rank for the data in each of the five classes of palindromic restriction systems (i thru v, Methods) was calculated. *Figure* 1 plots this data along with the theoretical frequency of restriction sequences in the host bacterial genome as a function of G+C content of the bacterial genome. Note that all but one of the curves are highly sensitive to the bacterial genomic G+C content.

Tetranucleotides

Restriction systems with palindromic tetranucleotide recognition sequences containing two G:C base pairs and two A:T base pairs (such as Mbo I GATC) are distributed fairly evenly among species with varying G+C contents, with a median at a host genomic G+C content of 51% (*Figure* 1 and 2A). The frequency of this kind of recognition sequence is relatively insensitive to G+C content. All these restriction systems have expected frequencies of once every 256 to 310 base pairs in their respective genomes.

In contrast, the frequency of tetranucleotide recognition sequences with four G:C base pairs (such as Hae III GGCC) is profoundly influenced by the G+C content of the genome (*Figure* 1). However, 77% (35 of 45), of these restriction systems are found in species with A+T rich genomes with a median at a host genomic G+C content of 42% (*Figure* 2B). The expected frequency of G+C rich tetranucleotide recognition sequences such as GGCC in a genome with a 42% G+C content is once every 514 base pairs. Only 3% of G+C rich tetranucleotide restriction systems are found in species with genomic G+C contents over 55%,

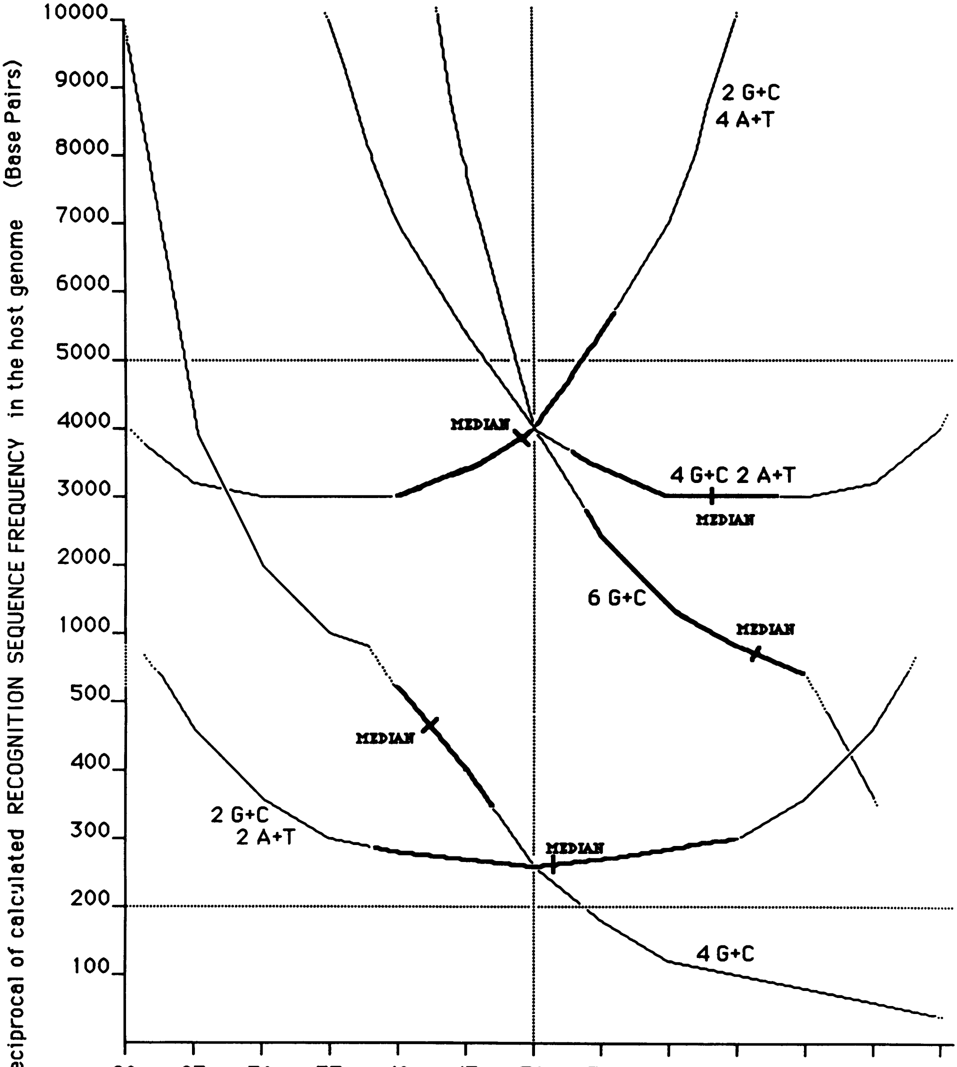


*Figure* 1. **Relationship Between The Host Genomic G+C Content And Restriction Recognition Sequence Frequencies.**

This figure includes the median bacterial genomic G+C content and the range of G+C content corresponding to the 25th to 75th percentiles (in **bold**) for the species that contain each of the five categories of palindromic restriction system: (i) tetranucleotides with four G:C base pairs; (ii) tetranucleotides with two G:C base pairs and two A:T base pairs; (iii) hexanucleotides with two G:C base pairs and four A:T base pairs; (iv) hexanucleotides with four G:C base pairs and two A:T base pairs; (v) hexanucleotides with six G:C base pairs.

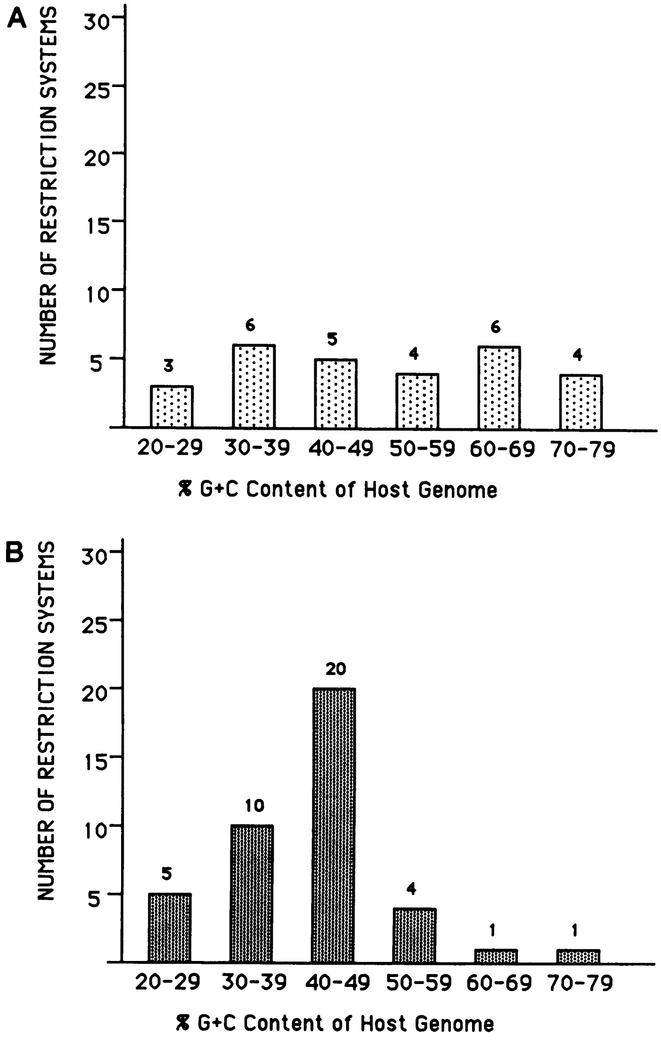


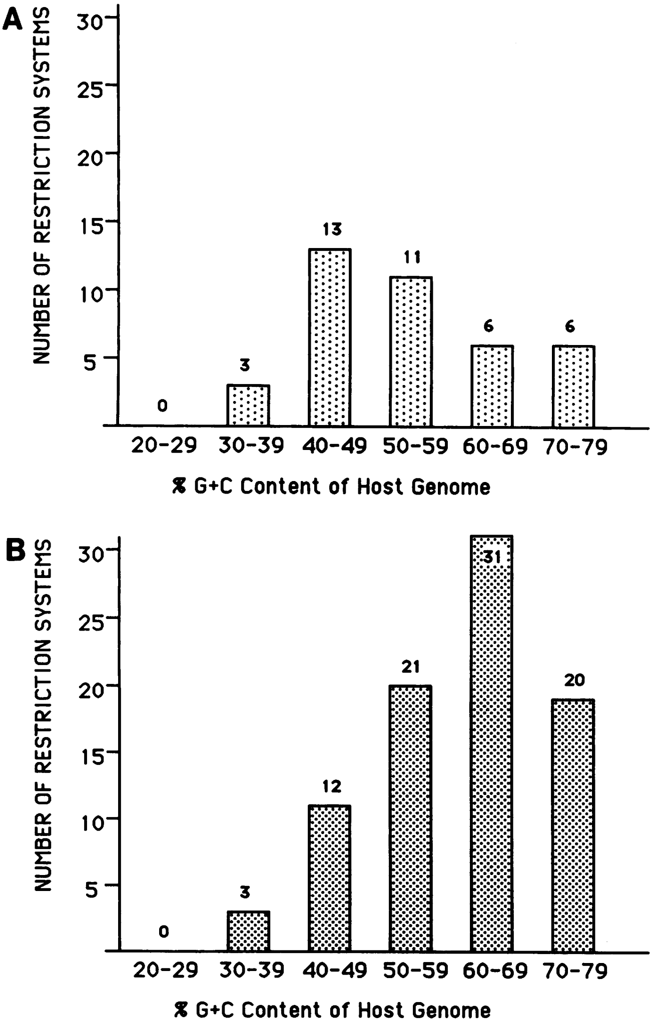
*Figure* **2A. Tetranucleotide Restriction Systems with Two G:C and Two A:T Base Pairs. B. Tetranucleotide Restriction Systems with Four G:C Base Pairs.**

where their recognition sequences are expected to occur less than once every 200 base pairs.

### Hexanucleotides

Restriction systems with A+T rich palindromic hexanucleotide recognition sequences, such as Eco RI (GAATTC), with two G:C base pairs and four A:T base pairs, and those with six A:T base pairs, represent 50% of the hexanucleotide restriction systems in species with A+T rich genomes (17 of 34) but only 19% of restriction systems in species with G+C rich genomes (24 of 159). Nevertheless, restriction systems with A+T rich hexamers are symmetrically distributed (*Figure* 3A) with a median at a genomic G+C content of 50% (*Figure* 1). This distribution may be accounted for by the fact that 79% of hexanucleotide recognition sequences (125 of 159) are found in species with G+C rich genomes. One wonders why hexanucleotide restriction systems are found mainly in species with G+C rich genomes and why A+T rich hexanucleotide specificities are relatively uncommon.

Almost all (96%) of hexanucleotide recognition sequences with 6 G:C base pairs, such as Sma I (CCCGGG), and 83% of hexanucleotide recognition sequences of four G:C and two A:T



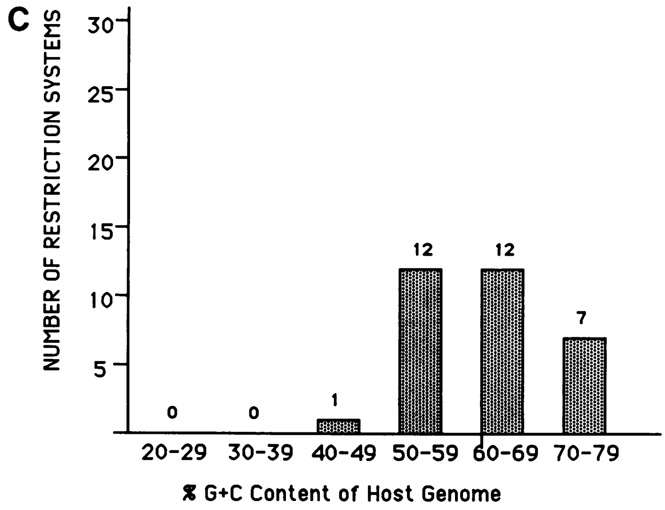


*Figure* 3A. **Hexanucleotide Restriction Systems with Two G:C and Four A:T Base Pairs. B. Hexanucleotide Restriction Systems with Four G:C and Two A:T Base C. Hexanucleotide Restriction Systems with Six G:C Base Pairs.**

base pairs, such as Bam HI (GGATCC), are found in G+C rich species (*Figure* 3B and C). These G+C rich hexamers are distributed with median bacterial genomic G+C contents of 64% and 63%, respectively (*Figure* 1).

The predicted frequency of these recognition sequences is profoundly influenced by the G+C content of the bacterial host. For instance, the expected frequency of CCCGGG in a 70% G+C genome is once every 544 base pairs and in a 30% G+C genome is once every 87,791 base pairs. However, because of the observed distribution, only 8% of the G+C rich hexanucleotide restriction systems occur in species with genomes where the frequency of the recognition sequence is expected to be more than once every 5,000 base pairs.

Intermediate Restriction Recognition Sequence Lengths

Restriction systems with an effective recognition sequence length intermediate between four and six base pairs (such as Mbo II GAAGA, Acc I GTMKAC, Eco RII CCWGG) are most often found in species with genomic G+C contents near 50% (*Figure* 4). Of these restriction systems 73% (58 of 80) are found in species with genomic G+C contents between 40 and 60%. In contrast, only 45% of tetranucleotide and hexanucleotide restriction systems are found in this range of genomic G+C content (103 of 228). For restriction systems with intermediate recognition sequence length, 96% are predicted to occur between once every 300 and once every 4,000 base pairs in the host genome.

Restriction Recognition Sequence Lengths of Over Six Base Pairs

The dependence of Type II restriction recognition sequence on the G+C content of the

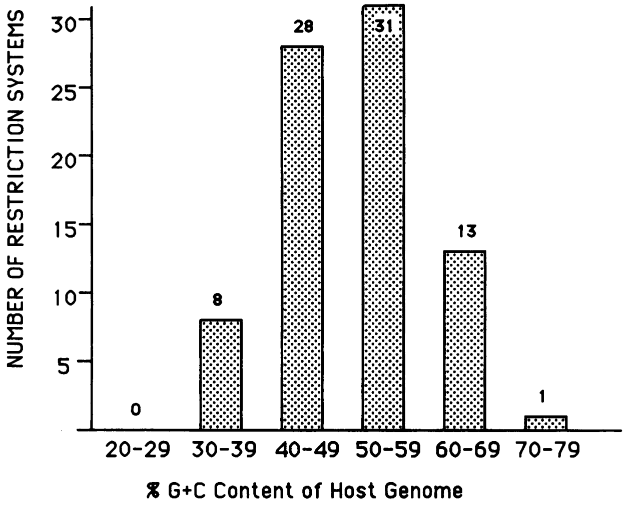


*Figure* 4. **Restriction Systems with Intermediate Recognition Sequence Length.**

bacterial genome extends to the Type II restriction systems with recognition sequences of length greater than six base pairs, Not I, Sfi I, Rsr II, Cpo I, Aci I and Sri I (R. Morgan, unpublished) and Csp I (McClelland and Nelson, unpublished). All have G+C rich recognition sequences and are found in species with G+C contents of at least 60%. Their recognition sequences are expected to occur in their respective bacterial genomes with frequencies between once every 2,277 and once every 4,982 base pairs. For instance, Sfi I (GGCCN$_5$GGCC) is found in a species with a G+C content of 70%[8]. This eight base pair recognition sequence is calculated to occur once every 2,560 ($1/0.375^8$) base pairs in this genome, compared to once every 65,000 base pairs in a genome of 50% G+C. These long recognition sequences are quite common in DNA with the host genomic G+C content. In this respect hepta- and octanucleotide restriction systems are similar to most other Type II restriction systems.

### Restriction Recognition Sequence Frequency in the Host Genome

Restriction recognition sequence frequency is dependent on the host genomic G+C content, the length of the recognition sequence, and the G+C content of the restriction sequence. However, the observed non-random distribution of restriction systems among different bacteria, based on genomic G+C content, reduces substantially the variation that these factors introduce. The expected frequencies of restriction recognition sequences in the host genome lie between once every 200 to 5,000 base pairs for a much larger percentage of restriction systems than can be accounted for by chance.

In *Figure* 5 the data on the expected restriction recognition sequence frequency in the genomes of the hosts is compared to a random model. The random model assigns the palindromic restriction systems in the sample to all the species in the sample that carry

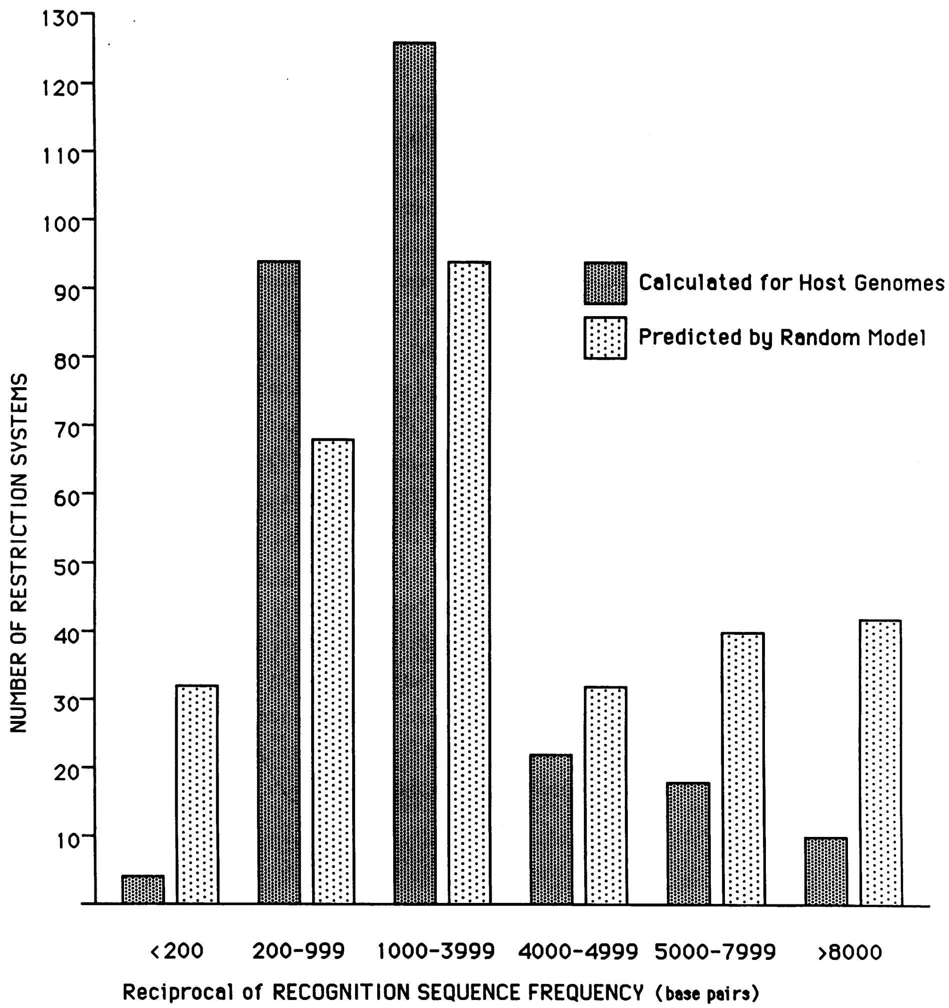


*Figure* 5. **Recognition Sequence Frequencies Calculated For Host Genomes Compared To Frequencies Predicted Under The Random Model.**

palindromic restriction systems. Only four restriction systems in the sample are calculated to have recognition sequences with a frequency of greater than once every 200 base pairs, whereas the random model predicts 31 such restriction systems. Similarly, the data show 31 restriction specificities calculated to occur in the host genome less frequently than once every 5,000 base pairs, compared to 83 predicted by the random model.

Exceptions

A few (35 of 308) of the restriction systems in the sample are exceptions to the general observation that the calculated frequency of restriction recognition sequences in the host genome lies between once every 200 and once every 5,000 base pairs. One possible explanation is that the calculations of frequency of restriction specificities in the host genome are based on bacterial G+C content alone. Factors such as selection within a genome either for or against certain sequences may substantially alter the actual frequency of the recognition sequences in question (6,9,10). This possibility has not been addressed here due to a lack of data.

Exceptions calculated to be most rare in the host (and host-specific phage) genomes, based on G+C content, include the four enzymes found in the *Sphaerotilus* genus which has a 66% G+C rich genome (11, confirmed in this lab). This genus contains restriction systems with A+T rich recognition sequences, such as Sna I, Sna BI, and Spe I with predicted frequencies in the genome of once every 16,125 base pairs, and Ssp I with a predicted frequency of once every 87,791 base pairs. Other extreme examples are Xba I (TCTAGA) from *Xanthomonas badrii* with a predicted frequency of once every 6,558 base pairs and Dra I (TTTAAA) from *Deinococcus radiophilus* with a predicted frequency of once every 21,256 base pairs. One possible reason for the rarity of these endonucleases is that they perform a function very different from other restriction systems. For example, these enzymes may not be used to restrict phage infection.

Restriction recognition sequences that occur more frequently than once every 200 base pairs in the genome include Blu I (GGCC) from *Brevibacterium luteum* with a predicted frequency of once every 63 base pairs. It is possible that the frequency of this sequence has been overestimated as there may have been selection against this sequence in the genome. For example, the sequence GGCC may be methylated at $^{5m}$C, which is hypermutable (12). During evolution, this would result in the elimination of many GGCC sites in the *Brevibacterium* genome.

## DISCUSSION

The distribution of restriction systems in bacteria is non-random with respect to G+C content. Restriction systems tend to occur in species where the frequency of restriction sites in the host genome will fall in a narrow range of between once every 200 and once every 5,000 base pairs.

In summary, five general rules may be formulated:

(1) Restriction systems with tetranucleotide recognition sequences that are G+C rich tend to be found in species with A+T rich genomes, where the recognition sequences occur less frequently than once every 256 ($4^4$) base pairs.

(2) Restriction systems with hexanucleotide recognition sequences that are G+C rich tend to be found in species with G+C rich genomes, where the recognition sequences occur more frequently than once every 4,096 ($4^6$) base pairs.

(3) Restriction systems with hexanucleotide recognition sequences that are A+T rich are the

majority of hexanucleotide restriction systems in species with A+T rich genomes, where the recognition sequences occur more frequently than once every 4,096 ($4^6$) base pairs.

(4) Restriction systems with recognition sequences that are intermediate between four and six base pairs tend to be found in species with genomes that are about 50% G+C, where the recognition sequences occur between once every 256 ($4^4$) and 4,096 ($4^6$) base pairs.

(5) Restriction systems with recognition sequences longer than six base pairs are found in species in the genomes of which their recognition sequences will occur frequently. For example, all such endonucleases discovered so far have G+C rich recognition sequences and are found in species with G+C rich genomic DNA.

Although exceptions exist, the observations presented here have predictive power. Bacterial species with a particular genomic G+C content tend to have restriction systems with certain recognition sequence lengths or G+C contents. For instance, it seems likely that most Type II restriction systems with recognition sequences of length over six base pairs will tend to be found in species with very G+C rich genomes or very A+T rich genomes, and will tend to have G+C rich or A+T rich recognition sequences, respectively.

From a practical standpoint, endonucleases with long recognition sequences will be invaluable for the production of very large DNA fragments. For example, Not I GCGGCCGC cleaves the A+T rich human genome about once every 1,000,000 base pairs. These large DNA fragments can be separated by Pulsed Field Gel Electrophoresis (13).

However, another conclusion is that there may be very few Type II restriction endonucleases with specificities over eight base pairs in length; there are very few bacterial species with genomic G+C contents over 75% or less than 25%, as would be required for the appropriate G+C rich or A+T rich nine or ten base pair recognition sequences to occur frequently in the genome. For example, in a species with a genomic G+C content of 70%, the frequency of a nine base pair G+C sequence is 50,800 ($0.3^9$) base pairs; for a G+C content of 75% the same sequence has a frequency of 6,820 ($0.35^9$) base pairs. Methods other than the use of restriction endonuclease digestion will be required in order to achieve cleavage specificities in excess of eight base pairs.

The G+C contents of bacteriophages are usually similar to the G+C content of their bacterial host (ref. 5 and McClelland, unpub.). The data presented here indicates that there may be selective pressure for a narrow range of Type II restriction recognition frequency in host or host-specific phage genomes that extends to restriction systems with long recognition sequences. The lower limit in frequency of sites for most restriction systems, once every 5,000 base pairs, may be maintained so that these restriction systems can function effectively against phage. The upper limit in frequency, about once every 300 base pairs, is more difficult to explain. Perhaps there is selection against a heavily methylated genome or against damage created by nicking or cleavage of a few residual unmethylated or hemimethylated sites after replication.

"Unusual" restriction systems have been detected that are only exceptional when one considers them in the context of bacterial genomic G+C content. Such systems, with recognition sequences that may be very rare or very common in the host genome, might not be involved in typical functions such as phage restriction. These potential exceptions deserve further investigation.

## ACKNOWLEDGEMENTS

I thank Mike Nelson for invaluable discussions and Joan Brooks, Tom Terwilliger, John Hanish and, of course, Werner Zumft for critical reading of the manuscript. I am a Lucille P. Markey Biomedical Scholar. This work was supported by the Lucille P. Markey Charitable Trust and the Goldblatt Research Foundation.

## REFERENCES

1. Dussoix D. and Arber W. (1962) J. Mol. Biol. 5:37-49.
2. Price C. and Bickle T.A. (1986) Microbiological Sciences 3:296-299.
3. Roberts R.J. (1985) Nucleic Acids Res. 13:r165-r200.
4. Bergeys Manual of Systematic Bacteriology (1984) Ed Holt J.G., Williams and Wilkins, Baltimore/London.
5. Gibbs A. and Primrose S. (1976) Intervirology 7:351-355.
6. Ehrlich M., Gama-Sosa M.A., Carreira L.H., Ljungdahl L.G., Kuo K.C. and Gehrke C.W. (1985) Nucleic Acids Res. 13:1399-1412.
7. Huang L.-H., Farnet C.M., Ehrlich K.C. and Ehrlich M. (1982) Nucleic Acids Res. 10:1579-1591.
8. Quiang B.Q. and Schildkraut I. (1984) Nucleic Acids Res.12:4507-4516.
9. McClelland M., Jones R., Patel Y. and Nelson M. (1987) Nucleic Acids Res. 15:5985-6005.
10. Sharp P.M. (1986) Mol. Biol. Evol. 3:75-83.
11. Mandel M., Johnson A. and Stokes J.L. (1966) J. Bacteriol. 91:1657-1658.
12. Coulondre C. and Miller J.H. (1978) Nature 274:775-780.
13. Schwartz D. and Cantor C.R. (1984) Cell 37:67-75.