J. theor. Biol. (1982) **98**, 165–169

LETTER TO THE EDITOR

The Average Spacing of Restriction Enzyme Recognition Sites in DNA

The discovery of naturally occurring enzymes which cleave DNA at sites specific to particular nucleotide sequences has had a great impact on molecular biology. The function of these enzymes *in vivo* is to protect bacterial cells from viral invasion by degradation of foreign DNA. Several hundred of these "restriction" enzymes are known and they are a very common tool both for analysis and manipulation of DNA. The overwhelming majority of restriction enzyme recognition sites are four or six nucleotides in length and have the remarkable property of internal symmetry with respect to nucleotide sequence (i.e. diad symmetry). A major practical offshoot of the characterization of restriction enzymes has been their use in DNA cloning. In addition, comparison of restriction enzyme cleavage patterns has been used analytically to assess the similarity of DNA from related organisms—particularly mitochondrial DNA. As an aid in these studies, a number of statistical methods have been devised to analyze data generated by comparison of restriction enzyme digestion patterns. The first of these studies was carried out by Upholt (1977) and Upholt & Dawid (1977). This work was revised by Nei & Li (1979) and by Gotoh *et al.* (1979). The intent of this note is to add practical detail to these analyses.

One of the parameters addressed in the studies referenced above is an estimate of the average spacing of restriction enzyme cleavage sites in a molecule of DNA. This issue is of importance in selecting enzymes with which to analyze a particular DNA and also in assessing the randomness of site distribution within a given DNA sequence. If the distribution of sites is random, the average spacing is a function of (1) the particular nucleotide recognition sequence and (2) the G-C/A-T content of the DNA, which varies significantly in DNA from different sources. Deviation from the expected distribution of restriction sites may indicate the presence of distinctive sequence features such as those discussed below.

The derivation of an equation which predicts the average spacing of cleavage sites is relatively simple if we make three assumptions of a nature common in probability problems: (1) the sites occur randomly along the DNA strand; (2) the strands are very long, comprising many nucleotides; (3) the occurrence of any given base pair on the strand may be considered

0022-5193/82/170165+05 $03.00/0

an independent event uninfluenced by the occurrence of others. Although assumed to occur at random, sites are not all equally probable. Rather, the probability varies with the total G-C/A-T content of the DNA. For example, the probabilities of occurrence, P_G or P_C, of the bases G or C on one strand is given by

$$P_G = P_C F_{GC}/2$$

and the corresponding probability of A or T is

$$P_A = P_T = F_{AT}/2$$

Here the F_{GC} and F_{AT} are the respective fractions of G-C and A-T in the DNA. The factor 2 takes into account the fact that the DNA is double stranded, with occurrence on either strand equally likely but mutually exclusive. Clearly $F_{GC}+F_{AT}=1$. The assumption of independence then allows us to write the probability for any given sequence of bases as the product of the independent probabilities, i.e.

$$P = \prod_i P_i$$

For a restriction site consisting of m_{GC} base pairs of G-C and m_{AT} pairs of A-T (including both strands)

$$P = (F_{GC}/2)^{m_{GC}}(F_{AT}/2)^{m_{AT}}.$$

$m_{GC}+m_{AT}$ is the total number of base pairs in both strands of the recognition site. Then

$$\text{Average site spacing} = 1/P.$$

This equation was stated previously by Upholt (1977) without proof.

To make the result more generally useful, we have graphed the equation in a form which allows prediction of site spacing for enzymes which recognize four and six nucleotide sequences with diad symmetry (Figs 1 and 2). This is appropriate since most known restriction enzymes fall into one of these two catagories. Application of the curves involves no numerical calculation and they are, therefore, rapid and convenient to use in the laboratory.

In cases where the G-C/A-T content of a particular DNA is known, comparison of the restriction site spacing predicted by the curves of Figs 1 and 2 to experimentally determined site positions can be informative. Randomness of the DNA sequence can be assessed from the extent to which the experimental data fit the predicted spacing. Blocks of sequence which vary significantly from the spacing predictions are non-random and often contain distinctive features. For example, regions which are A-T or

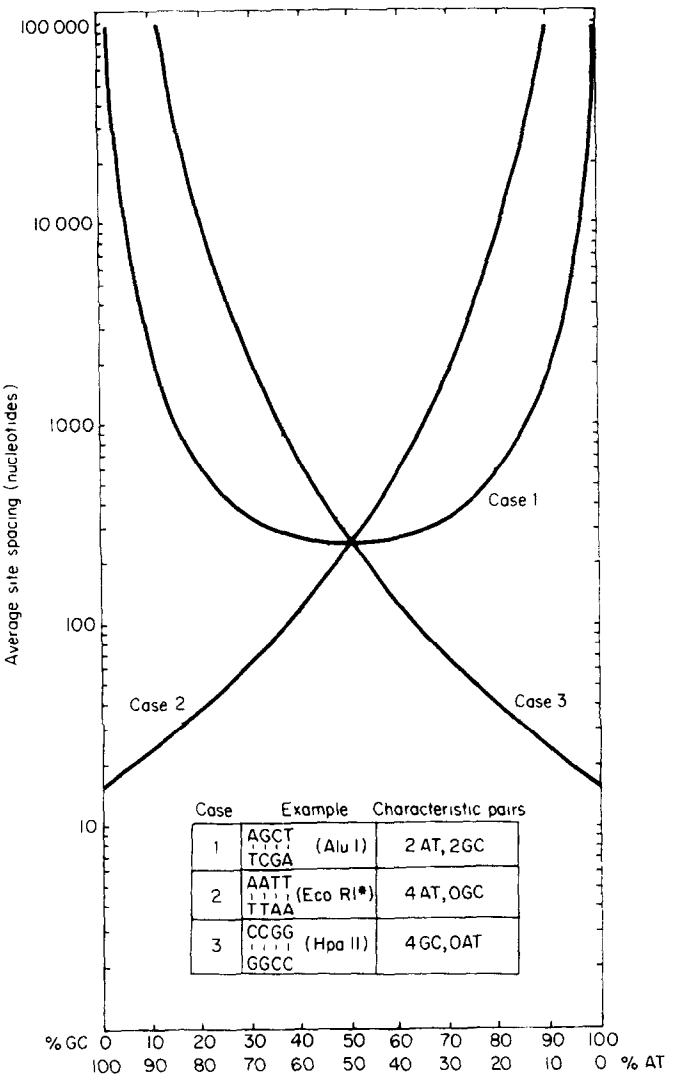

FIG. 1. Graphical representation of the equation for restriction enzyme recognition sequences consisting of four nucleotides. For these cases $m_{GC} + m_{AT} = 4$.

G-C rich can be identified since they will be cleaved preferentially by enzymes which recognize A-T or G-C rich sites. Also, repeated DNA sequences result in an unusually large number of sites for enzymes whose recognition sequence is contained within the repeat element. DNA which contains such repeat elements is anomalous with respect to site spacing

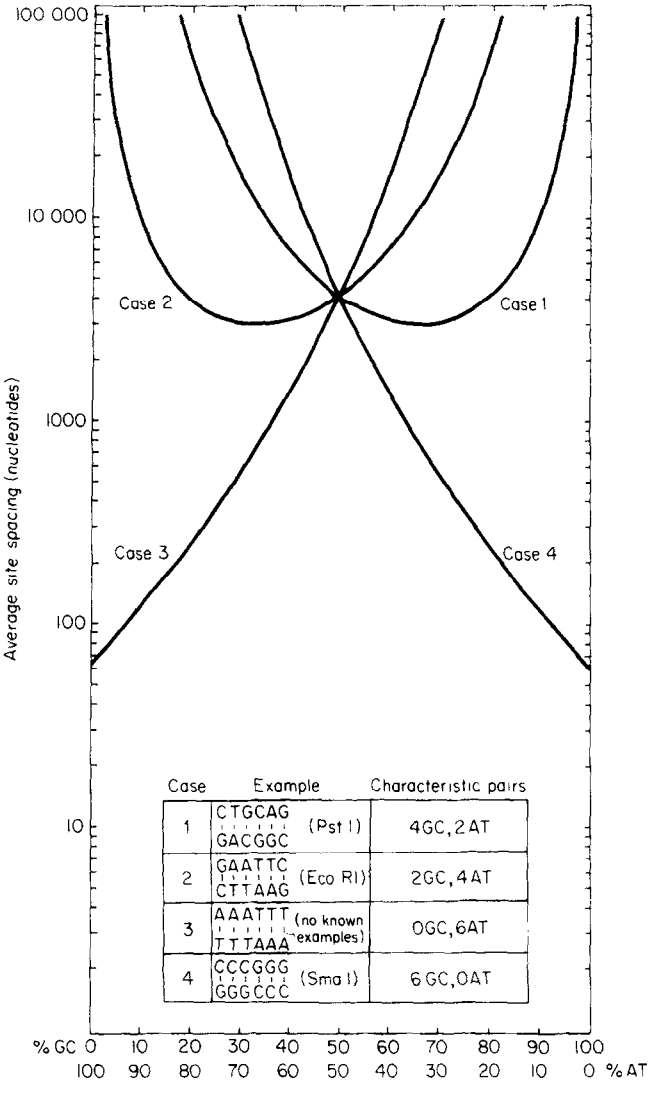

FIG. 2. Graphical representation of the equation for restriction enzyme recognition sequences consisting of six nucleotides. For these cases $m_{GC} + m_{AT} = 6$.

and may thus be identified. An example is the "Alu family" of repeats, which was first identified in the human genome because of the non-random abundance and spacing of recognition sites for the enzyme Alu 1 in human DNA (Houck, Rinehart & Schmid, 1979). Similarly, highly methylated DNA is cleaved in a distinctive, non-random fashion since some restriction enzymes recognize methylated nucleotides while others do not.

Non-random DNA sequences can be identified by lack of conformity of experimental data to the spacing predictions of Figs 1 and 2. This may explain (or predict) physical characteristics of particular DNAs. An example would be a relatively short, but highly G-C rich region which stabilizes a long molecule of DNA. This would result in elevation of the temperature at which the strands separate beyond the melting point predicted by the overall G-C content. Non-random features of DNA sequence such as those described above may, or may not, be of biological importance.

In cases where the G-C/A-T content is not known, it can be estimated from experimentally derived site spacing data using the curves of Figs 1 and 2. As a hypothetical example, consider a 2000 nucleotide long fragment of DNA experimentally shown to cleave at six positions with Alu 1, 22 positions with Eco RI* and twice using Hpa II. Applying the curve of Fig. 1, it is evident that the fragment is about 35% G-C and about 65% A-T. The G-C/A-T content can also be experimentally determined by study of various physical properties of DNA but the method described above provides a rapid alternative. Clearly the more enzymes whose cleavage positions are mapped, the greater the accuracy of the estimate.

The curves shown here are useful in a variety of ways. They can be used to assess the randomness of DNA sequence, to identify unusual regions such as those rich in A-T or G-C, to help locate repeated DNA sequences and to provide an estimate of the G-C/A-T content. They can also help to assess which enzymes should be utilized in a restriction analysis to yield a manageable number of DNA fragments. We present these curves as a simplifying tool for those involved in related studies.

Division of Biological Sciences,
The University of Michigan,
Ann Arbor, Michigan 48109, *U.S.A.*

GORDON P. MOORE†

R.C.A. Laboratories,
Princeton, New Jersey 08540, *U.S.A.*

ARNOLD R. MOORE

(*Received* 27 *May* 1981, *and in revised form* 23 *January* 1982, *and in final form* 6 *February* 1982)

REFERENCES

GOTOH, O., HAYASHI, J., YONEKAWA, H. & TAGASHIRA, Y. (1979). *J. mol. Evol.* **14,** 301.
HOUCK, C. M., REINHART, F. P. & SCHMID, C. W. (1979). *J. mol. Biol.* **132,** 289.
NEI, M. & LI, W. (1979). *Proc. natn. Acad. Sci. U.S.A.* **76,** 5269.
UPHOLT, W. B. (1977). *Nucleic Acids Res.* **4,** 1257.
UPHOLT, W. B. & DAWID, I. B. (1977). *Cell* **11,** 571.

† To whom reprint requests should be sent.