The Specificity of Deoxyribonucleases and their Use in Nucleotide Sequence Studies

GIORGIO BERNARDI, STANISLAV D. EHRLICH* & JEAN-PAUL THIERY

Laboratoire de Génétique Moléculaire, Institut de Biologie Moléculaire, Paris 5°, France

DNases split specific sets of short nucleotide sequences and can be used to assess the frequency of such sequences in DNA.

WE present here recent findings from our laboratory showing that deoxyribonucleases (DNases) hydrolyse specific sets of short nucleotide sequences, and that they can be used to obtain information on the frequency of such sequences. As shown in Table 1, which summarises some different approaches to the study of DNA sequences, the new method is a frequency method, and is basically different both from the indirect method relying on sequence dependent properties and from the direct sequence methods.

To study the specificity of DNases, it is necessary to determine the composition of the nucleotides adjacent to the breaks introduced by the enzymes. The methods we have used for the isolation and analysis of the termini $WX \downarrow YZ$ (the sequence being written in the usual 5′→3′ direction and the vertical arrow indicating the position of the breaks) have been described elsewhere[1]. They can be outlined as follows: for the 3′ terminal nucleotide (X)[2], oligonucleotides are first dephosphorylated with spleen phosphatase B[3], then hydrolysed with spleen exonuclease[4], an enzyme which releases 3′ nucleotides starting from the 5′ end; the 3′ terminal nucleotide is released as a nucleoside, and can be separated from the 3′ nucleotides formed. For 5′ terminal and penultimate nucleotides (YZ)[5], oligonucleotides are dephosphorylated, degraded with pancreatic DNase in order to decrease their average chain length, and digested with snake venom exonuclease;

* Present address: Department of Genetics, Stanford University Medical School, Stanford, California 94305.

Table 1 Methods for Studying Nucleotide Sequences in DNAs*

A, Indirect methods
- (1) Renaturation kinetics[22,23]
- (2) Cyclization after exonuclease treatment[24]
- (3) Sequence-dependent properties
  - (a) Circular dichroism[25]
  - (b) Silver binding[26,27]

B, Direct methods
- (1) Frequency methods
  - (a) Depurination[28]
  - (b) Nearest neighbour analysis[29,30]
  - (c) Analysis of termini released by DNases †
- (2) Sequence methods
  - (a) DNA polymerase repair[31]
  - (b) RNA polymerase copy[32]
  - (c) Direct sequencing[33]

* This table is by no means exhaustive; it provides examples of different approaches.
† Present work.

this enzyme splits off 5′ nucleotides starting from the 3′ end; 5′ terminal dinucleoside monophosphates, being very resistant to digestion by venom exonuclease because they lack the 5′ phosphate, accumulate in the digestion mixture up to levels higher than 90% of the theoretical yield and can be separated from the 5′ nucleotides formed; they are then split with spleen exonuclease into 3′ nucleotides (corresponding to the 5′ terminals, Y) and nucleosides (corresponding to the 5′ penultimates, Z), which can be separated from each other. For 3′ penultimate nucleotides (W), the two methods we have developed[6] are more complex and have not been routinely used; they will not be described here. The termini isolated by the above procedures have been analysed as nucleosides by one of the micromethods developed in our laboratory[7–10], the error being of the order of 1% for the 3′ terminal and 2% for the 5′ terminal and penultimate nucleotides.

This methodology can be applied to both 3′ and 5′ phosphate-ended oligonucleotides. It relies heavily on the quality of the isolation and analytical techniques as well as on the purity of the enzymes used, particularly of the phosphatase and exonucleases; failure to satisfy very exacting criteria has previously led to unreliable results and is responsible to a great extent for the view that DNases have no specificity towards nucleotide sequences.

Specificity of DNases

Four DNases have been investigated so far: acid DNase B from hog spleen[11], acid DNase from the hepatopancreas of the snail *Helix aspersa*, Müll[12], bovine pancreatic DNase, and *Escherichia coli* endonuclease I[13]. The data obtained for the composition of the termini released from calf thymus DNA are shown in Table 2. It can be seen that the base composition of each terminus differs with the enzyme which releases it; it is never equal to the base composition of the DNA, a finding ruling out the possibility that enzymatic degradation is random, since if this was the case all termini should have the base composition of the DNA. The possibility that some termini have a base composition different from that of DNA simply because they are the nearest neighbours of termini specifically recognised by the enzymes can be checked by comparing the experimental results with those expected from the nearest neighbour data. This type of control (Table 3) has shown that all termini released have a composition which differs from that expected from the nearest neighbour data and are indeed recognised. The only exception is given by the 5′ penultimate nucleotides released by the snail enzyme; these have the composition expected for the nearest neighbours of the 5′ terminal nucleotides and are therefore not recognised[14].

On the basis of the results just mentioned, we can conclude that the minimum length of the nucleotide sequences recognised by the enzymes investigated is 4 for spleen acid DNase (in which case the 3′ penultimate position was also analysed), 2 for snail DNase, and 3 for pancreatic DNase and *E. coli* endonuclease I. It is likely that the actual length of the recognised sequences is not much greater than the minimum length for two reasons: first, the presence in the final digests of oligonucleotides as short as di- and/or trinucleotides suggests that the shortest sequences which can be split are four to six nucleotides long; second, the enzymes under consideration have small molecular weights ranging from 38,000 for spleen acid DNase[15], to 30,000 for snail acid DNase[13], 31,000 for pancreatic DNase[16], and about 24,000 for *E. coli* endonuclease I[17], and it is unlikely that they make contact with long nucleotide sequences as do restriction enzymes which have much higher molecular weights.

The conclusion that the DNases investigated here have a sequence specificity, as opposed to the single-base specificity of RNases, is very probably of general significance. It is well known that restriction enzymes are also sequence specific; the main differences with the DNases considered here are the greater length of the recognised sequence and the much higher specificity of the restriction enzymes.

The results of Table 2 show that the DNases tested here split specific sets of nucleotide sequences and that these sets are different for each enzyme. A rough estimate of the number of sequences forming the specific sets recognised by each enzyme can be obtained by considering the average chain length of the final DNase digests. In the case of calf thymus DNA the average chain length comprises between 2 (pancreatic DNase in the presence of Mn^{2+}) and 4.5 (spleen acid DNase[18]), indicating therefore that the number of sequences which can be split by these enzymes is in the range of 50% to 20% of all sequences. This estimate is correct only if all sequences are equally frequent; since this is not the case, the percentage of sequences which can be split is, in fact, higher or lower than that indicated above, but very probably not too far from it. The percentages of sequences which can be split may, however, be higher than those just mentioned since the average size of the final digest might be lower than those indicated above, at least for some enzymes, if the inhibitory effect of the reaction products could be completely eliminated; in addition, susceptible sequences are likely to overlap to some extent and breaks introduced by the enzymes may hinder hydrolysis of neighbouring phosphodiester bonds. It seems safe to conclude, therefore, that the number of sequences split by the enzymes is so large that the

Table 2 Termini Liberated from Calf Thymus DNA by Four DNases

		3′ Terminals	5′ Terminals	5′ Penultimates
Spleen DNase*	T	20	11	14
	G	43	43	26
	A	29	18	52
	C	8	28	8
Snail DNase	T	16	14	38
	G	6	45	24
	A	78	10	21
	C	1	31	17
Pancreatic DNase†	T	36	38	13
	G	15	22	36
	A	31	15	30
	C	18	25	21
E. coli endonuclease I	T	41	24	28
	G	8	35	29
	A	35	17	29
	C	16	23	14

* In this case, the average chain length of oligonucleotides was equal to fifteen. 3′ penultimate nucleotides were also analysed; they were T 22%, G 16%, A 46%, C 16%.
† Digestion was carried out in the presence of Mg^{2+}. Results obtained in the presence of Mn^{2+} are very slightly different. Experimental details are given elsewhere[10,34–36].

Table 3 Average Composition of Sequences Split by Acid DNase in Calf Thymus DNA

	3′ P penultimate		3′ P terminal			5′ OH terminal			5′ OH penultimate	
	Found	Calculated	Calculated	Found	Calculated	Calculated	Found	Calculated	Calculated	Found
T	22	(31)	(29)	20	(32)	(29)	11	(29)	(29)	14
G	16	(21)	(22)	43	(21)	(23)	43	(23)	(19)	26
A	46	(30)	(29)	29	(30)	(29)	18	(30)	(31)	52
C	16	(19)	(20)	8	(17)	(18)	28	(18)	(20)	8

Values in parentheses indicate the composition of each terminus as calculated from its nearest neighbour(s). For the terminal positions calculated values could be obtained from both neighbouring positions.

* Average chain length of oligonucleotides was 15.

sequence sets recognised by them overlap with each other to some extent. Yet the fact that the composition of termini released by the four DNases differs according to the enzyme used indicates that the sets of sequences recognised still largely differ for different enzymes.

The data of Table 2 represent the apparent average base composition of all the sequences split by the enzymes. In calf thymus this is related to the actual average base composition of all sequences split through the set of K_M and/or V_{max} values associated with the split sequences. As an example of the effect of such differences in the K_M and/or V_{max}, we can quote results obtained with poly(dA-dT)·poly(dA-dT). This polymer, which contains equal amounts of the tetranucleotides ATAT and TATA, releases, upon spleen acid DNase degradation, 3′ terminal A and T in a ratio 4:1, thus indicating that the two sequences are split with different K_M and/or V_{max}. The relative importance of these two factors is unknown so far, but the finding that the affinity of all DNases used (except for the pancreatic enzyme) for polyribonucleotides is comparable with, or higher than, that for DNA[12,14,19], in spite of the large structural differences between these polymers and DNA, suggests that K_M values for different sequences may not differ very much from each other and that the main factor is V_{max}. It is important to note that differences in K_M and/or V_{max} for the two sequences present in poly(dA-dT)·poly(dA-dT) do not lead to a variation in the percentage of terminal nucleotides released in the average chain length range 40 to 15, indicating that both sequences are still present at saturating levels (at the substrate concentrations used) when as much as 7% of all sequences have been split.

The lack of variation in the relative amounts of different termini as released by different enzymes seems to be the rule, and reflects the fact that termini derive from a very large number of sequences and are therefore associated with average K_M and V_{max} values. An additional reason may be that the majority of sequences have K_M and V_{max} values in a narrow range. In the range of average oligonucleotide chain length 50 to 8, no change has been seen in the composition of any of the termini released from calf thymus DNA by any of the DNases investigated, with the single exception of the 3′ terminals released by spleen acid DNase.

Another implication of the invariance of the composition of termini in the range of average chain length, 50 to 88, of

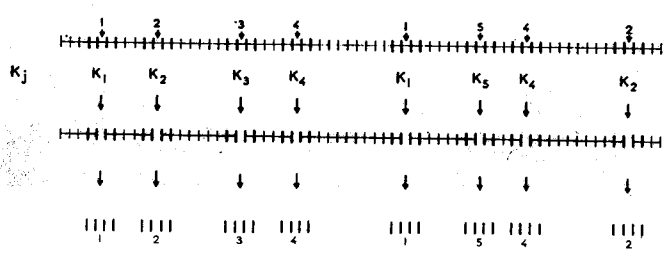

Fig. 1 Analysis of termini: a number of sequences, indicated as tetranucleotides and numbered 1 to 5, are recognised and split with different K_M and/or V_{max}, indicated by K_1, K_2, etc. Termini are isolated from the resulting oligonucleotides, and the base compositions of termini, WXYZ, are determined.

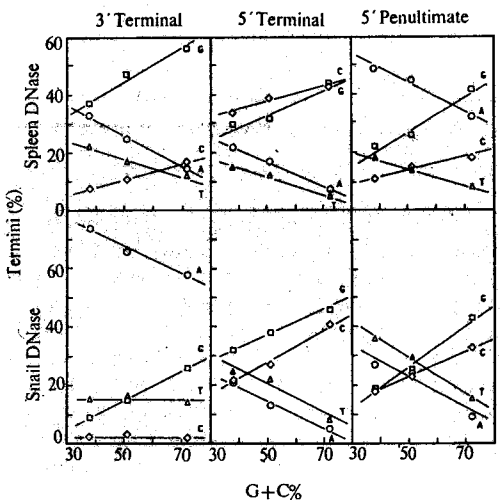

Fig. 2 The percentages of A (○), G (□), C (◇) and T (△) in the 3′-terminal, 5′-terminal and 5′-penultimate nucleotides formed by the spleen and the snail DNase from bacterial DNAs (*Haemophilus influenzae*, 38% G+C; *Escherichia coli*, 51% G+C; *Micrococcus luteus*, 72% G+C), are plotted against the G+C content of DNAs. Values obtained at an average chain length of fifteen nucleotides were used.

the digests is that the specificity of the DNases used is not affected by the secondary structure of the substrate. In fact, melting of double-stranded DNA fragments takes place in the average range of chain length explored, under the experimental conditions used; this causes a striking slowing down of the reaction rate, due to the lower affinity of the enzymes for single stranded than for double stranded DNA, but no change in the composition of termini. The special case of spleen acid DNase is discussed elsewhere[10].

Analysis of Termini

Since DNases split specific sets of sequences, the analysis of termini provides information on the frequency of these sequences in a given DNA. In fact, the composition of termini is related, first, to the average composition of the sequences which can be split by the enzymes; second, to the K_M and V_{max} values associated with each sequence; and third, to their relative amount in the DNA under consideration (Fig. 1).

The easiest way to check the latter effect is to determine the composition of termini obtained from DNAs having different G+C contents and, therefore, different relative amounts of various sequences. Indeed, the compositions of termini released by bacterial DNAs having different G+C contents are different. Furthermore, if the percentages of A, G, C and T in the termini of bacterial DNAs are plotted against their G+C contents, linear relationships are obtained (Fig. 2). The choice of bacterial DNAs for looking at the effect of changes in the relative amount of sequences on the composition of termini is based on the absence of repetitive sequences in bacterial DNAs and on the fact that the doublet frequencies of

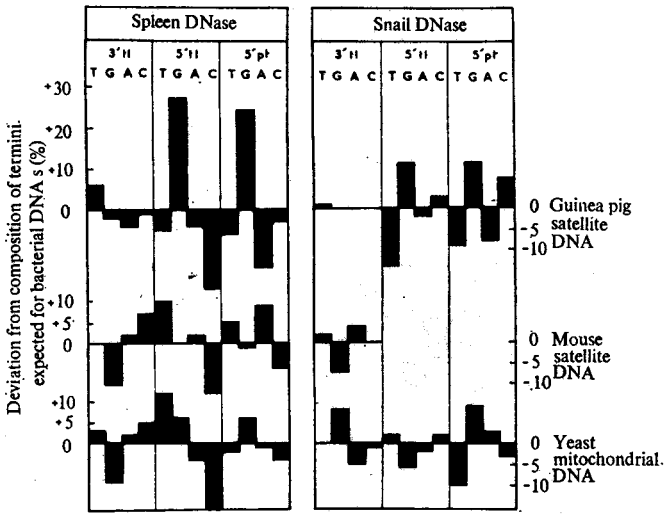

Fig. 3 Deviation patterns of three repetitive DNAs. The histograms show the differences between the composition of termini formed from guinea pig satellite, mouse satellite and yeast mitochondrial DNAs by spleen and snail DNases and the compositions expected for bacterial DNAs having the same G+C contents; tl, terminal; pt, penultimate.

bacterial DNAs, as determined by the nearest neighbour analysis, show essentially linear relationships with the frequencies predicted for random association[20], indicating a common type of doublet distribution in these DNAs.

As expected, the composition of termini released from DNAs containing repetitive nucleotide sequences deviates, in either direction, from the linear relationship obtained with non-repetitive (bacterial) DNAs. The deviation patterns which can thus be obtained represent a novel way of characterising repetitive DNAs or, more generally, DNAs having sequence distributions different from those of the bacterial DNAs examined here. Figure 3 shows the deviation patterns of three DNAs containing short repeated sequences: the satellite DNAs from mouse and guinea pig and the mitochondrial DNA from yeast. Expectedly, deviation patterns obtained with different enzymes on the same DNAs are different from each other, as are deviation patterns obtained with the same enzyme on different DNAs.

As another example, Fig. 4 shows deviation patterns obtained with calf, mouse and guinea pig DNAs and with yeast nuclear DNA. It is evident that in this case, too, a number of sequences are present in excess or are deficient in the eukaryotic DNAs compared with bacterial DNAs of identical G+C composition. Interestingly, the mammalian

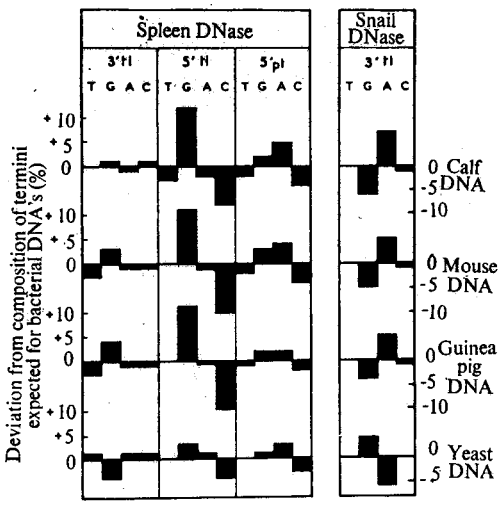

Fig. 4 Deviation patterns of four eukaryotic DNAs. In the case of yeast DNA a nuclear DNA preparation free of mitochondrial DNA was used. See legend of Fig. 3 for the presentation of data.

DNAs have similar deviation patterns whereas the yeast nuclear DNA pattern is quite different; mammalian DNAs seem to share the sequence features which are responsible for the similarity of their deviation patterns and which do not exist in yeast nuclear DNA. The possibility that the deviations observed in the mammalian DNAs arise from their satellite DNAs is ruled out by the completely different deviation patterns exhibited by the latter (Fig. 3). The meaning of the deviation patterns of eukaryotic DNAs will be discussed elsewhere.

In conclusion, the analysis of termini represents a new, direct approach to the study of nucleotide sequences in DNAs which is particularly useful when a comparison of sequences of different DNAs is needed. The method is rigorously analytical; our present ignorance of the K_M and V_{max} values associated with each sequence is irrelevant as far as the practical use of the method is concerned, because these values are simple proportionality constants, and because the analysis bears on termini derived from a number of different sequences. The method is also very flexible. Different enzymes with different specificities can be used to analyse different sets of sequences, thereby permitting the use of the most sensitive deviation patterns obtained; the use of several enzymes permits the analysis of the majority, if not all, of the nucleotide sequences in DNAs. In addition, the analysis can be limited to only one, or extended to several termini according to the degree of information required; 5′ terminal doublets can be analysed as such, yielding more information than obtained after splitting; methods for the labelling of 3′ terminal[2] and of 5′ terminal nucleotides (G. Bernardi, unpublished) have been developed. The main differences existing between the present method and another frequency method, nearest neighbour analysis, are that the latter measures the frequencies of all doublets, whereas the analysis of termini is based on the sequences selected by the enzyme used; and that the sequences recognised by the enzymes are longer than the dinucleotides investigated by the nearest neighbour analysis. More information can therefore be obtained from the analysis of termini than from nearest neighbour analysis.

This work would have been impossible without the previous contributions of Alberto Bernardi, M. Carrara, A. Chersi, C. Cordonnier, M. Griffé and H. Stebler.

Received July 9, 1973.

[1] Bernardi, G., Ehrlich, S. D., and Thiery, J. P., *Meth. Enzym.* (in the press).
[2] Carrara, M., and Bernardi, G., *Biochemistry*, **7**, 1121 (1968).
[3] Chersi, A., Bernardi, A., and Bernardi, G., *Biochim. biophys. Acta*, **246**, 51 (1971).
[4] Bernardi, A., Bernardi, G., *Biochim. biophys. Acta*, **155**, 360 (1968).
[5] Ehrlich, S. D., Torti, G., and Bernardi, G., *Biochemistry*, **10**, 2000 (1971).
[6] Devillers-Thiery, A., Ehrlich, S. D., and Bernardi, G., *Eur. J. Biochem.*, **38**, 416 (1973).
[7] Carrara, M., and Bernardi, G., *Biochim. biophys. Acta*, **155**, 1 (1968).
[8] Piperno, G., Bernardi, G., *Biochim. biophys. Acta*, **238**, 388 (1971).
[9] Ehrlich, S. D., Thiery, J. P., and Bernardi, G., *Biochim. biophys. Acta*, **246**, 161 (1971).
[10] Thiery, J. P., Ehrlich, S. D., Devillers-Thiery, A., and Bernardi, G., *Eur. J. Biochem.*, **38**, 434 (1973).
[11] Bernardi, G., and Griffé, M., *Biochemistry*, **3**, 1419 (1964).
[12] Laval, J., and Paoletti, C., *Biochemistry*, **11**, 3596 (1972).
[13] Lehman, I. R., Roussos, G. C., and Pratt, E. A., *J. biol. Chem.*, **237**, 819 (1962).
[14] Ehrlich, S. D., Devillers-Thiery, A., and Bernardi, G., *Eur. J. Biochem.* (in the press)
[15] Bernardi, G., Appella, E., Zito, R., *Biochemistry*, **4**, 1725 (1965).
[16] Lindberg, U., *Biochemistry*, **6**, 335 (1967).
[17] Cordonnier, C., and Bernardi, G., *Biochem. biophys. Res. Commun.*, **20**, 555 (1965).
[18] Soane, C., Thiery, J. P., Ehrlich, S. D., and Bernardi, G., *Eur. J. Biochem.*, **38**, 422 (1973).
[19] Bernardi, G., *Biochem. biophys. Res. Commun.*, **17**, 573 (1964).
[20] Baldwin, R. L., and Kaiser, A. D., *J. molec. Biol.*, **4**, 418 (1962).
[21] Bertazzoni, U., Ehrlich, S. D., and Bernardi, G., *Biochim. biophys. Acta*, **312**, 192 (1973).
[22] Britten, R. J., and Kohne, D. E., *Science, N.Y.*, **161**, 529 (1968).

[23] Wetmur, J. G., and Davidson, N., *J. molec. Biol.*, **31**, 349 (1968).
[24] Thomas, jun., C. A., Hamkalo, B. A., Misra, D. N., and Lee, C. S., *J. molec. Biol.*, **51**, 621 (1970).
[25] Allen, F. S., Gray, D. M., Roberts, G. D., and Tinoco, jun., I., *Biopolymers*, **11**, 853 (1972).
[26] Corneo, G., Ginelli, E., Soave, C., and Bernardi, G., *Biochemistry*, **7**, 4373 (1968).
[27] Filipski, J., Thiery, J. P., and Bernardi, G., *J. molec. Biol.* (in the press).
[28] Chargaff, E., in *The Nucleic Acids* (edit. by Chargaff, E., and Davidson, N.), **1**, 307 (Academic Press, London, New York, 1955).
[29] Josse, J., Kaiser, A. D., and Kornberg, A., *J. biol. Chem.*, **236**, 864 (1961).
[30] Swartz, M. N., Trautner, T. A., and Kornberg, A., *J. biol. Chem.*, **237**, 1961 (1962).
[31] Wu, R., and Taylor, E., *J. molec. Biol.*, **57**, 491 (1971).
[32] Takanami, M., Okamoto, T., and Sugiura, M., *Cold Spring Harb. Symp. quant. Biol.*, **35**, 179 (1970).
[33] Ziff, E. B., Sedat, J. W., and Galibert, F., *Nature new Biol.*, **241**, 3 (1973).
[34] Laval, J., Thiery, J. P., Ehrlich, S. D., Paoletti, C., and Bernardi, G., *Eur. J. Biochem.* (in the press).
[35] Ehrlich, S. D., Bertazzoni, U., and Bernardi, G., *Eur. J. Biochem.* (in the press).
[36] Ehrlich, S. D., Bertazzoni, U., and Bernardi, G., *Eur. J. Biochem.* (in the press).