Volume 13 Number 12 1985 Nucleic Acids Research

Structural junctions in DNA: the influence of flanking sequence on nuclease digestion specificities

Horace R.Drew and Andrew A.Travers

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Received 29 April 1985; Revised and Accepted 24 May 1985

ABSTRACT

When a protein binds to DNA, the affinity of this protein for its primary site of interaction may be influenced by the nature of flanking sequences. This is thought to be a consequence of local cooperativity in the DNA molecule, where the conformation at one point along the helix can influence the conformation at another, and thereby modulate the free energy of protein-DNA recognition.

In order to learn more about this process, we have carried out experiments of two sorts. First, we have constructed sequences of the type $(dA)_{11}(dG)_8$, where the conformational preferences of the DNA molecule switch from one extreme to another over just a single base pair, and subjected them to digestion by DNAase I and DNAase II. This is to learn whether the structure changes abruptly at the junction point, or more gradually with an influence extending into residues on either side. Secondly, we have subjected long plasmid DNA to digestion by restriction enzymes *Fnu* DII, *Hae* III, *Hha* I and *Msp* I, to look for correlations between cutting rate and the identity of nucleotides on either side of the restriction site.

The influence of flanking sequence on nuclease digestion specificities is clearly evident in both kinds of experiment, but the rules governing this seem complex and not easily formulated. The best that can be done at present is to divide the problem into two parts, "analogue" and "digital", representing sugar-phosphate and base components of recognition.

INTRODUCTION

Following the suggestion of Klug *et al*. (1) that enzymes such as DNAase I might be profitably used to study DNA structure in solution, we have been investigating various aspects of this problem.

The initial work was carried out on simple sequences such as poly(dA-dT) and poly(dG-dC), and provided evidence that the helical conformation of DNA could vary locally, through alteration in phosphate torsion angle, in a way recognizable to DNAase I (2). A more extensive analysis, using DNA sequences of both synthetic and natural origin, showed that the structure could vary globally as well: runs of A plus T, and likewise runs of G plus C, adopt helical conformations different from that of mixed-sequence DNA; and these conformations are recognizable to a wide variety of nucleases (3). The few

© IRL Press Limited, Oxford, England.

single-crystal x-ray studies which have been done (on short DNA molecules of defined sequence) corroborate the digestion data, and further show that the helical conformations of AT-rich and GC-rich regions are not equivalent: the double-helical minor groove is very narrow in the former but very wide in the latter (4,5).

After some reflection, it became apparent to us that the juxtaposition of such AT-rich and GC-rich sequences in a long DNA molecule would produce a structural discontinuity. Over the course of just a single base pair, the preferred spacing across the minor groove would have to change from narrow to wide. The question then arose as to whether this structural change need be abrupt or, as seems more likely, whether it might be spread over several bonds on either side of the junction. Any sort of delocalized transition would be of considerable interest, since this would provide a mechanism whereby the affinity of a protein for its primary site of interaction could be modulated by sequences on either side.

In order to investigate this possibility further, we thought it might be useful to construct several AT/GC junctions and then determine their structure in solution, through use of the enzymes DNAase I and DNAase II. These enzymes are particularly appropiate for such a purpose, since the rate at which they cut a bond depends primarily on the conformation observed rather than the base sequence per se. DNAase I (Figure 1a) recognizes the spacing between sugar-phosphate chains as measured across the minor groove (3,6,7). It cuts very poorly at either end of a helix where one chain protrudes beyond the other, marginally well where the spacing is especially narrow or wide due to sequence or environmental effects, and very well where the spacing is intermediate between the two extremes. DNAase II (Figure 1b) recognizes some aspect of single-strand conformation, although the mechanics of this are not well-understood (3,6). Sometimes it cuts on just one strand of the helix and other times on both strands; for many indirect reasons we think that it prefers to cut when a series of phosphates lie close together in space.

The model for DNAase I cleavage has already proved useful in understanding how antibiotics such as echinomycin (8) and distamycin (9) influence the helical conformation of DNA in regions adjoining their binding sites. Echinomycin (Figure 2a) unwinds the helix by 48° so as to open minor grooves nearby; for sequences of the type (dA).(dT) which prefer a narrow groove in the absence of antibiotic (4), this opening of the groove to an intermediate state encourages DNAase I to cut more rapidly (10-12). By way of contrast, distamycin (Figure 2b) lies deep within a narrow minor groove and

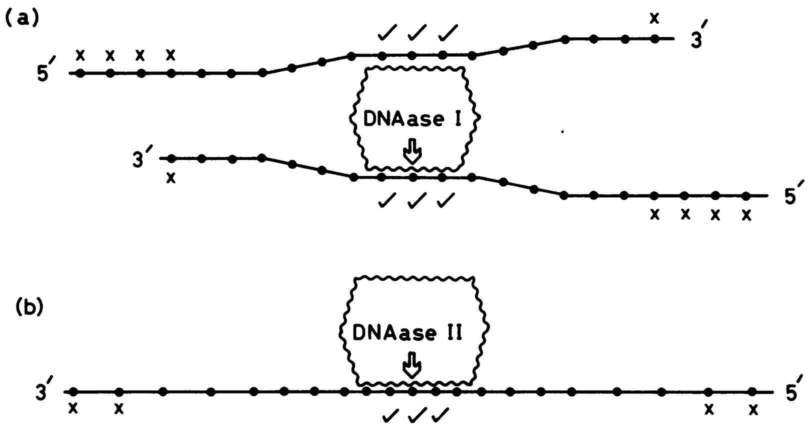


Figure 1. Models for the specificity of: (a) DNAase I and (b) DNAase II. Each of the small dots represents a phosphate group, and the lines connecting them represent a series of connecting bonds in the deoxyribose sugar. The surface of the protein in contact with the DNA is shown in wavy lines, with an open arrow indicating the protein cutting-function. Checks identify phosphates which are especially sensitive to cleavage, while crosses identify phosphates which are especially resistant to cleavage.

makes van der Waals contact with sugar-phosphate chains on either side; for sequences of the type (dG).(dC) which prefer a wide groove in the absence of antibiotic (5), this narrowing of the groove to an intermediate state similarly encourages DNAase I to cut more rapidly (11).

Here we probe the nature of the structural junction between AT-rich and GC-rich sequences by use of the enzymes DNAase I and DNAase II. In the case of DNAase I, an abrupt change in conformation (e.g., a "kink") should disrupt the structure of the DNA at the junction so as to reduce the rate of cleavage; whereas a gradual change in conformation should cause the minor groove to become intermediate in size, and therefore more amenable to cleavage (Figure 2c). We also study four restriction enzymes: Fnu DII, Hae III, Hha I and Msp I, to see how their rates of cleavage are influenced by flanking sequence.

MATERIALS AND METHODS

Synthesis, Purification and Labelling

The 20-base-pair double helix shown in Figures 3 and 4 was assembled in the following way: two strands $GA_{11}G_8$ and $C_8T_{11}C$ were synthesized by an automated phosphotriester method, and purified on 20% polyacrylamide gels containing 7 M urea. One of the two strands, depending upon the experiment, was then labelled at its 5'-end using polynucleotide kinase and [γ-^{32}P] ATP,

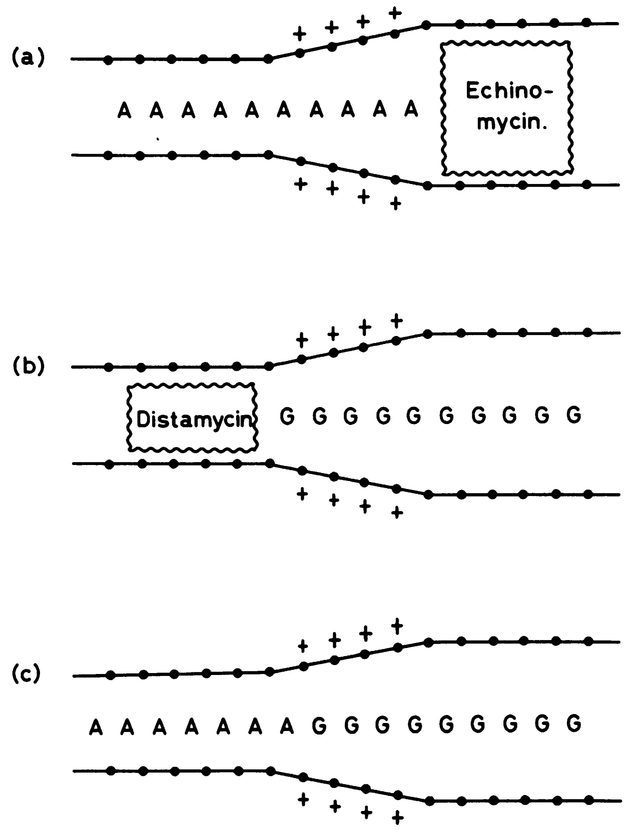


Figure 2. Variations in DNAase I cleavage specificity induced by: (a) the binding of echinomycin; (b) the binding of distamycin; or (c) a continuous AT/GC junction. The width of the minor groove is postulated to be a key feature in controlling DNAase I specificity. Although no rigorous formula to calculate this parameter is yet available, we have been using an approximate relation of the form: groove width (Å) = 16.5 + 0.33 (roll - propellor twist), where "roll" and "propellor twist" are in degrees and must be averaged over several steps to reach a mean value.

purified on a 20% gel containing no urea, and mixed with a 2:1 molar excess of non-radioactive partner strand to form a double helix. The change in conformation from single strand to double helix could be monitored by a reduction of mobility on nondenaturing polyacrylamide gels, run at low voltage to minimize heating. By electrophoretic criteria, this synthetic duplex begins to melt at 40°C.

The sequences shown in Figures 6b and 6c were constructed by ligation of this 20-mer duplex into the Sma I site of plasmid pUC13. Three recombinant plasmids ("S3", "S4" and "S6") were isolated from E. coli, cut with Hae II and Hind III to yield a mixture of 230-mer ligation product and long plasmid DNA, then applied to a 6% gel. The 230-mers were isolated from the gel, labelled at their 5'-ends using polynucleotide kinase and [γ-^{32}P] ATP, cut with Eco RI to yield a mixture of large (145-165) and small (65-85) fragments, then applied to a 10% gel. In each case the small fragment, containing the original site of ligation, was isolated from the gel and sequenced by the Maxam-Gilbert procedure. Clones S3 and S6 were found to be identical in sequence, each containing one copy of a $C_8T_{11}C$ insert. Clone S4 was found to contain a two-copy $C_8T_{11}C_9T_4AT_6C$ insert. This dimer differs slightly from the original sequence by a T-to-A transversion in the second run of T_{11}.

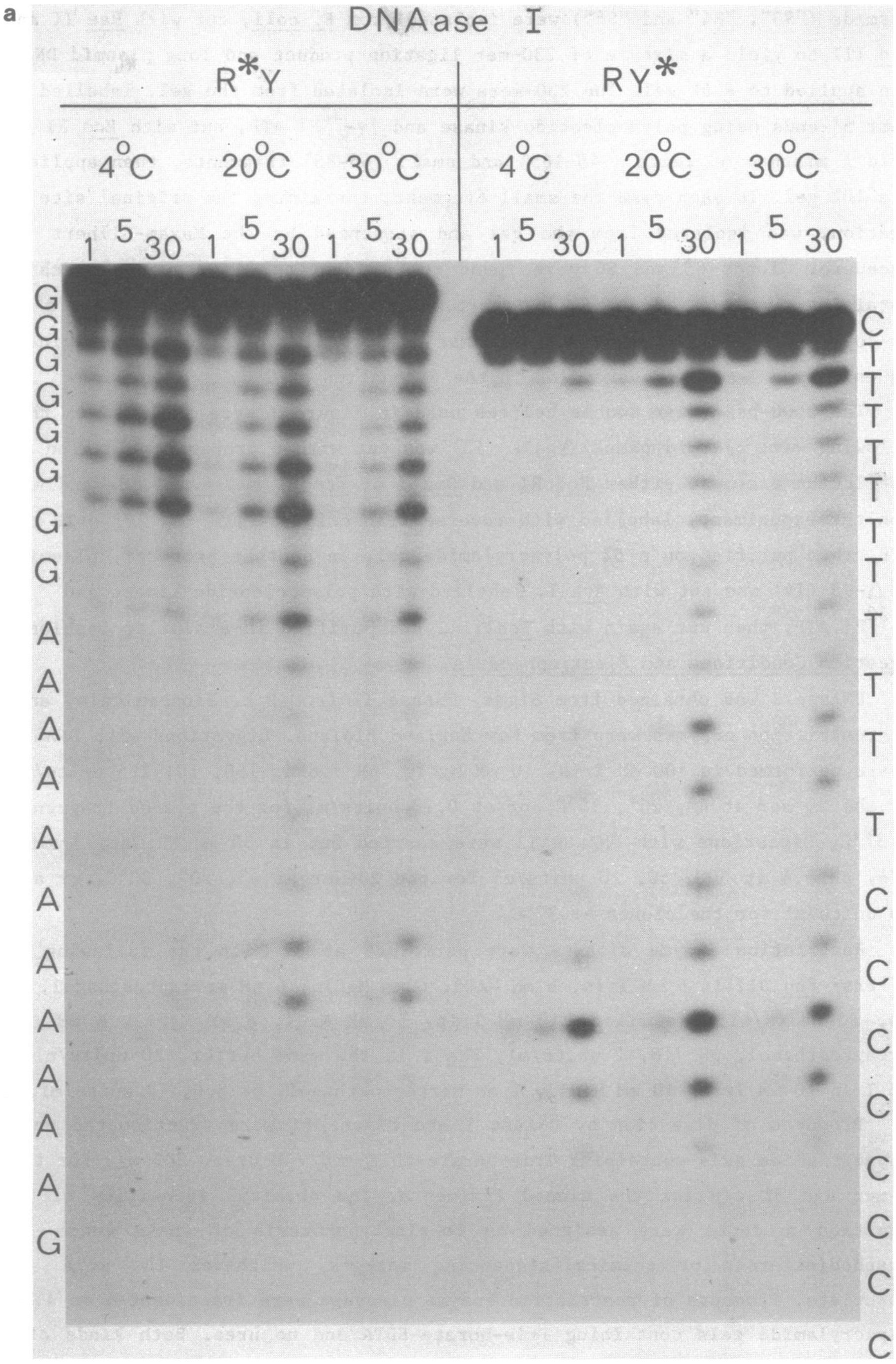
The 2000-base-pair double helices used in Figure 7 were isolated in the following way: plasmid pKMp27ΔgalK (13) was cut with a proper combination of restriction enzymes, either Eco RI and Sma I or Eco RI and Ava I, depending upon the experiment, labelled with reverse transcriptase and [α-^{32}P] dATP or dCTP, then purified on a 5% polyacrylamide gel. In another protocol, plasmid pKMΔ-98 (14) was cut with Ava I, labelled with polynucleotide kinase and [γ-^{32}P] ATP, then cut again with Tth111 I and purified in a similar fashion.

Digestion Conditions and Electrophoresis

DNAase I was obtained from Sigma, DNAase II from P.L. Biochemicals, and all restriction enzymes were from New England Biolabs. Digestions with DNAase I were performed in 100 mM Tris, 10 mM $MgCl_2$, pH 7.6 at 150, 10, 2.5 units/ml for the 20-mer at 4°, 20°, 30°C, or at 0.10 units/ml for the cloned fragments at 37°C. Digestions with DNAase II were carried out in 50 mM NH_4OAc, 1 mM EDTA, pH 5.4 at 500, 50, 10 units/ml for the 20-mer at 4°, 20°, 30°C, or at 0.5 units/ml for the clones at 37°C.

Restriction enzyme digests were performed at 37°C in the following buffers: Fnu DII in 6 mM Tris, 6 mM NaCl, 6 mM $MgCl_2$, 6 mM mercaptoethanol, pH 7.4, 10 units/ml; Hae III in 10 mM Tris, 50 mM NaCl, 6 mM $MgCl_2$, 6 mM mercaptoethanol, pH 7.8, 2 units/ml; Hha I in the same buffer, 20 units/ml; Msp I in 10 mM Tris, 10 mM $MgCl_2$, 1 mM mercaptoethanol, pH 8.0, 12 units/ml.

Products of digestion by DNAase I and DNAase II were fractionated on polyacrylamide gels containing Tris-borate-EDTA and 7 M urea, 20% w/v for the 20-mer and 9% w/v for the cloned fragments. The chemical identities of digestion products were assigned by co-electrophoresis of snake venom phosphodiesterase or chemical-sequencing markers, whichever the more appropiate. Products of restriction enzyme cleavage were fractionated on 4.5% polyacrylamide gels containing Tris-borate-EDTA and no urea. Both kinds of


a
DNAase I
R*Y
RY*
4°C
20°C
30°C
4°C
20°C
30°C
1 5 30 1 5 30 1 5 30
1 5 30 1 5 30 1 5 30
G
G
G
G
G
G
G
G
A
A
A
A
A
A
A
A
A
A
A
G
C
T
T
T
T
T
T
T
T
T
T
T
C
C
C
C
C
C
C
C

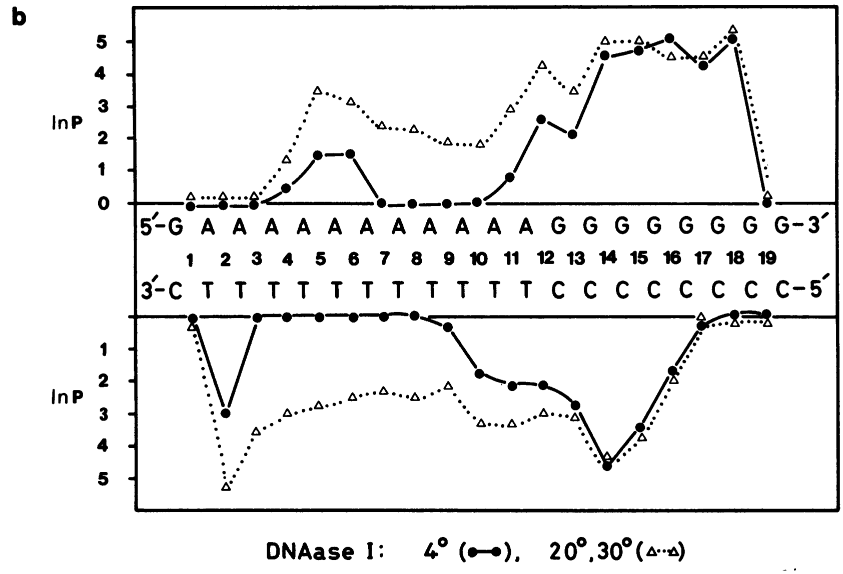


Figure 3. Digestion of the synthetic 20-mer by DNAase I: (a) electrophoretic profiles of digestion for the purine R*Y or pyrimidine RY* strand, with times 1, 5, 30 in minutes; (b) the corresponding probability plot. Probabilities of cleavage for upper and lower strands appear staggered in the 3'-direction by three bonds, and this is because phosphates which face one another across the minor groove lie displaced in a similar fashion.

low-percentage gel (9% and 4.5%) were fixed in 10% acetic acid, transferred to Whatman 3MM paper, dried under vacuum and subjected to autoradiography with an intensifying screen for 6-24 hours.

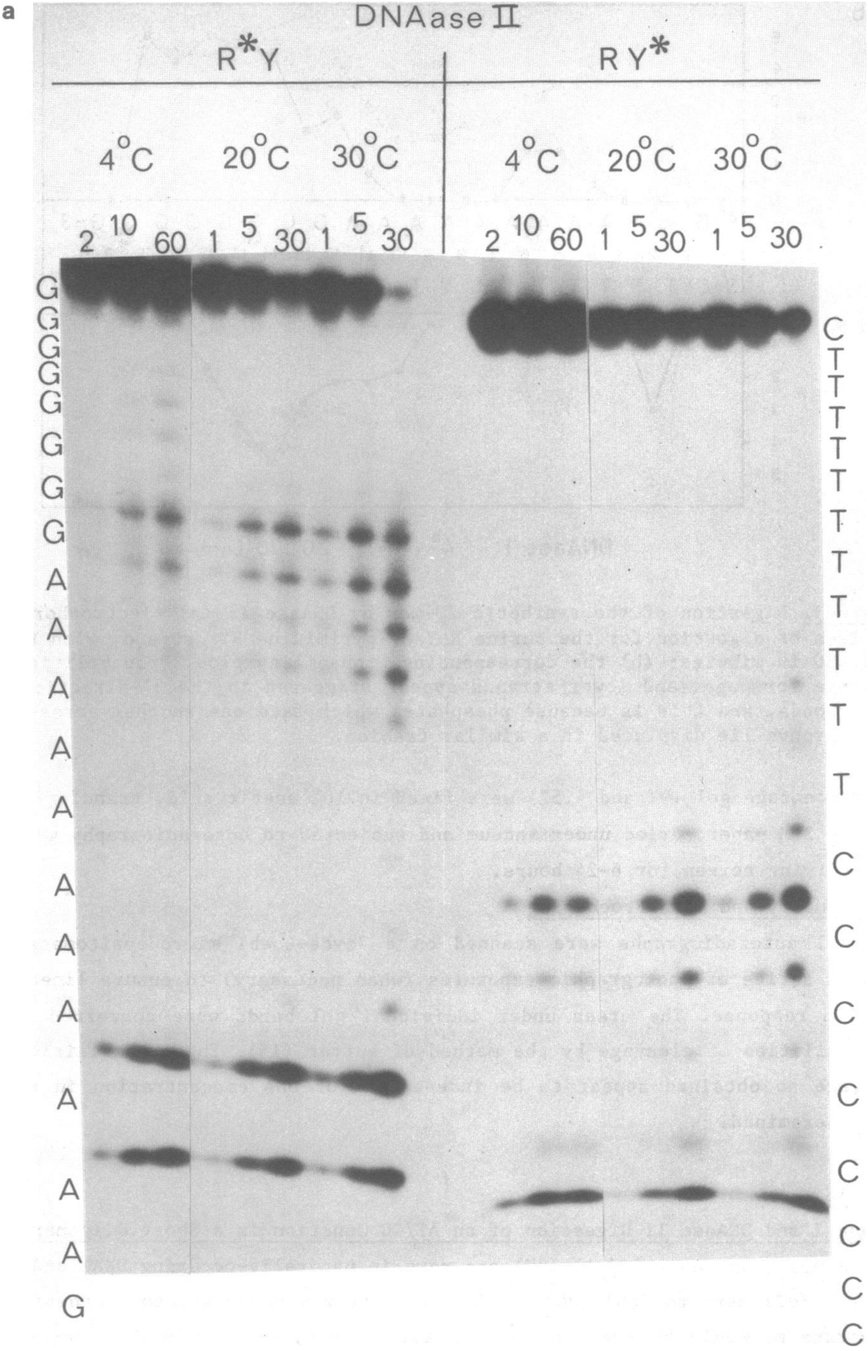
Densitometry and Data Processing

All autoradiographs were scanned on a Joyce-Loebl microdensitometer, using a series of photographic exposures (when necessary) to ensure linearity of film response. The areas under individual gel bands were converted to probablilities of cleavage by the method of Lutter (15). The probabilities of cleavage so obtained appear to be independent of DNA concentration in the region examined.

RESULTS

DNAase I and DNAase II Digestion of an AT/GC Junction in a Short Oligomer

Since long runs of (dA).(dT) are rare in naturally-occuring DNA, and runs of (dA).(dT) next to (dG).(dC) doubly rare, it was necessary to synthesize a substrate molecule by chemical means. Accordingly, we made a 20-base-pair double helix of sequence GA$_{11}$G$_8$, with the initial guanine intended to


a
DNAase II
R*Y
RY*
4°C
20°C
30°C
4°C
20°C
30°C
2
10
60
1
5
30
1
5
30
2
10
60
1
5
30
1
5
30
G
G
G
G
G
G
G
G
A
A
A
A
A
A
A
A
A
A
A
G
C
T
T
T
T
T
T
T
T
T
T
T
C
C
C
C
C
C
C
C

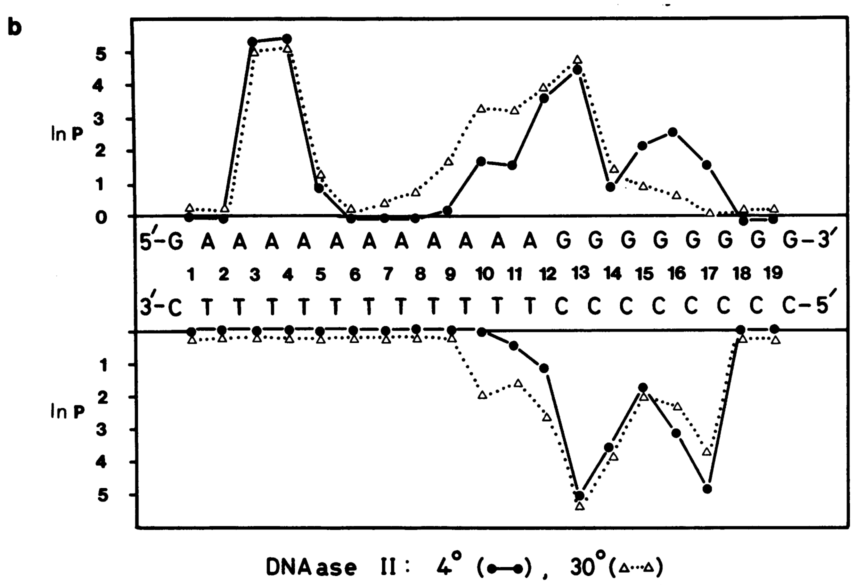
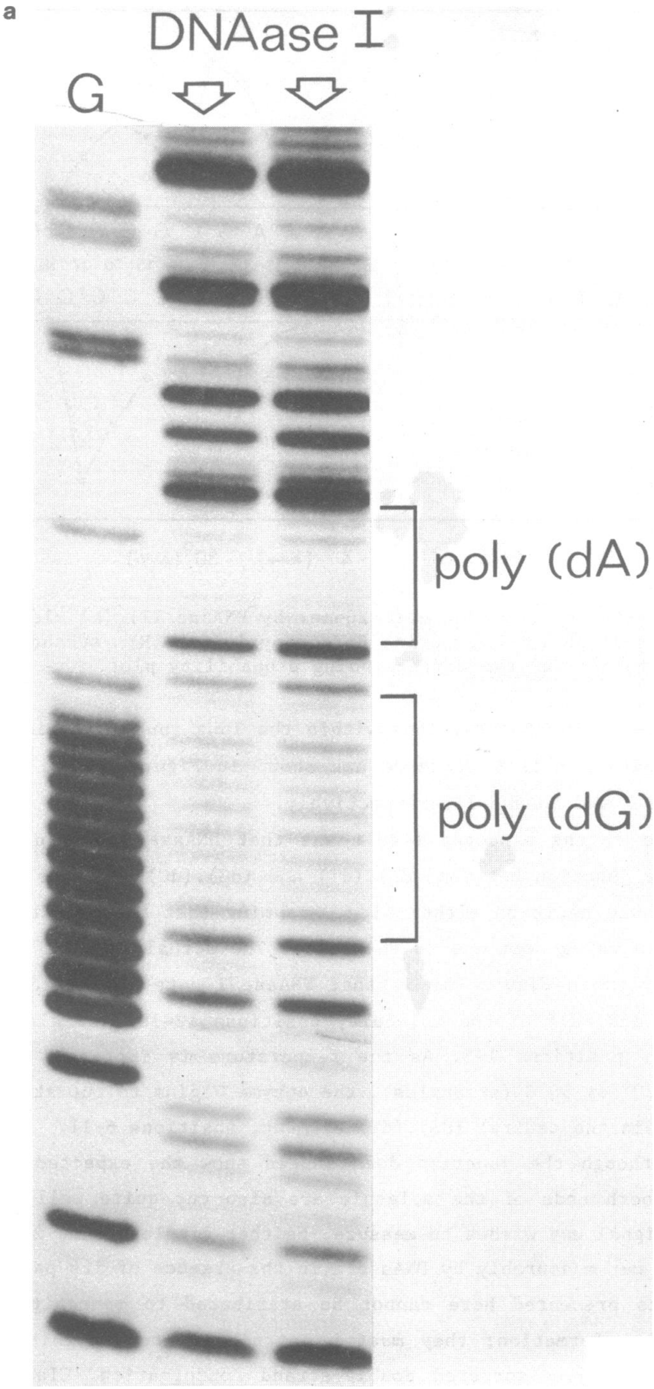


Figure 4. Digestion of the synthetic 20-mer by DNAase II: (a) electrophoretic profiles of digestion for the purine R*Y or pyrimidine RY* strand, with times 1, 5, 30 in minutes; (b) the corresponding probability plot.

stabilize base-pairing interactions within the long run of adenine residues. Digestion results for this sequence are shown in Figures 3 and 4 for the enzymes DNAase I and DNAase II, respectively.

In Figure 3, the expected result was that DNAase I should cut more rapidly at the junction between (dA).(dT) and (dG).(dC) segments than in the regions of double helix on either side, assuming that the structure changes continuously in going from one to the other. The actual result, as indicated by filled circles in Figure 3b, is that DNAase I cuts well throughout the entire right-hand half of the molecule, positions 12-18, and also near the left-hand end, positions 1-5. As the temperature is increased from 4°C (circles) to 20° or 30°C (triangles), the enzyme begins to cut at a measurable rate even within the central (dA).(dT) segment, positions 6-11.

Thus, although the junction does indeed show the expected DNAase I sensitivity, both ends of the molecule are also cut quite well so as to obscure the signal one wishes to measure. Neither single-strand 20-mer, GA$_{11}$G$_8$ or C$_8$T$_{11}$C, is cut measurably by DNAase I in the absence of its partner strand, so the results presented here cannot be attributed to recognition of a single-strand conformation; they must be attributed to recognition of an altered or thermally-disordered double-strand conformation. Clearly it is


a
DNAase I
G
poly (dA)
poly (dG)

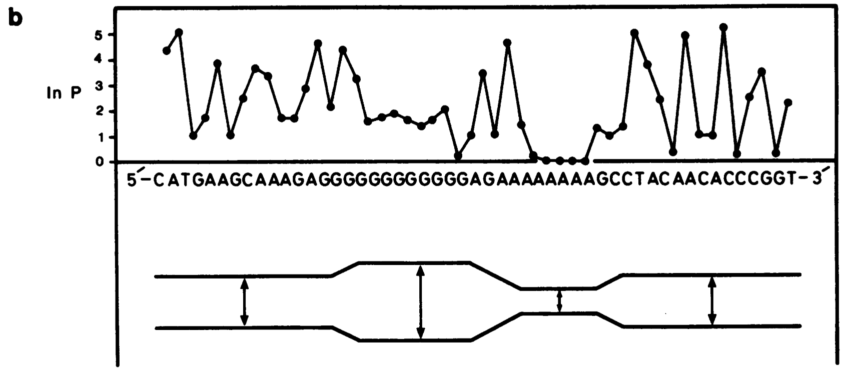


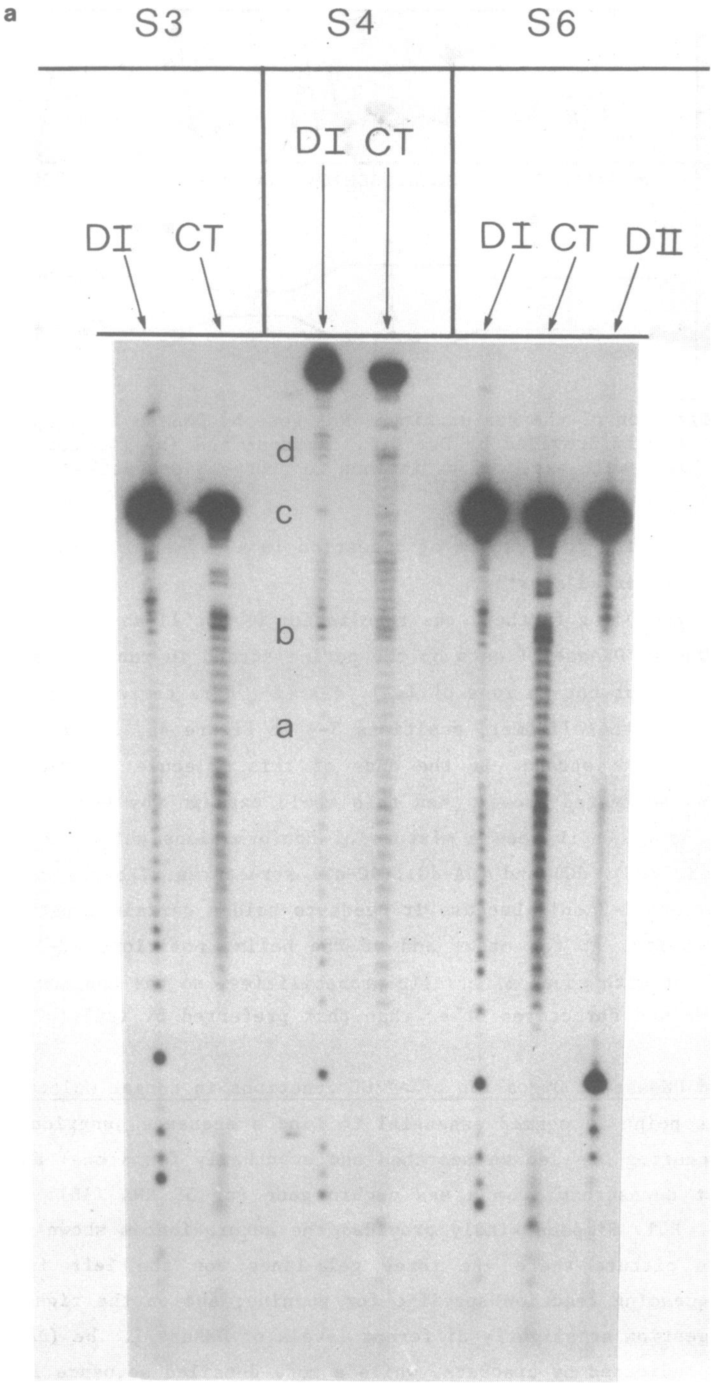
Figure 5. Digestion of the sea urchin 5S RNA gene by DNAase I: (a) the profile of digestion kindly provided by Dr. R.T. Simpson; and (b) the corresponding probability plot, with expected variations in groove width marked below.

desirable to examine the results of digestion in a longer molecule, and these data will be presented shortly.

Before proceeding further, the results for DNAase II should be mentioned. In long polymers, DNAase II cuts on the purine strand in runs of (dG).(dC) or (dA-dG).(dC-dT) but not in runs of (dA).(dT) (3). Here certain bonds near the (dA).(dT) end of the oligomer, positions 3-4 in Figure 4b, are cut well but only on the purine strand. If the ends of this molecule are thermally disordered as suggested above, then this would explain why the enzyme cuts these bonds so well: it sees a mixture of conformations which includes the characteristic (dG).(dC) and (dA-dG).(dC-dT) structures. The enzyme fails to cut at positions 1-2 only because it needs to hold a certain length of chain for full activity. At the other end of the helix, positions 12-18, both strands are cut with equal if erratic probabilities, so the conformation there must also include structures other than that preferred by (dG).(dC) in long DNA.

DNAase I and DNAase II Digestion of AT/GC Junctions in Longer Molecules

At this point it seemed essential to find a sequence junction in some naturally-occuring DNA; so we searched and eventually found one: a sequence $G_{12}AGA_8$ just downstream from a sea urchin gene for 5S RNA (16). At our request, Dr. R.T. Simpson kindly provided the autoradiogram shown in Figure 5a. In this picture there are three gel lanes: on the left is a chemical-sequencing reaction specific for guanine; and on the right are two lanes of digestion at slightly different levels of DNAase I. The (dA) and (dG) regions are indicated by brackets, while a more detailed sequence is shown in


a
S3
S4
S6
DI
CT
DI
CT
DI
CT
DII
d
c
b
a

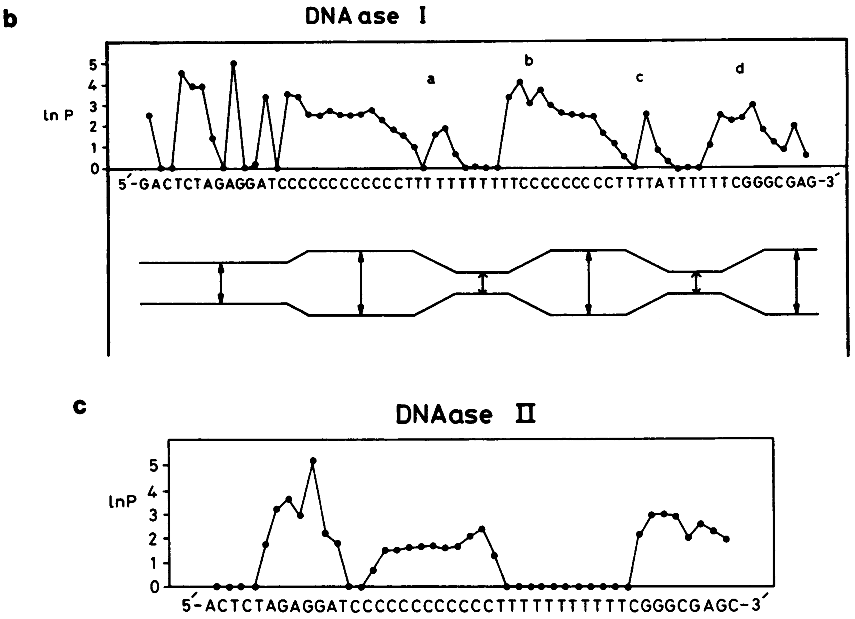


Figure 6. Digestion of the cloned 20-mer in several molecules: (a) profiles of digestion for clones S3, S4, and S6 by DNAase I ("DI"), DNAase II ("DII"), along with various hydrazine-piperidine sequencing tracks ("C plus T"); (b) the probability plot for DNAase I cleavage of clone S4, with junctions (a, b, c, d) marked above data points, and expected variations in groove width drawn below; (c) probability plot for DNAase II cleavage of clone S3. It is important to note that all probability values listed here are on a natural logarithmic scale.

Figure 5b along with a densitometer analysis of gel intensities.

As expected, both (dA).(dT) and (dG).(dC) segments in this 260-base-pair molecule strongly resist DNAase I cleavage; but several bonds which lie at the junction between these segments are cut rather well. In the lower half of Figure 5b, we have drawn a set of lines to explain our present interpretation of these data. We think that the (dG).(dC) segment is resistant to DNAase I on account of its wide minor groove, while the (dA).(dT) segment is resistant because of its narrow minor groove. All of the mixed-sequence DNA on either side, and the junction DNA in the middle, is sensitive to DNAase I digestion because the groove is intermediate in size. The many step-to-step variations in cutting rate are attributed to local variation in phosphate accessibility (2,3).

The simplicity of this interpretation encouraged us to search for other sequence junctions in long DNA. Finding no other naturally occurring examples, we decided to clone the synthetic 20-mer studied above into plasmid pUC13.

Working from the blunt-ended Sma I site, we isolated three recombinant molecules which will be designated "S3", "S4" and "S6". Upon isolation and sequencing, S3 and S6 were found to be identical, each containing a single copy of the insert $C_8T_{11}C$ in the same orientation. Both are 65 base pairs long with the insert in the center. Clone S4 was found to contain a dimer insert $C_8T_{11}C$ - $C_8T_4AT_6C$, mutated from T_{11} to T_4AT_6 in the second copy. It is 85 base pairs long with the insert at center.

Patterns of DNAase I and DNAase II digestion for these three clones are shown in Figure 6a; corresponding densitometer analyses are shown in Figures 6b and 6c. In the sequence of S4 there are actually four sequence junctions, marked as "a", "b", "c" and "d", so it is on this molecule that we will focus our attention. (Digestion profiles of S3 and S6 are closely equivalent to that of S4 in regions of sequence homology.) Junctions "b" and "d", in the center of Figure 6a and the top half of 6b, are cut as expected by DNAase I: strongly at the intersection of (dT) and (dC) segments and less well on either side. Junctions "a" and "c" are also cut preferentially near the junction, though less strongly and somewhat to one side.

The expected theoretical interpretation is pictured below the plot in Figure 6b. Runs of (dC), having a wide minor groove, should be relatively insensitive to DNAase I cleavage as observed; runs of (dT), having a narrow minor groove, should also be cut poorly as observed. At the junction between each (dC) and (dT) segment there should be a short zone of DNAase I sensitivity where the groove becomes intermediate in size. This interpretation is clearly supported by examples "b" and "d", but less so by examples "a" and "c". It seems that for sequences (dT) and (dC) joined by a TC step, DNAase I cuts rather well, thereby supporting the idea that variation in groove width is spread over several bonds. By contrast, for sequences (dC) and (dT) joined by a CT step, DNAase I cuts rather poorly at the junction; nevertheless a region of enhanced cleavage does lie within the (dT) sequence 3-4 bonds distant. One possible explanation for this result is that the helical structure has become rather discontinuous or "kinked" at the CT junction, thereby reducing the rate of DNAase I cleavage. Full explanation of the effect will require more structural data from x-ray crystallography concerning the conformation of such sequences.

Digestion of clone S6 by the enzyme DNAase II appears in the rightmost lane of Figure 6a, and in the plot of Figure 6c. As expected from previous studies of DNAase II specificity, the enzyme cuts preferentially within both runs of purine, AGAGGA and GGGCGAG, less well with the long run of C and not at all within the long run of T. Slight enhancement of cleavage is detectable at the (dC).(dT) junction.

Comparison of these plots (6b and 6c) with those for the short oligomer above (Figures 3b and 4b) confirms that patterns of cleavage within the short oligomer are strongly influenced by end effects. The cutting profiles for DNAase I are quite dissimilar, and those for DNAase II show similarity only at positions 1-15.

Restriction Enzymes

If flanking sequence can influence the rate of DNAase I cleavage over many base-pair steps as just seen above, then can it influence the rate of restriction enzyme cleavage in a similar fashion? Armstrong and Bauer (17), and similarly Alves et al. (18), have studied the influence of nearby sequence on rates of cleavage by Eco RI (GAATTC), Hinf I (GANTC) and Pst I (CTGCAG), and found that the activities of all three enzymes could be inhibited by long flanking runs of (dG).(dC). We suggested that this result might usefully be interpreted in terms of an altered helix conformation at runs of (dG).(dC), which carries over into the protein binding site so as to disturb its structure (3). A selective inhibition of cleavage by long flanking runs of (dA).(dT) has also been observed and will be reported elsewhere (12). Here let us present some additional data on the subject.

We chose four enzymes: Fnu DII (CGCG), Hae III (GGCC), Hha I (GCGC) and Msp I (CCGG), and measured their relative probabilities of cleavage over 4000 nucleotides of plasmid DNA (14). Sites were mapped on two separate fragments, so that products of digestion from each 2000-base-pair molecule could be resolved in a 4.5% polyacrylamide gel. Typical results are shown in Figure 7 and numerical data are listed in Tables 1-4. Each site has been assigned a letter (a, b, c, ..., y, z, aa, bb) according to its position in the plasmid.

Some of these enzymes are not as well-behaved as DNAase I and DNAase II: their product distributions look different when the radioactive label lies at one end of the molecule or the other. One can look at the gel patterns in Figure 7 and immediately see evidence for this behavior. In the Msp I tracks on the right, for example, the lowest-molecular-weight band "n" is by far the most intense even at slight levels of digestion. This could be because site "n" is cut especially well by Msp I, but if one looks again at the same pattern of digestion when the radioactive label lies at the other end of the molecule (not shown), then band "f" appears most intense and band "n" looks like all the rest. In the Fnu DII tracks on the left, and the Hae III tracks at center, the lowest-molecular-weight bands "a" and "a" are not especially

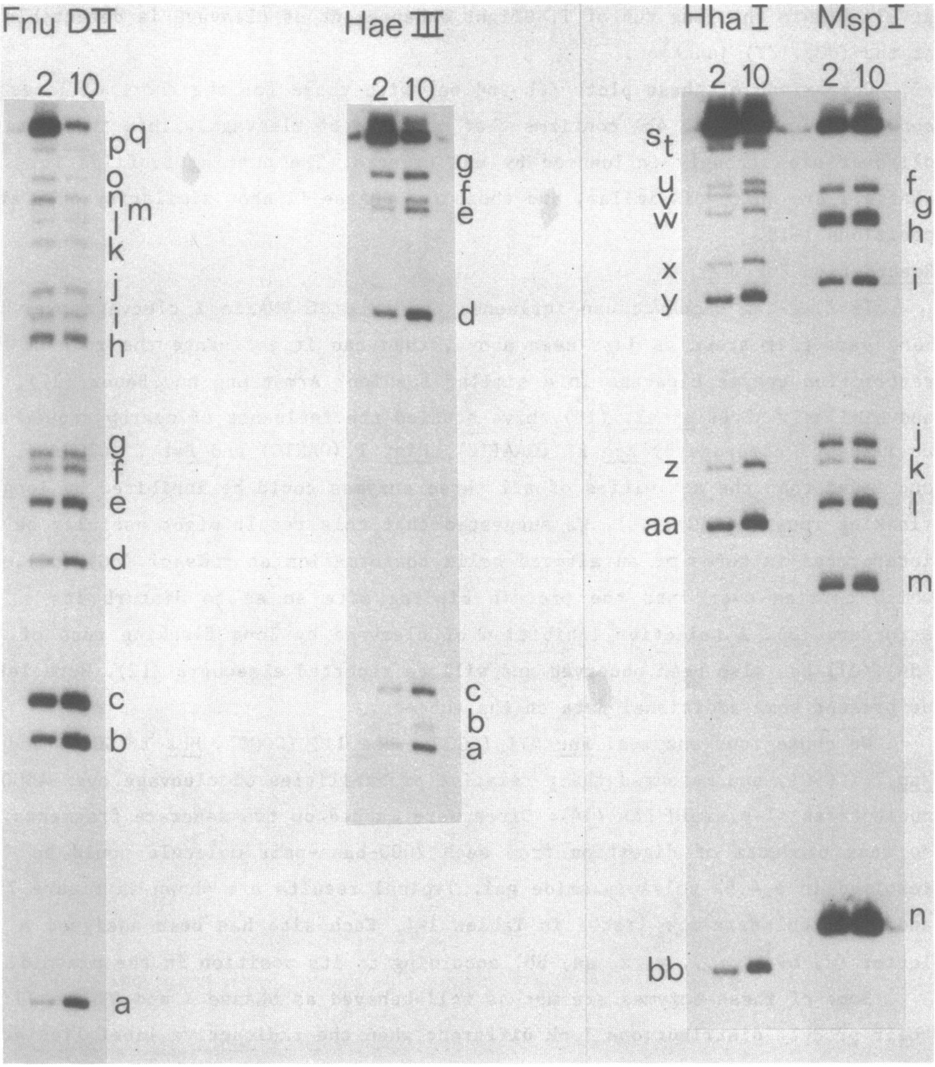


Figure 7. Profiles of digestion by restriction enzymes Fnu DII, Hae III, Hha I and Msp I on plasmid DNA, with times 2, 10 in minutes. Sites (a, b, c, ...) correspond to the probability values listed in Tables 1-4.

intense (at least in 2-minute lanes) and the pattern looks much the same when the label lies at the other end.

Thus, some care is necessary in interpreting data from enzymes such as Msp I: once bound to the DNA it slides along and seldom comes off, thereby producing a subpopulation of very short fragments. Fnu DII and Hae III seem to

Table 1. Probabilities of Cleavage: Fnu DII

Site	Numbering	Probability	5'-flanking		3'-flanking
c	871	169	ATTTT	[CGCG]	AATCC
b	847	89	GTTTG		CAGTC
e	1029	79	CCCTG		ATTGA
h	1296	77	GAAGT		GTCGG
t	3449	70	GATAC		AGACC
s	3119	67	GATTA		CAGAA
r	2538	52	AAGGC		TTGCT
n	1795	49	CTATG		ATGAT
g	1084	41	AACTT		TGATG
f	1066	38	CACCA		ATGAC
j	1425	38	ATTTC		CTCGG
o	1914	31	ATCGT		CTGAT
d	973	31	CTGGC		TGAAT
i	1353	30	CAAAT		CTTAA
a	729	21	CTGCA		CACTT
u	3942	21	AATAC		CCACA
k	1592	19	CGGTG		TTTCT
1	1681	12	CAAAA		TGCGT
m	1731	6	GCCAG		CTGGA

Not resolved: p(2092), q(2094)

find their cutting sites in a more random fashion, while Hha I appears to be an intermediate case. Most of the probabilities listed in Tables 1-4 were measured twice in separate experiments with the radioactive label at either end; bands such as "n" for Msp I were assigned their apparent value when most distant from a labelled end.

As shown in Table 1, Fnu DII exhibits a 28-fold variation in probability of cleavage. Some of the most-favored sites are of a type Y[CGCG]R, for example (c, h, t); whereas many of the least-favored sites are of a type R[CGCG]Y, for example (m, 1, k). A notable exception to this rule is site "b", which is cut quite well with a sequence R[CGCG]Y, but possibly this is because site "b" lies adjacent to most-favored site "c".

There are not enough positions of Hae III cleavage in the DNA sequence examined here (Table 2) to understand the role of flanking sequence in enzyme specificity; the observed variation in probability is roughly 20-fold, and the least-favored site "b" has a run of purine bases on either side.

The enzyme Hha I exhibits only a 7-fold variation in probability of cleavage (Table 3). The five most-favored sites (aa, y, m, 1, p) are all of a type Y[GCGC]R, while the least-favored site "v" is of a type R[GCGC]Y.

Table 2. Probabilities of Cleavage: Hae III

Site	Numbering	Probability	5'-flanking		3'-flanking
g	3774	315	AAGTT	[GGCC]	GCAGT
d	2969	260	GTGGT		TAACT
c	2535	115	AAAAA		GCGTT
e	3427	98	CATCT		CCAGT
a	2506	85	CAAAA		AGCAA
f	3507	78	GGAAG		GAGCG
b	2517	15	CAAAA		AGGAA

The behavior of *Msp* I is so processive that it is difficult to know how much of the measured 7-fold variation in probability is meaningful. The least-favored site, "k" in Table 4, has several purine bases on either side.

In summary, the influence of flanking sequence on the activities of these restriction enzymes is complex and varied. For *Fnu* DII and *Hha* I, a good correlation can be found in terms of the bases closest to the primary site of recognition; for *Hae* III no interpretation is offered; and for *Msp* I the identity of nearby sequences is of little importance.

Table 3. Probabilities of Cleavage: Hha I

Site	Numbering	Probability	5'-flanking		3'-flanking
aa	3606	100	AGTTT	[GCGC]	AACGT
y	3120	80	ATTAC		AGAAA
m	1651	60	CTGTT		ATGAA
l	1591	55	CCGGT		GTTTC
p	1794	50	TCTAT		GATGA
r	2093	50	GCCTC		GTTTC
bb	3943	50	ATACC		CACAT
h	1135	45	TCGAT		CCATT
q	1915	40	TCGTC		TGATC
f	964	35	TTCAG		CTGGC
g	1028	35	GCCCT		GATTG
k	1426	30	TTTCC		TCGGC
u	2670	30	CTCGT		TCTCC
z	3513	30	GCCGA		AGAAG
c	753	25	CTGCT		TCCGC
x	3011	22	TATCT		TCTGC
t	2400	20	TCGCT		TCGGT
w	2837	17	CCGCT		CTTAT
v	2737	15	GCGTG		TTTCT

Not resolved: a(721), b(730), d(846), e(848), i(1343), j(1354), n(1730), o(1732), s(2367), cc(4275)

Table 4. Probabilities of Cleavage: Msp I

Site	Numbering	Probability	5'-flanking		3'-flanking
h	2872	215	CCAAC	[CCGG]	TAAGA
b	1265	210	GGGTG		GTTAA
c	1586	170	CGAAA		TGCGC
g	2846	170	CTTAT		TAACT
m	3677	165	CAGCT		TTCCC
l	3567	120	TGTTG		GAAGC
i	3062	115	TTGAT		CAAAC
n	3919	109	CTTGC		CGTCA
j	3466	100	GCTCA		CTCCA
f	2699	90	GCTTA		ATACC
k	3500	30	GCCAG		AAGGG

Not resolved: a(878), d(1886), e(1924)

DISCUSSION

It is now clear that different sequences in a DNA molecule can be distinguished in either of two ways: by contacts with the base pairs, or by contacts with the sugar-phosphate backbones. The hydrogen-bonding positions on the exposed edges of base pairs vary in a discrete and discontinuous fashion according to the sequence of nucleotides, whereas the conformation and spacing of sugar-phosphate chains vary in a smooth and continuous fashion, according to base-pair overlap geometries in the core of the helix (19). These two kinds of recognition have been given various names: "sequence" versus "structure" (2), or "extrinsic" versus "intrinsic" (20). Here we suggest the terms "digital" and "analogue" in order to emphasize that one process is discrete and the other continuous.

In Figure 8 is shown a very simple approach to understanding this problem. The various forms of recognition have been divided into two overriding categories "analogue" and "digital", and then further subdivided within the analogue category into two more sections which one might call "exposure" and "conformation". The protein or antibiotic must first see the structure to which it desires to bind, and this requires that the sugar-phosphate chains be exposed and not covered by another DNA helix or, most notably, by a set of histone proteins. If the proper structure is exposed, then the protein must check to see if the conformation is correct; e.g. check the groove width for a small protein, or the global curvature for a large one (21). Finally, if both the exposure and conformation are correct, then the protein may proceed to consider the identity of base pairs. The hierarchy of steps in this scheme corresponds to the chemical activity of

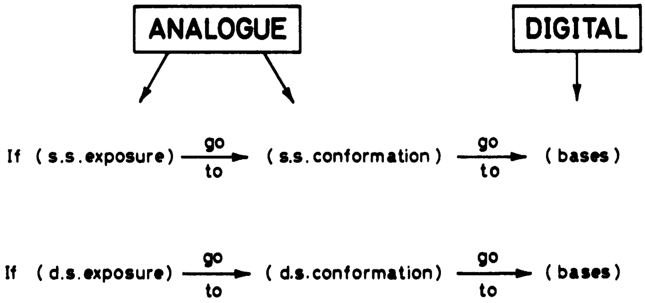

ANALOGUE
DIGITAL
If (s.s.exposure) go to (s.s.conformation) go to (bases)
If (d.s.exposure) go to (d.s.conformation) go to (bases)

Figure 8. A general model for protein-DNA specificity: "analogue" recognition of sugar-phosphate backbones, combined with "digital" recognition of base pairs or individual bases. Dr. C.R. Calladine has suggested that the process of protein-DNA recognition may be likened to that of a blind man in a library: first the blind man runs his hands over many volumes to check their size and shape; then for each volume which seems suitable, he turns to the title page and reads the characters there in Braille script.

reacting groups: phosphates first and bases second.

When considering the influence of flanking sequence on protein-DNA affinity, clearly one should think of this as an analogue phenomenon. The free energy of interaction will vary continuously according to the relative disposition of sugar-phosphate chains at the primary site of recognition, and this in turn will be influenced by sequences on either side.

DNA Structure in Crystal and Solution

Since analogue recognition is a rather commonly-used way of distinguishing base sequences in DNA, it is important to know what kind of backbone conformations are preferred by these different sequences. The only reliable source of information in this regard is single-crystal x-ray analysis. Whenever x-ray crystallography enters a new field, solution chemists are often able to point out differences between crystal and solution. In the case of DNA, however, all reported x-ray structures (5,8,9,22) seem to be in good agreement with nuclease digestion studies (2,3,10,11), measurements of helical repeat (3,5,23), Raman spectroscopy (24), solution x-ray scattering (25,26) and many NMR results (27).

The only apparent disagreement concerns the conformation of sequences such as r(GCG)TATACGC, which adopt a classical RNA-DNA hybrid conformation in the crystal but seem to adopt a somewhat different conformation in solution (28). One possible explanation is that the molecule in solution is "wiggling around" more than in the crystal, and thus what is observed by NMR is a time-averaged mixture of conformations that differs somewhat from the single conformation in a crystalline lattice. Indeed, we have found by nuclease digestion measurements that both ends of a 20-base-pair double helix adopt an altered double-helical conformation relative to the inner part of the same molecule, and relative to a double helix of the same sequence imbedded in long DNA. In a similar fashion, Alves _et al._ (18) have found that relative rates of cleavage by the restriction enzyme _Eco_ RI increase by a factor of 10 in going from the inner to the outer parts of a 24-base-pair molecule.

A second point does not relate directly to any sequences which have yet been examined by crystallography, but should serve as a warning for the future. Although sequences such as (dA).(dT) and (dG).(dC) normally adopt conformations that differ significantly from that of mixed-sequence DNA, the present study shows that these conformations can be modified by the influence of flanking sequence, particularly in the vicinity of a structural junction. Such effects might be expected to appear in the x-ray structures of somewhat longer DNA molecules (16-40 base pairs) than those which have been studied so far (8-12 base pairs).

It appears then, that x-ray analysis of DNA fragments can indeed provide useful information concerning sequence-dependent conformation in solution, as long as certain caveats concerning thermal motion and sequence environment are borne in mind.

DNAase I and DNAase II

Of the proteins studied here, DNAase I and DNAase II may be considered as purely "analogue" reagents, since they seem to recognize only the continuous variation in sugar-phosphate backbones; whereas the restriction enzymes _Fnu_ DII, _Hae_ III, _Hha_ I and _Msp_ I might better be considered as "analogue-digital" hybrids, making use of both forms of recognition to obtain their very-high sequence selectivities.

The enzyme DNAase I cuts poorly at sequences such as (dA).(dT) or (dG).(dC) as compared to mixed-sequence DNA, but it will cut preferentially between these sequences in many instances; it will cut at (dA).(dT) if the groove is opened nearby, and it will cut at (dG).(dC) if the groove is closed nearby. We also know that DNAase I binds to two strands rather than one (6), and this can be explained in terms of the protein folding (7). The only model logically consistent with all of these data is that shown in Figure 1a above: runs of (dA).(dT) are cut poorly on account of their narrow minor groove; runs of (dG).(dC) are cut poorly because of their wide minor groove; while the junction between AT-rich and GC-rich segments is cut well if the groove width varies symmetrically and continuously, but poorly if the structure changes abruptly.

DNAase II cuts on the purine strand of sequences such as (dG).(dC) or (dA-dG).(dC-dT) in normal duplex DNA, and also on the purine strand of (dA).(dT) at the end of a helix or if grooves are opened nearby (10). Sometimes, however, it cuts both strands equally: for example positions 12-17 in Figure 4b above, or when the helix is unwound by organic solvent (3). We also know that DNAase II requires just one strand for its cleavage activity (6), and from its behavior near helix ends we may deduce that it holds about five phosphates. The simplest model to explain these data is that shown in Figure 1b, where DNAase II recognizes some aspect of single-strand conformation. In principle, the protein could prefer either a compressed single-strand or an extended single-strand but, for reasons discussed previously (3), the compressed-strand model is currently more suitable.

Restriction Enzymes

Proteins such as Fnu DII, Hae III, Hha I and Msp I are reagents of high digital specificity: they cut only at double-helical sequences CGCG, GGCC, GCGC and CCGG, respectively. Yet relative rates of cleavage by Fnu DII, for example, are not constant for all CGCG sequences but vary 30-fold according to the identity of bases on either side. Fnu DII cuts especially well at sequences YCGCGR and especially poorly at sequences RCGCGY, although this is clearly not the sole factor involved; other long-range influences can also be important (12,17). Hha I exhibits a similar preference for YGCGCR over RGCGCY; while the preferences of Hae III and Msp I are not yet clear.

All of these results can be rationalized in terms of DNA helix structure at the recognition site, and how it is influenced by the bases on either side. In a local sense, switching purines for pyrimidines on either side of the restriction site will alter base-pair roll angles, and therefore helix groove widths, by a mechanism discussed elsewhere (20). In a more global sense, long flanking runs of (dG).(dC) or (dA).(dT) can impose their own kinds of structure from a somewhat longer distance, and these structures need not be equivalent to that preferred by the enzyme (3).

In summary, the influence of flanking sequence on nuclease digestion specificities seems to be complex, and not yet interpretable in terms of any single model. The best that can be done at present is to divide the problem of protein-DNA recognition into two parts, analogue and digital, and continue to collect data on both components.

ACKNOWLEDGEMENTS

Thanks are due to R.T. Simpson, K. Fox, P. Caspers and P. Geiduschek for use of their gel photographs, and especially to M. McCall for an explanation of the G_4C_4 structure. This research was supported by PHS Grant Number CA06971-03 of the National Cancer Institute, DHHS.

REFERENCES

1. Klug, A., Jack, A., Viswamitra, M.A., Kennard, O., Shakked, Z. and Steitz, T.A. (1979) J. Mol. Biol. 131, 669-680.
2. Lomonossoff, G.P., Butler, P.J.G. and Klug, A. (1981) J. Mol. Biol. 149, 745-760.
3. Drew, H.R. and Travers, A.A. (1984) Cell 37, 491-502.
4. Fratini, A.V., Kopka, M.L., Drew, H.R. and Dickerson, R.E. (1982) J. Biol. Chem. 257, 14686-14707.
5. McCall, M., Brown, T. and Kennard, O. (1985) J. Mol. Biol. 183, in press.
6. Drew, H.R. (1984) J. Mol. Biol. 176, 535-557.
7. Suck, D., Oefner, C. and Kabsch, W. (1984) EMBO J. 3, 2423-2430.
8. Wang, A.H.-J., Ughetto, G., Quigley, G.J., Hakoshima, T., van der Marel, G.A., van Boom, J.H. and Rich, A. (1984) Science 225, 1115-1121.
9. Kopka, M.L., Yoon, C., Goodsell, D., Pjura, P. and Dickerson, R.E. (1985) Proc. Nat. Acad. Sci. USA 82, 1376-1380.
10. Low, L., Drew, H. and Waring, M. (1984) Nucleic Acids Res. 12, 4865-4879.
11. Fox, K. and Waring M. (1984) Nucleic Acids Res. 12, 9271-9285.
12. K. Fox, personal communication.
13. Drew, H.R., Weeks, J. and Travers, A.A. (1985) EMBO J. 4, 1025-1032.
14. Lamond, A.I. and Travers, A.A. (1983) Nature 305, 248-250.
15. Lutter, L.C. (1978) J. Mol. Biol. 124, 391-420.
16. Simpson, R.T. and Stafford, D.W. (1983) Proc. Nat. Acad. Sci. USA 80, 51-55.
17. Armstrong, K.A. and Bauer, W.R. (1983) Nucleic Acids Res. 11, 4109-4126.
18. Alves, J., Pingoud, A., Haupt, W., Langowski, J., Peters, F., Maass, G. and Wolff, C. (1984) Eur. J. Biochem. 140, 83-92.
19. Calladine, C.R. and Drew, H.R. (1984) J. Mol. Biol. 178, 773-782.
20. Dickerson, R.E. (1983) Scientific American (Dec.) 249, 94-111.
21. Widom, J. (1984) Bioessays 2, 11-14.
22. Dickerson, R.E., Drew, H.R., Conner, B.N., Kopka, M.L. and Pjura, P. (1982) Cold Spring Harbor Symp. Quant. Biol. 47, 13-24.
23. Rhodes, D.R. (1983) in Topics in Nucleic Acid Structure, Part 2, Neidle, S. Ed., pp. 287-304, Macmillan Press, London.
24. Nishimura, Y., Torigoe, C., Katahira, M. and Tsuboi, M. (1985) Nucleic Acids Res. Symp. Series 15, 147-150.
25. Bram, S. (1971) Nature New Biol. 232, 174-176.
26. Bram, S. (1972) Nature New Biol. 233, 161-164.
27. Patel, D.J., Pardi, A. and Itakura, K. (1982) Science 216, 581-590.
28. Mellema, J.-R., Haasnoot, C., van der Marel, G., Wille, G., van Boeckel, C., van Boom, J. and Altona, C. (1983) Nucleic Acids Res. 11, 5717-5738.