Volume 13 Number 5 1985 Nucleic Acids Research

Concordance of experimentally mapped or predicted Z-DNA sites with positions of selected alternating purine-pyrimidine tracts

Andrzej K.Konopka, Johanes Reiter[1]*, Manfred Jung[2], David A.Zarling+ and Thomas M.Jovin

Abteilung Molekulare Biologie, Max-Planck-Institut für biophysikalische Chemie, D-3400 Göttingen, FRG, [1]Institut für theoretische Chemie und Strahlchemie der Universität, A-1090 Wien, Austria, and [2]Abteilung Biochemische Kinetik, Max-Planck-Institut für biophysikalische Chemie, D-3400 Göttingen, FRG

Received 7 December 1984; Revised and Accepted 4 February 1985

ABSTRACT

The recent electronmicroscopic and biochemical mapping of Z-DNA sites in ϕX174, SV40, pBR322 and PM2 DNAs has been used to determine two sets of criteria for identification of potential Z-DNA sequences in natural DNA genomes. The prediction of potential Z-DNA tracts and corresponding statistical analysis of their occurrence have been made on a sample of 14 DNA genomes.

Alternating purine and pyrimidine tracts longer than 5 base pairs in length and their clusters (quasi alternating fragments) in the 14 genomes studied are under-represented compared to the expectation from corresponding random sequences. The fragments $[d(G\cdot C)]_n$ and $[d(C\cdot G)]_n$ $(n\geq 3)$ in general do not occur in circular DNA genomes and are under-represented in the linear DNAs of phages λ and T7, whereas in linear genomes of adenoviruses they are strongly over-represented. With minor exceptions, potential Z-DNA sites are also under-represented compared to random sequences.

In the 14 genomes studied, predicted Z-DNA tracts occur in non-coding as well as in protein coding regions. The predicted Z-DNA sites in ϕX174, SV40, pBR322 and PM2 correspond well with those mapped experimentally. A complete listing together with a compact graphical representation of alternating purine-pyrimidine fragments and their Z-forming potential are presented.

INTRODUCTION

The alternation of purines and pyrimidines in DNA sequences constitutes one of the most important factors potentiating the transition from the right-handed B to the left-handed Z helical conformation *in vitro* (1-4).

Topological stress in the form of negative supercoiling promotes the B to Z transition of protein-free covalently closed circular DNA (ccc DNA) (5-12). Studies of anti-Z-DNA-IgG binding to chromosomal DNA (11-16) have established the existence of potential Z-DNA tracts *in vivo*. It is probable that the combined effects of nucleotide sequence, topological stress, and interactions with ions, proteins and polyamines (5, 11, 12, 17) determine the physiological distribution and functions of left-handed DNA (for a review see 4).

The biological significance of Z-DNA is unknown. It has been suggested that some potential Z-DNA loci in the SV40 genome can play a role in the

control of transcription or in genetic recombination (5, 11, 12 18). Other studies with cytological material have emphasized potential structural roles for left-handed DNA in chromosomal organization (4, 11, 12, 15, 16).

Different alternating purine-pyrimidine sequences in linear synthetic polymers exhibit a hierarchy in the potential for undergoing the B-Z transition (4, 20). The minimum length of a linear alternating oligonucleotide required for the establishment of the Z form in solution has been evaluated as 6 base pairs (20). Although the precise sequence-dependence of the transition equilibrium remains to be experimentally established, it is clear that the G·C basepair is much more effective than the A·T basepair in stabilizing the Z conformation (4, 20). Thus, we can identify two primary factors which determine the Z-forming potential of a natural sequence: length and base composition. In addition, studies of anti-Z-DNA-Ig binding sites in pBR322 (8, 19) indicate that DNA fragments including bases out of alternation may also assume the left-handed conformation at high superhelix density. It follows that a favorable (clustered) distribution of potential Z-forming tracts may lead to a cooperative and collective behaviour.

The mapping of anti-Z-DNA immunoglobulin binding sites by immunoelectron microscopy provides several examples of naturally ocurring Z-DNA tracts in ϕX174 (23), PM2 (24-26) and SV40 (18, 27) DNAs. Corresponding data also exist for the cloning vector pBR322 (8). On the basis of the available data, the minimal length of an alternating purine-pyrimidine fragment required for stabilization of the Z conformation in natural sequences is on the order of 8 base-pairs, although shorter tracts composed exclusively of G and C may also be effective (23). The results of these experimental studies have been used to define empirical criteria for identifying potential Z-DNA tracts on the basis of nucleotide sequence (in the next section two working definitions of sequences with the potential for adopting the left-handed conformation are presented). The algorithms based on these definitions have been applied to several viral and episomal DNA genomes. The results of the search together with a corresponding statistical analysis are presented and discussed in this paper.

BASIC PRINCIPLES

Terminology. DNA sequences can be analysed for simple dinucleotide repeating units. In particular, we will consider alternating repetitions of a purine (R) and a pyrimidine (Y). A sequence of purines and pyrimidines in alternation is referred to as an *alternating fragment* (AF). Formally we can

consider two kinds of AFs: those which are alternations of only two bases and those which consist of more than two bases. An AF of the first kind will be referred to as a *uniform alternating fragment* (uAF). Examples of uAFs are the sequences: GTGTGTGTGTG, ACACACAC, CGCGCGCGCG, ATATATATAT. The other kind of AF will be referred to as a *mixed alternating fragment* (mAF). Examples of mAFs are the sequences: GCGTACGT, GCACATGTA, ACACGTACATG, ACGTACGTACGT.

In a long DNA tract, AFs can be separated or clustered. The obvious criterion for establishing whether a block of AFs constitute a cluster is based on the distances between the AFs. From the viewpoint of Z-forming potential, we regard one base-pair as a reasonable maximum distance between AFs in a cluster. A cluster of AFs will be referred as a *clustered alternating fragment* (cAF). An example is the sequence: ATACGT TGTGTGT T CGATCGTG. (Here and elsewhere a space will be used to denote the separation between AFs constituting a cAF.) This cluster consists of three AFs. The first two are contiguous (distance = 0). The second and third AFs are separated by one base (distance = 1). Thus, we consider a cAF as equivalent to an AF with a few bases out of perfect alternation. The length of a cAF will be taken as the difference between the positions of the first base of the first AF and of the last base of the last AF in the cluster. Under the assumption that the Z-forming potential can be a property of AFs as well as cAFs, we define a *quasi-alternating fragment* (qAF) as a DNA sequence which is either an AF or a cAF.

Potential Z-DNA. Taking into account the facts briefly described in the Introduction, we should not expect that every qAF has the potential for adopting the Z conformation. In light of available experimental data, the two following definitions of a potential Z-DNA fragment seem to be appropriate. The first definition evaluates the criteria of length and composition for a qAF as a whole, whereas the second considers the length of the longest subfragment of a given qAF.

DEFINITION 1: A potential Z-DNA fragment is a qAF fulfilling the following conditions:

i. The total length (base-pair units) is > a.

ii. The fraction of the sequence consisting of A and T in alternating repetition is ≤ b.

iii. If the fragment is a cAF, the constituent AFs have a length > c.

DEFINITION 2: A potential Z-DNA fragment is a qAF fulfilling conditions (i) and (iii) from the previous definition and containing a subfragment which fulfills conditions (i) and (ii).

For the reasons given previously, we have applied the search algorithm with

the parameters a = 7, b = 0.3, and c = 4. These values lead to the identification of the binding sites for anti-Z DNA immunoglobulins reported to date (with minor exceptions; see Discussion).

Let us consider as examples the following qAFs:

Sequence 1: ATATCG TGTGTG GCATATATAT

Sequence 2: GTATATAT TATATAT T CACAC

Sequence 3: CGCGCG T CATGTG ACACACAT

Sequence 4: CACGTATGTGTGTATATGTGCA

Sequence 5: GTGTA

Sequence 1 is a potential Z-DNA fragment according to definition 2 but not definition 1, due to violation of condition (ii). Sequences 2 and 5 are not potential Z-DNA fragments according to both definitions [sequence 5 violates all conditions whereas sequence 2 violates condition (ii)]. Sequences 3 and 4 are potential Z-DNA fragments according to both definitions.

DNA sequences studied. The DNA sequences chosen from the EMBL Nucleotide Sequence Library are as follows (ssc, single-stranded circular; dsc, double-stranded circular; dsl, double-stranded linear. Lengths are in base or base-pair units): cloning vector pBR322 (dsc 4362); bacteriophages: ϕX174 (ssc 5386), M13 (ssc 6407), T7(dsl 39936), λ (dsl 48502); papovaviruses: SV40 (simian; dsc 5243), BKV (human, strain Dunlop, dsc 5153), Polyoma-A2 (strain A2; dsc 5292); adenoviruses: Adeno-7 l (type 7; dsl 6707; left 0-18.5%), Adeno-2 l (type 2; dsl 11600; left 0-32%) Adeno-2 r (type 2; dsl 10305; right 70.7-100%); mitochondria: bovine (*Bos taurus*, dsc 16338), murine (*Mus musculus*, dsc 16295), human (dsc 16569). The sequence of a purine-pyrimidine rich region of phage PM2 (dsc 1757) has been taken from reference 26. The single-stranded DNA genomes have, of course, double-stranded DNA replicative intermediates. Linear genomes can circularize during replication.

Occurrence and average length of qAFs in a long DNA sequence. Let us consider the Y/R tracts found in a fragment of the SV40 genome (Fig. 1a-c). Figure 1a shows all AFs not shorter than 4 bases. There are two such fragments in the sequence studied. Both of them are separated AFs. The AFs which are not shorter than 3 bases are shown in Figure 1b. There are four such fragments, two of which make a cluster with two bases out of perfect alternation. Figure 1c shows all possible AFs (including doublets of alternating Ys and Rs) in the DNA fragment studied. In this case 12 AFs are found. Two of them constitute a cluster with two bases out of alternation and another six are involved in a cluster of total length 14.

These examples suggest that in general, a tendency of AFs to cluster is stronger in the case of short AFs (Figure 1c) compared to long AFs (Figure 1a

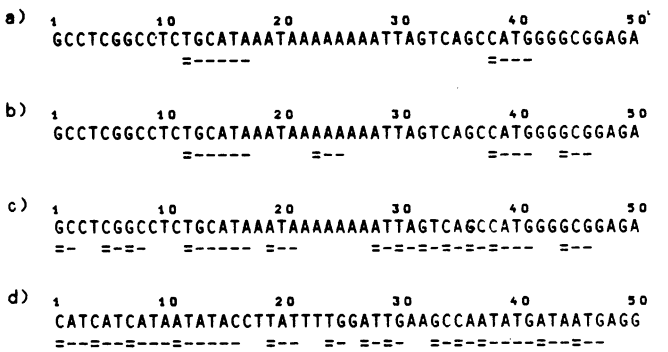

Fig. 1. Alternating Y and R repetitions in the first 50 bases of SV40 (a-c) and Adeno-2·l (d). a) AFs of length ⩾ 4; b) AFs of length ⩾ 3; c) and d) AFs of length ⩾ 2. Every AF is underlined. The = indicates the first base of an AF.

and 1b, which have a smaller number of underlined bases than Figure 1c). When we analyse all possible AFs from different natural DNAs, the tendency to cluster differs. An example is provided by a comparison of the SV40 genome fragment (Figure 1c) with a corresponding Adeno-2·l fragment (Figure 1d). There are 13 AFs in the first 50 bases of the Adeno-2·l but only one is clearly isolated. Another 12 AFs are involved in three clusters. We know from the previous example that there are 12 AFs in the first 50 bases of the SV40 genome; two of them are separated and the other 10 occur in three clusters. Thus, the tendency of short AFs to cluster is greater in the Adeno-2·l than in the SV40 fragment. The clusters in Adeno-2·l are, in general, longer than in SV40 (the lengths of the clusters are 16, 7 and 15 in Adeno-2·l whereas these lengths in SV40 are equal to 4, 10 and 14 bases).

We require a quantitative measure of the clustering tendency. Such a measure consists of the average length of qAFs in a given DNA tract. In order to define this quantity let us assume that a sequence under consideration is of length N bases, and that it contains a number k of qAFs. In addition, let $n_2, n_3, \ldots, n_m$ be the numbers of qAFs of length 2, 3,..., m bases, respectively. The average length of a qAF is defined by the following general expression:

$$\langle L \rangle = \sum_{i=2}^{m} i \cdot p_i = (1/k) \cdot \sum_{i=2}^{m} i \cdot n_i \qquad (1)$$

where n_i is the number of qAFs of length i and the probabilities p_i are equal to n_i/k.

Let us return to our example of the 50 bp regions of SV40 and Adeno-2. We have already pointed out that the tendency of AFs to cluster is greater in the adenovirus than in the SV40 fragment. We compute the <L> values for both these fragments by using (1) and distinguish the <L> values for AFs as $\langle L_{AF} \rangle$ and the values for qAFs as $\langle L_{qAF} \rangle$. For the SV40 fragment $\langle L_{AF} \rangle = 2.7$ and $\langle L_{qAF} \rangle = 6.6$, whereas the corresponding values for the adenovirus fragment are equal to 3.1 and 10.3. It appears from this calculation that although the two $\langle L_{AF} \rangle$ values are similar, the $\langle L_{qAF} \rangle$ values are considerably different. This result suggests that the quantity $\langle L_{qAF} \rangle - \langle L_{AF} \rangle$ is a good measure of the tendency of AFs to cluster.

Random DNA sequence. Random DNA sequences have been generated and representative examples chosen for Y/R searches. In order to verify our definitions of potential Z-DNA we have generated three categories of random sequences, i.e. those with an equiprobable base composition, those rich in A and T (30% each) and those rich in G and C (30% each).

The expected number of AFs longer or equal to 2k bases in a fragment of length L and a given base composition (N_A adenines, N_C cytosines, N_G guanines and N_T thymines) can be calculated in the following way: The frequencies of the bases are: $p_A = N_A/L$, $p_C = N_C/L$, $p_G = N_G/L$, $p_T = N_T/L$. Let α be the frequency of a fragment RY (R = purine, Y = Pyrimidine). If $p(R) = p_A + p_G$ and $p(Y) = p_T + p_C$, we have $\alpha = p_{RY} = p_{YR} = p(R) \cdot p(Y)$. Then the probability of an AF not shorter than 2k bases is equal to $P = [2\alpha + p^3(Y) + p^3(R)] \cdot \alpha^k/(1-\alpha)$ and the expected number A_{exp} of such fragments equals $L \cdot P$.

The expected number of uAFs of a given kind is calculated in a similar way: Let $\beta(AT) = p_A \cdot p_T$, $\beta(AC) = p_A \cdot p_C$, $\beta(GC) = p_G \cdot p_C$ and $\beta(GT) = p_G \cdot p_T$. Then the probabilities of uAFs are:

$$
\begin{aligned}
P(AT) &= [2\alpha + p^2(R) \cdot p_A + p^2(Y) \cdot p_T] \cdot \beta^k(AT) \cdot [1-\beta(AT)]^{-1} \\
P(AC) &= [2\alpha + p^2(R) \cdot p_A + p^2(Y) \cdot p_C] \cdot \beta^k(AC) \cdot [1-\beta(AC)]^{-1} \\
P(GC) &= [2\alpha + p^2(R) \cdot p_G + p^2(Y) \cdot p_C] \cdot \beta^k(GC) \cdot [1-\beta(GC)]^{-1} \\
P(GT) &= [2\alpha + p^2(R) \cdot p_G + p^2(Y) \cdot p_T] \cdot \beta^k(GT) \cdot [1-\beta(GT)]^{-1}
\end{aligned}
\qquad (2)
$$

The expected numbers of uAFs are then equal to: $A_{exp}(AT) = L \cdot P(AT)$, $A_{exp}(AC) = L \cdot P(AC)$, etc.

Occurrence of given fragments. Comparison between natural and random sequences. Let the number of fragments of a given kind (for example AFs, qAFs or potential Z-DNA sequences) found in a natural sequence be equal to A. The quantity $F = (A-A_{exp}) \cdot (A_{exp})^{-1/2}$ (analogous to a coefficient of variation) measures the degree to which the frequency of a fragment in a natural sequence differs from that calculated for the corresponding random sequence. If $F < 0$, we state that the fragment in the natural sequence is

F-fold under-represented. If F > 0 we state that a fragment is F-fold over-represented.

RESULTS

Occurrence of AFs. Table 1 shows the frequencies of occurrence of AFs longer than 5 bases. In each case we also show the corresponding values in the random sequences (second row of every case listed). It is evident that AFs

TABLE 1. Occurrence of uAFs and mAFs longer than 5 bases.[a]

GENOME	LENGTH	uAFs GC+CG	AC+CA	GT+TG	AT+TA	mAFS	All number	AFs F[b]
SV40	5243	0	3	2	0	42	47	-3.9
		0.2	0.8	0.8	2.5	77.5	81.8	
BKV	5153	0	2	0	8	40	51	-3.3
		0.2	0.7	0.7	2.8	76.0	80.5	
Polyoma	5292	0	1	2	2	54	59	-2.6
		0.6	0.9	0.8	1.2	79.1	82.7	
Human mito.	16569	0	15	1	12	143	171	-4.9
		0.7	10.8	0.3	5.0	231.2	248.1	
Murine mito.	16295	0	22	0	29	111	162	-5.7
		0.3	6.8	0.5	11.6	232.5	251.6	
Bovine mito.	16338	0	14	0	21	131	166	-5.5
		0.4	7.5	0.5	8.7	235.1	253.	
pBR322	4363	1	0	1	2	46	50	-2.2
		1.1	0.7	0.7	0.4	64.9	67.9	
Adeno-2· l	11600	36	1	9	3	113	162	-1.4
		4.4	0.9	3.2	0.7	171.2	180.3	
Adeno-2· r	10305	3	10	3	3	93	112	-3.8
		1.4	3.4	0.7	1.8	153.2	160.6	
Adeno-7· l	6707	12	1	3	3	33	42	-6.1
		1.0	0.6	1.9	1.1	99.2	103.8	
M13	6407	0	0	0	5	55	60	-3.8
		0.3	0.5	1.6	2.8	92.0	97.2	
ϕX174	5386	1	0	2	0	48	51	-3.6
		0.4	.5	1.4	1.6	79.5	83.	
T7	39936	3	16	13	3	459	494	-7.9
		6.4	6.3	8.2	8.1	673.6	702.7	
λ	48502	7	20	16	15	579	637	-4.3
		7.7	6.8	9.1	8.0	723.1	754.	

[a] The second row of every case listed shows the values expected for random sequences. The first rows show the numbers found in natural sequences.
[b] The last column lists the F-values defined in the text.

TABLE 2. Tendency of alternating fragments to cluster within DNA genomes[a]

GENOME	$\langle L_{AF} \rangle$	$\langle L_{qAF} \rangle$	$\langle L_{qAF} \rangle - \langle L_{AF} \rangle$
SV40	2.85	9.48	6.61
BKV	2.90	8.70	5.80
Polyoma-A2	2.91	9.69	6.78
Human Mito.	2.87	10.97	8.10
Murine Mito.	2.85	11.48	8.63
Bovine Mito.	2.88	11.60	8.72
pBR322	3.00	13.60	10.60
Adeno-2· l	3.01	11.70	8.69
Adeno-2· r	2.99	11.55	8.56
Adeno-7· l	2.95	10.72	7.37
M13	2.82	11.60	8.78
ϕX174	2.90	12.79	9.89
T7	2.93	12.38	9.45
λ	3.01	14.02	11.01
RANDOM 8	2.97	12.96	9.99
RANDOM 4	3.02	13.69	10.67

[a] $\langle L_{AF} \rangle$ and $\langle L_{qAF} \rangle$ are the average lengths of AFs and qAFs, respectively.

are under-represented compared to the expectation for random sequence (see the last column of Table 1 where all F values are negative). The same observation appears for mAFs which constitute about 90% of all AFs.

The above conclusion does not hold in the comparison of the occurrences of uAFs in the genomes studied compared to the corresponding random sequences. Thus, $[d(G\text{-}C)]_n$ and $[d(C\text{-}G)]_n$ fragments are strongly over-represented in adenoviruses (F = 15 for Adeno-2· l and F = 11 for Adeno-7· l), whereas in all circular DNAs they are generally absent (which is in agreement with the expectation for random sequences: |F| close to 0). However, in phages T7 and λ these uAFs are under-represented (F = -1.3 and -2.5, respectively). This suggests that in circular DNAs, long $[d(G\text{-}C)]_n$ and $[d(C\text{-}G)]_n$ tracts are avoided and that this circumstance arises at least in part from the base composition (correlation coefficient between the fractions of G and C and the F value is equal to -0.73 in these genomes).

Very different patterns of occurrence are displayed by the $[d(A\text{-}C)]_n$ and $[d(C\text{-}A)]_n$ sequences (and the complementary $[d(G\text{-}T]_n$ and $[d(T\text{-}G)]_n$). The occurrence of these fragments seems to vary from one genome to another. Even

TABLE 3. Occurrence of qAFs and potential Z-DNA sites.[a]

GENOME	qAFs		Potential Z-DNA			
			Definition 1		Definition 2	
	Number	% bases	Number	% bases	Number	% bases
SV40	13	2	10	1	8	1
	34	5	19	2	18	2
BKV	17	3	7	1	6	1
	32	6	19	3	19	3
Polyoma-A2	22	5	12	2	14	3
	33	8	19	3	18	4
Human Mito.	44	3	28	2	25	2
	102	7	61	4	61	5
Murine Mito.	51	4	21	1	23	2
	100	8	58	3	56	2
Bovine Mito.	53	4	21	2	22	2
	101	8	58	6	58	5
pBR322	22	5	17	4	16	4
	26	6	16	4	16	4
Adeno-2· l	52	5	46	4	46	4
	72	7	43	4	42	4
Adeno-2· r	39	4	28	3	29	3
	64	7	38	4	38	4
Adeno-7· l	32	6	22	4	21	4
	41	8	25	5	24	5
M13	14	2	6	1	4	0.7
	39	6	23	4	23	4
ϕX174	15	3	12	2	11	2
	33	7	20	3	20	3
T7	159	5	107	4	110	4
	246	8	151	5	151	5
λ	215	5	163	4	149	3
	299	7	184	5	184	5

[a]The first row of every case corresponds to a natural sequence whereas the second row corresponds to random sequences.

in the papovavirus genomes (SV40, BKV and Polyoma) the F values vary between 0.1 and 2.5. The same observations apply to the sequences of $[d(A-T)]_n$ and $[d(T-A)]_n$.

We can also see from Table 1 that mAFs are about ten-fold more frequent

than uAFs in both natural and random sequences. This means that although AFs are under-represented in natural DNAs, the expected proportion of uAFs to mAFs (about 1:10) is conserved in natural DNA tracts.

Tendency of AFs to cluster. Table 2 shows values for the average lengths for AFs ($<L_{AF}>$) and qAFs ($<L_{qAF}>$). The values of $<L_{AF}>$ are almost the same for all genomes studied (i.e. about 3 bases/fragment). In contrast, $<L_{qAF}>$ varies from 8.7 in BKV to 14.2 in the case of phage λ. It appears from Table 2 that the $<L_{qAF}>$ values are higher in procaryotic than in eucaryotic systems ($<L_{qAF}>$ is highest for pBR322 and phage λ, and lowest for eucaryotic papovaviruses SV40, BKV and Polyoma). It is also evident that in the case of single-stranded circular genomes (M13 and ϕX174) the tendency of AFs to cluster is stronger than in double-stranded circular eucaryotic viruses. The same observations appear from an analysis of the difference $<L_{qAF}> - <L_{AF}>$ (last column of Table 2). In random sequences, there is a very strong tendency of short AFs to cluster (last two rows of Table 2).

The conclusion which can be drawn in this section is that the tendency of AFs to cluster decreases in the order: random sequence > procaryotic DNAs > mitochondrial and linear eucaryotic DNA > circular DNAs of eucaryotic viruses.

Distribution of qAFs and potential Z-DNA within genomes. The frequencies of occurrence of qAFs and potential Z-DNA sites in the DNAs studied (including the corresponding random sequences) are shown in Table 3. The linear maps of qAFs and potential Z-DNA sites for the genomes studied and the two random sequences are shown in the Figures 2 and 3. The plots for random sequences illustrate that our criteria for potential Z-sites restrict the amount of A and T in qAFs.

The general conclusion which appears from Figures 2-4 and Table 3 is that qAFs are under-represented in every case studied. This under-representation is not so universal for potential Z-DNA tracts. Except for the cloning vector pBR322, the circular DNAs studied display strong under-representation of potential Z-DNA sites (particularly in the case of mitochondrial DNAs). The same observation holds for the linear DNAs of phages T7 and λ, whereas potential Z-DNA sites occur with almost random frequency in the linear genomes of adenoviruses (Adeno-2 and Adeno-7). Complete listings of all qAFs longer than 7 bases found in the DNAs studied are presented in Figures 5 - 8.

It is interesting that potential Z-DNA occurs both within coding and non-coding regions of the genomes studied. This finding is in agreement with existing experimental data (see Discussion).

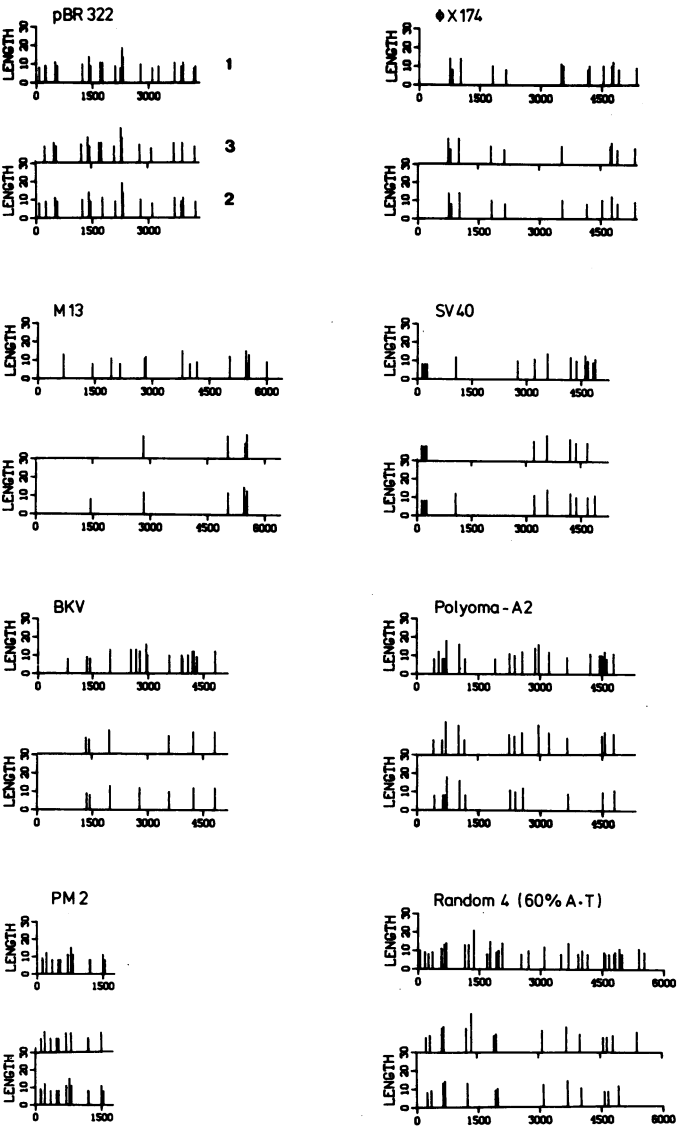

Fig. 2. Linear maps of qAFs and potential Z-DNA sequences in cloning vector pBR322, papovaviruses, dsc-phage PM2 (partial sequence), and ssc-DNA phages. The maps are rendered linearly starting at the origin defined in the Data Bank (EMBL Nucleotide Sequence Library) and extending to the right. The fragments are shown as vertical lines with lengths in bases given by the ordinate. The top horizontal line (e.g. number 1 in the case of pBR322) corresponds to qAFs, the botom line (number 2) to potential Z-DNA tracts fulfilling definition 1 and the middle line (number 3) to potential Z-DNA fulfilling definition 2. Random 4 is a random sequence rich in A and T (30% each).

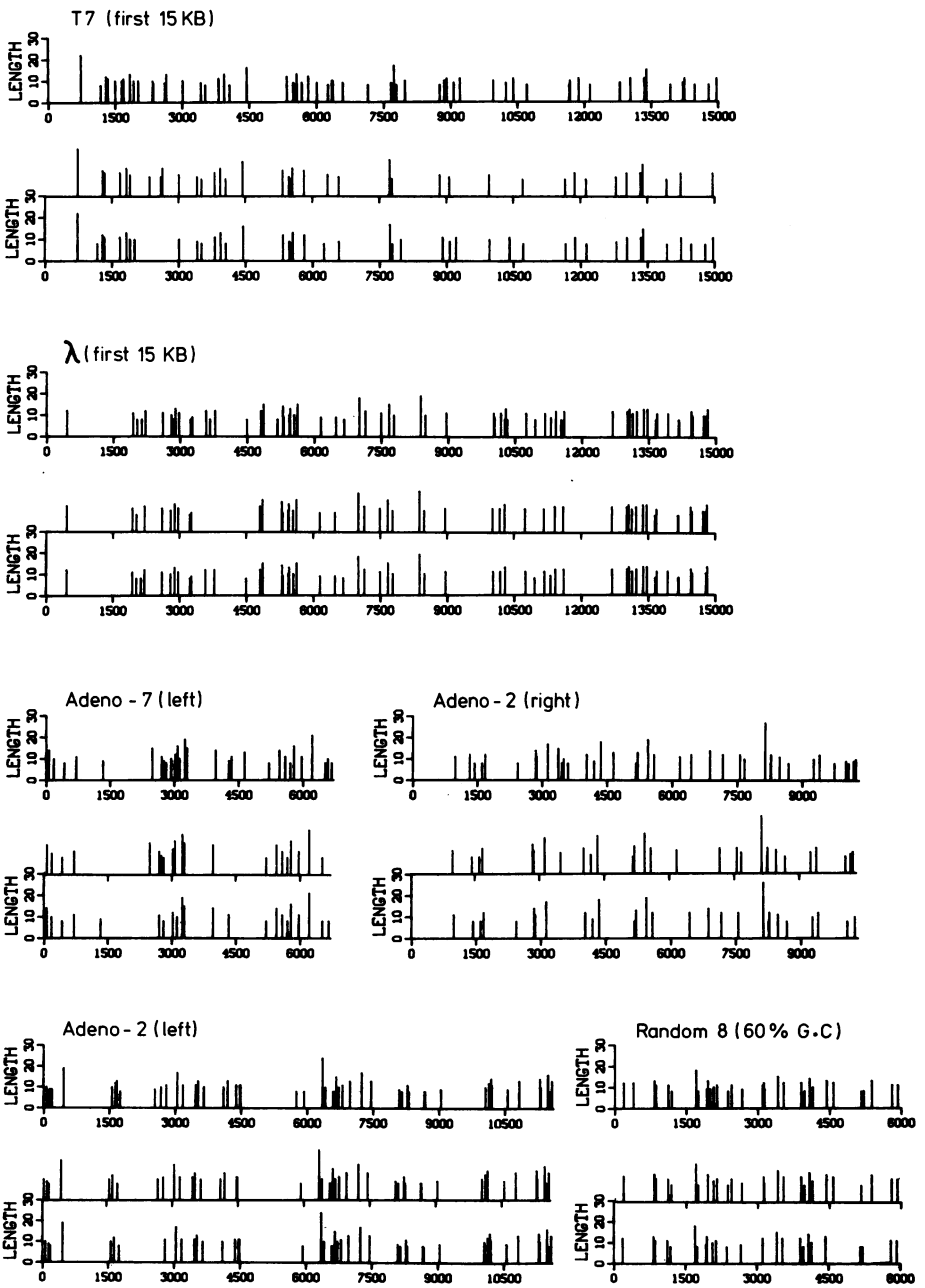

Fig. 3. Linear maps of qAFs and potential Z-DNA sites in dsl-DNA of phages T7 and λ and adenoviruses. Conventions as in Fig. 2. Random 8 is a random sequence rich in G and C (30% each).

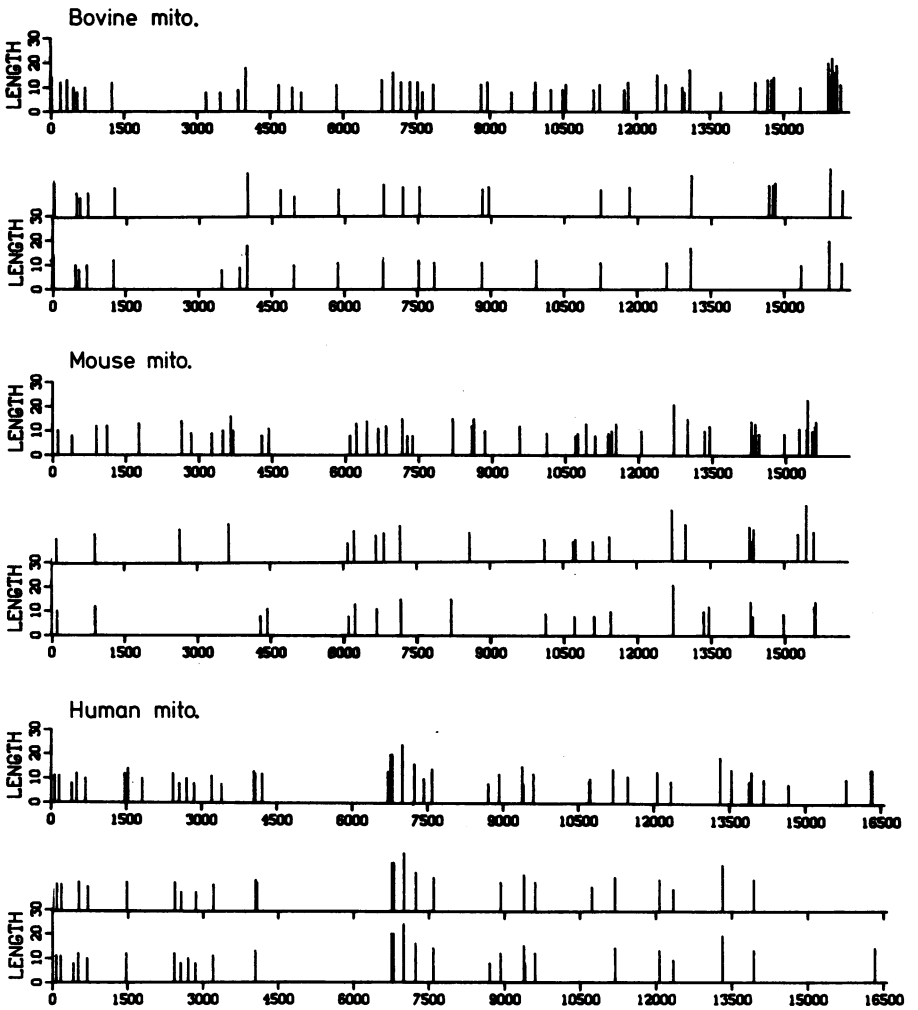

Fig. 4. Linear maps of qAFs and potential Z-DNA sites in mitochondrial genomes. Conventions as in Fig. 2.

DISCUSSION

We find that in the DNAs studied:

1) Alternating purine pyrimidine tracts are under-represented compared to the expectation for random sequences.

2) Uniform alternating fragments $[d(G-C)]_n$ and $[d(C-G)]_n$ are in general absent in circular DNAs and seem to be over-represented in DNAs of linear eucaryotic viruses.

pBR322

81	8	1	-	CGTGTATG
237	9	-	-	tatatgcgt
258	9	1	2	taTGCGCAC
507	11	1	2	GCGTG G GTatg
560	9	1	2	TGCATGCAC
1233	10	1	2	CGCGT TGCGT
1410	14	1	2	CGCACGCG GCGCAt
1452	9	1	2	GTGCGCATG
1709	11	-	2	tACAT C TGTat
1767	11	1	2	CGCAT C CAtac
2107	9	1	2	ACACATGCA
2243	8	-	-	gtgtatac
2290	19	1	2	GTGCAC CAtaTGCG GTGTG
2315	14	1	2	CGCACA G ATGCGTa
2785	10	1	2	TGTGTGCACG
3099	8	1	2	tACGCGCA
3266	9	-	-	gtatatatg
3696	11	1	2	CATGT TGTGCA
3873	9	1	-	GTGTATGCG
3924	11	1	2	CGCGC CACAta
4212	8	-	-	atacatat
4256	9	1	2	CGCGCACAt

øX174

763	14	1	2	TGCGTGTACGCGCA
811	8	1	2	ACGTGCGT
826	8	1	2	tACGTGCG
1027	14	1	2	gcgtaTGC CGCATG
1816	10	1	2	TGTGC CGCGT
2146	8	1	2	ACGCGTAC
3504	11	-	-	gtatatgcaca
3555	10	1	2	CGCGT TGCGT
4161	8	1	-	CGTATGCA
4192	10	-	-	atatgtatgt
4537	10	1	-	ATGCA ATGCG
4742	10	-	2	gtaTG GTACA
4773	12	1	2	CATGCG GTGCAC
4911	8	1	2	tACACGCA
5345	9	1	2	taTGCGCGC

M13

676	13	-	-	tatgtat c tgcat
1439	8	1	-	TATGCGTG
1930	11	-	-	tatac t tatat
2157	8	-	-	catgtatg
2802	11	-	-	tatgtatgtat
2828	12	1	2	acatAC TGCGTa
3794	15	-	-	tgtata acgcatatg
3990	8	-	-	tacatata
4173	9	-	-	cacatatat
5032	12	1	2	TGTGC CACGTat
5454	15	1	-	GCACGT TATACGTGC
5486	8	1	2	GTACGCGC
5531	13	1	2	tACGCGCA GCGTG
6000	9	-	-	gtacatatg

SV40

126	8	1	2	aTGCATGC
198	8	1	2	aTGCATGC
258	8	1	2	ACACACAt
1056	12	1	-	GCGTGTA ATACA
2752	10	-	-	atgtg gtatg
3218	11	1	2	GCACAC TGTAC
3575	14	1	2	TGTACAC CATGCAt
4208	12	1	2	taTGC C TGTGTG
4367	10	1	2	ACACA GCATG
4606	13	-	-	acacat tatacac
4675	10	1	2	TGTGT CACAC
4825	9	-	-	gcatatgca
4876	11	1	-	GTACAT TGCAT

PM2

118	9	1	-	TGCGCATAC
129	8	1	2	CGTGTGTa
212	12	1	2	CGCGC CAtacgc
345	8	1	2	CGCGCACA
483	8	1	2	GCACACGC
528	8	1	2	ACACGCGC
699	11	1	2	TGTGC T TGCGT
770	15	1	-	TGTAT TGCAT TGCGT
812	11	1	2	ACGCAT CACGC
1194	8	1	2	GCGCGCGT
1205	8	1	2	GTGCGCAt
1494	11	1	2	CGCGCACGTat
1542	8	1	-	CATACGCA

Polyoma-A2

420	8	1	2	tACATGCG
536	12	-	-	tgtatat tatgc
632	8	1	2	tACATGCA
675	8	1	-	TGTATGCA
723	18	1	2	aTGTGCACA GCGTGTata
1033	16	1	2	atACAC C CGCACAtac
1176	8	1	2	GTACATGC
1909	8	-	-	catgcata
2255	11	1	2	ACGCA ACAtac
2383	10	1	2	GTGTG GCACA
2568	12	1	2	TGTGTGTGTACA
2881	14	-	-	acatatgtg atata
2966	16	-	2	acatA GCGCGT CAtat
3220	12	-	2	tataT C TACACA
3657	9	1	2	CATGCACAt
4215	11	-	-	tgtatatgtgc
4447	10	-	-	atgtatgtgt
4495	10	1	2	TGTGT TGTGT
4524	9	-	-	cacacatat
4567	12	-	2	CATGTAC TACAt
4613	8	-	-	gtatatat
4777	11	1	2	ACACA G GCACA

BKV

822	8	-	-	acatatgc
1341	9	1	2	CGTGTACAt
1429	8	1	2	GTGTACAt
1969	13	1	2	GTGCATG A GCATG
2526	13	-	-	tatgtatg gtatg
2667	13	-	-	tgtacatatacat
2769	12	1	-	TACACA A ATGCA
2941	16	-	-	catgcatatat tatat
2972	10	-	-	gtatacatac
3569	10	1	2	atACA GTGCA
3907	10	-	-	tacat tgtat
3928	8	-	-	atacatac
4076	10	-	-	gtatg gtatg
4194	12	-	-	tatgt c tatgtg
4235	12	1	2	GCACACATGTGT
4316	9	-	-	gtatacaca
4812	12	1	2	acatA GCATGCA

Adeno-2·r

984	11	-	2	GCGCG G GCACA
1317	12	-	-	tgtatgctatat
1432	8	1	2	GCGCACGC
1607	8	1	2	tACATGTG
1679	12	1	2	tACATG A GCGCG
2431	8	1	-	TACACGTA
2843	14	1	2	ACGTACG A GTGCGT
2867	11	1	2	TGCGC C CACAC
3124	17	1	2	TGCACGCACGT t tgtac
3365	15	-	-	gtatgctgtatatgc
3449	8	-	-	tatgtata
3488	10	-	2	CATGTACATG
3591	8	-	-	gtatgtac
4024	12	-	2	GCGCA G ACGCGC
4194	9	1	2	TGTGCGTGC
4347	18	1	2	TGTGTGCGCat TGCGTAC
4642	13	-	-	catatatc catac
5162	8	1	2	CATGTGTG
5198	13	1	2	acata a ACACACA
5436	19	1	2	cgtgtatc cataTG ACACA
5576	12	1	2	GCATGC T TGCGC
6170	11	-	2	GCATGCGTata
6431	12	1	-	ATACA A ACACAT
6866	14	1	-	CATACACA A ATGCA
7163	12	1	2	TGCATGT tatgt
7556	12	1	2	CACGC C TACATG
7660	10	-	2	atACA ACATG
8135	27	1	2	GCACA ACACA G GCACACGTGCAtacac
8258	8	1	2	CGCACGTa
8269	12	1	2	CACGT TGTGCAt
8466	11	1	2	GTGCG G GCGTG
8674	8	1	2	CACACACG
9258	10	1	2	ACACGCAtac
9391	12	1	2	CATGC T CATGCA
9737	8	-	-	atacatac
10001	9	-	-	gtatatata
10061	8	1	2	CGCACGCG
10176	9	-	2	atACATGCA
10232	10	1	2	CGCGC CACGT

Adeno-2·l

65	10	1	2	ACGTG GCGCG
140	9	1	2	ACACATGTa
180	8	1	2	GTGTGCGC
190	9	-	-	gtgtatacg
457	19	1	2	ACGCGCA GTGTat t tatac
1559	10	1	2	TGCGTGTGTG
1634	12	1	2	TGCATG GCGTGT
1668	13	-	-	gtatata atgcgc
1746	8	1	2	TGTGCGTa
2532	9	-	-	gtatatgca
2666	10	-	2	aTGTA GCATG
2788	11	1	2	tACACG GTGTa
3041	17	1	2	gcata ACATG GTGTGTG
3170	11	1	2	GCACA ACAtac
3459	11	1	2	TGCAC C CGCGC
3513	13	1	2	aTGTGTG G GCGTG
3648	10	1	2	ACGCGCATGC
4095	10	1	2	TGCGCGCGCG
4189	13	-	2	atACATG G GCAta
4384	11	1	-	GTGCATACGTG
4465	11	1	2	CATGT TGTGCA
4485	11	1	2	GCACA GTGTat
5753	8	-	-	cgtataca
5925	8	1	2	ACACATGT
6340	24	1	2	TGCACGTat T CGCGCGCA ACGCAC
6405	10	1	2	GTGCACGCGC
6585	8	1	2	GCGCGCGT
6637	8	1	2	CATGCGCG
6653	15	1	2	GCGCGCGC T CGTatg
6712	10	1	2	GCGTACATGC
6799	11	1	2	GCGCGCACGTa
6966	13	1	2	CGCGT CACGCACG
7238	17	1	2	gcgcgtatGC C TGCGCG
7448	13	1	2	TGCGTGTG ATGCG
8085	9	1	2	ACGTGCGCA
8147	8	1	2	CGCGCACA
8274	11	1	2	CACGC CGCGCG
8304	8	1	2	CGCGCGCG
8656	8	1	2	CGCGCGCG
8679	8	1	2	TGCGCGCG
9036	9	1	2	GCGCGCATG
10049	10	1	2	CATGCGTGTG
10128	12	1	2	TGCAC C TGCGTG
10178	14	1	2	gtaTGCGC C CGTGT
10197	11	1	-	GTGTA A GTGCA
10559	9	1	2	GCGTGCGCA
10821	13	1	2	CGCGCGCG GCGTa
11294	14	1	2	GCGTG ACACGCGCG
11310	10	1	2	GCGTACATGC
11477	16	1	2	CGCGCGCGCACACGTG
11513	8	1	2	CGCGTACG
11568	13	1	2	CACGTGCGCACGC

Adeno-7·l

6	12	-	-	tatata atatac
65	14	1	2	GTGCGCGC TGTGTG
178	10	1	2	aTGTG ACGTa
425	8	1	2	CGCGTACG
704	11	1	2	tACGC TGCACG
1327	9	1	-	ACGTATGCA
2481	15	-	2	aTGCATGC TACAtat
2700	11	1	2	atACGTGTGTa
2750	9	-	2	taTGCATGC
2806	8	1	2	aTGCATGT
2923	10	-	-	gcata atatg
3007	9	-	-	cgtgcatat
3021	12	1	2	CACATGCACGCA
3052	10	-	-	acata atgtg
3070	16	-	2	GTGCAC CATGCAtata
3118	10	1	-	GTGTA ACATG
3245	19	1	2	GTGCACGCATGCG A atgcg
3290	15	1	2	GTGTGCGTG G ATGTG
3962	14	1	2	CGCGCGCG GTatgc
4267	9	-	-	tatatgcat
4330	11	1	-	CATAT TGTGCA
4635	13	-	-	cgatata t tgcat
5205	8	1	2	GTGCGCGC
5445	14	1	2	gtaTGCAT C CGCAC
5579	11	1	2	CGTGT C CACGC
5701	8	1	2	GCGCGTGT
5782	16	1	2	gtatgta a ACACATGT
5963	11	1	2	aTGCG G GCATG
6208	21	1	2	ACGTAC T CGCGCGC CACACAt
6515	8	1	2	GCGCGCGC
6574	10	-	-	gcatacatgc
6664	8	1	-	CGCACATA

Fig. 5. Complete listings of qAFs in papovaviruses (SV40, BKV and Polyoma-A2), adenoviruses (Adeno-2·l, Adeno-2·r and Adeno-7·l), cloning vector pBR322, dsc-phage PM2, and ssc-phages M13 and øX174. In every case, the first and second columns show the position and length of the qAF, respectively. A digit 1 in the third column means that the qAF is a potential Z-DNA sequence according to the definition 1. A digit 2 in the fourth column indicates a potential Z-DNA tract fulfilling definition 2. Non-concordance is indicated by a dash. The sequences are shown to the right. Spaces denote the separation between AFs constituting a cAF. Lower case letters denote bases which do not belong to a potential Z-DNA fragment or subfragment whereas capital letters denote the (sub)fragments with Z-forming potential.

Bovine

19	14	1	2	CATGC T CACACAta
204	12	-	-	tatat tatatat
337	13	-	-	atatata a acgca
473	10	1	2	tACACATGCA
545	8	1	2	GCACACAC
709	10	1	2	GTACG GCGTa
1256	12	1	2	GCACGCACACAC
3188	8	-	-	tatatata
3478	8	1	-	ATACGCAC
3846	9	1	-	CACACATAC
3999	18	1	2	TGTGCATGTG ACACGTat
4679	11	-	2	tataT C TGTAC
4957	10	1	2	CACACACATG
5142	8	-	-	tatatacg
5863	11	1	2	CGCACACGCAt
6792	13	1	2	acacatAC TACGT
7020	16	-	-	atgcatacaca atatg
7188	12	-	2	catat CACACAt
7371	12	-	-	catatg gcatat
7527	12	1	2	catACA A GCACG
7624	8	-	-	tatacata
7846	11	1	-	ACGTAT TACAC
8820	11	1	2	tACAC T TGCAC
8948	12	-	2	tataT C TGCATG
9440	8	-	-	gcatatat
9905	8	-	-	atgtatac
9929	12	1	-	CATACG A ATGTG
10248	9	-	-	gtatacata
10486	9	-	-	atacatatg
10503	8	-	-	tatgtaca
10555	11	-	-	tatac t tatac
11120	9	-	-	gcatgtata
11248	11	1	2	GTACG GTatgc
11751	9	-	-	tatacatgc
11829	12	-	2	aTGCAC T CAtat
12421	15	-	-	atatg atatatatac
12596	11	1	-	ATGCA A ACACA
12936	10	-	-	tacat tatgc
12976	8	-	-	atatgcgc
13090	17	1	2	CACAT C TGTAC C CACGC
13721	8	-	-	catacata
14422	12	-	-	ataca atgtata
14670	13	-	2	atACAC TACACAt
14753	13	-	2	atacatACACGCA
14794	14	-	2	tatataTGCACGTa
15341	10	1	-	TGCATACGCA
15912	20	1	2	acaatAACACGC C CAtacaca
15964	9	-	-	atgtacata
15990	22	-	-	acata atatgtatata gtacat
16026	10	-	-	catgcatata
16060	16	-	-	gtacata atacatata
16085	19	-	-	tgtacata gtacat tatgt
16168	11	1	2	CATGC CGCGTG

Murine

111	10	1	2	tACACATGCA
402	8	-	-	tatatgtg
895	12	1	2	GTACGCACACAC
1114	12	-	-	tacaca a atata
1767	13	-	-	acatacgcgtata
2642	14	-	2	atatatACGTACAC
2837	9	-	-	gtacataca
3260	9	-	-	acacatatg
3487	10	-	-	tatac tatat
3649	16	-	2	tatgtaTGTG ACAtat
3697	10	-	-	catacatata
4284	8	1	-	TGCACATA
4430	11	1	-	ACACA A ATACG
6091	8	1	2	CACATGTa
6220	13	1	2	aTGTA G ACACACG
6433	14	-	-	atacatac tatgta
6666	11	1	2	tACAC CACATG
6829	12	-	2	catat CACACAt
7163	15	1	2	acacatACA A GCACA
7263	8	-	-	tatatata
7371	8	-	-	catatata
8192	15	1	-	CACATACAT T TACAC
8585	12	-	2	tataT C TACATG
8622	15	-	-	catgcatat cacata
8846	10	-	-	atatg gtata
9562	12	-	-	catatg a atgcg
10118	9	1	2	ACACGTACG
10707	8	1	2	ACACACAC
10752	9	-	2	GCATGCAta
10925	13	-	-	atatata gcatac
11111	8	1	2	CGCACATG
11371	8	-	-	gtatatac
11382	9	-	-	atatacata
11446	10	1	2	CACACACACG
11535	13	-	-	atatgtg a atata
12053	10	-	-	atatatacac
12715	21	1	2	tACACAT C TGTAC C CACGCA
12995	15	-	2	taTGTACA GCAtacg
13338	10	1	-	CGCAT TACAC
13443	12	1	-	CACACA A ACATA
14296	14	1	2	catACAC TACACAt
14337	8	1	2	ACACACAC
14378	13	-	2	atatatACACGCA
14455	9	-	-	catatacat
14966	9	1	-	TGCATACGC
15274	11	-	2	tataT C CATGT
15429	11	-	-	tacata gtaca
15443	23	-	2	gtacat t tatgtataT CGTAC
15545	10	-	-	atatatatac
15596	12	1	2	acataT C TGTGT
15614	14	1	-	ACATACAC CATACA

Human

71	11	1	2	gtaTGCACGCG
162	11	1	2	.CGCAC C TACGT
413	8	1	-	GTATGCAC
513	12	1	2	GCACACACACAC
690	10	1	2	tACACATGCA
1472	12	1	2	GCGCGTACACAC
1535	14	-	-	tacgcat t tatata
1816	10	-	-	gcata atata
2430	12	1	2	ACACA G GCATGC
2554	8	1	2	ACACATGT
2702	10	1	-	GCATA ACACA
2849	8	1	2	GTACATGC
3200	11	1	2	tatAC C CACAC
3394	8	-	-	tatataca
4042	13	1	2	acataTG ACGCAC
4069	12	-	2	tACACA ACAtat
4209	12	-	-	tatatg atatgt
6710	13	-	-	atacata g gtatg
6761	20	1	2	CGTGTG A GCACAC CAtatat
6796	20	1	2	ACGTA G ACACACG A GCAtat
6998	24	1	2	CGTAC TACACG ACACGTAC TACGT
7237	16	1	2	atgcatACAC CACATG
7427	10	-	-	cgtatacata
7589	14	1	2	GCACATGCA GCGCA
8703	8	1	-	CATACACA
8921	12	1	2	GCACAC C TACAC
9380	15	1	2	GCGCG ATGTa acacg
9400	8	1	-	GCACATAC
9604	12	1	2	ACACAT C CGTat
10717	9	-	-	acacatatg
10734	10	-	2	tACGTACATa
11190	14	1	2	ACGCA G GCACAtac
11484	11	-	-	gtata atacgc
12067	13	1	2	CATGT T CAtacac
12342	9	1	2	CATGCACAC
13316	19	1	2	TGCACAT C TGTAC C CACGC
13535	14	-	-	acatat catacaca
13886	9	-	-	tatgcacat
13932	13	1	2	CACACAC CGCACA
14177	10	-	-	atatatacac
14663	8	-	-	gcatacat
15811	10	-	-	cgtac tatac
16303	14	-	-	gtacata gtacata
16328	14	1	-	CGTACATA GCACAT

Fig. 6. Complete listing of qAFs in mitochondrial DNAs. Conventions as in Fig. 5.

T7

731	22	1	2	ACACG GTACG ATGTAC CACATG
1174	8	1	-	GCGCGTAT
1291	12	1	2	ACGTGCAatacg
1336	11	1	2	ACATG GTGCGC
1496	10	-	-	tacgta atgc
1648	10	-	-	tacgt tatgc
1683	11	1	2	cgtaTGTACAC
1827	13	1	2	GCACAC Tatacgc
1909	10	1	2	GCGTG GCAta
2011	10	1	-	ACACA ATGCA
2333	10	-	-	tatac cgcat
2348	9	-	2	CGCATGTat
2604	9	-	2	ACATGCAta
2640	13	-	2	ACGTAC CAtacat
3007	10	1	2	CGCGC TGCAC
3416	9	1	2	TGCACGCAt
3513	8	1	2	GCGTACAt
3810	11	1	2	GTACGC TGCAt
3939	13	1	2	CGTGCA G GTGCGC
4059	8	1	2	GTGCGTAC
4443	16	1	2	cgtgt t tACGC TGTGT
5337	12	1	2	TGCGC TGTGCAt
5479	9	1	2	aTGCACACA
5516	8	1	2	TGTACACA
5556	13	1	2	GTGTG G GCACACG
5671	9	-	-	acacatatg
5810	12	1	2	CGCGT T CGCGTa
6005	9	-	-	gtatgcgta
6254	8	1	-	TGCATACG
6339	10	-	2	ataTGTGTAC
6362	10	-	-	catatgtatg
6584	9	1	2	GTGTACGCG
7147	8	-	-	atgtgtat
7661	9	-	-	tgcgtatat
7728	17	1	2	GTGTGTGC T TGCGTGTG
7782	8	1	2	GTACACAC
7971	10	1	-	CGTATGTGCG
8748	8	-	-	tatgcata
8849	10	-	2	aTGCGTACAt
8905	11	1	-	ACGTA A ATACA
9063	9	1	2	CGTGTGCGT
9203	11	1	-	TACGC T TACAT
9953	10	1	2	GTGTG GTGCA
10236	9	-	-	atatacgca
10404	11	1	-	CACAC T TACAC
10712	8	1	2	CGTGTACA
11660	8	1	2	ACACACGT
11673	10	-	-	tacgtatgcg
11872	11	1	2	GTGTG ATGTAC
12126	8	1	2	ACATGCGC
12795	9	1	2	tACATGCGC
13033	11	1	2	GCGCAC TACGC
13345	11	1	2	ACACAC CATGT
13391	15	1	2	cgtaTGTACA ACGTG
13926	8	1	2	CATGCGTG
14208	9	-	-	atatgcacc
14243	11	1	2	CACGCG GTGCA
14471	8	1	-	GTGCGTAT
14784	8	1	-	GCGTATGC
14958	11	1	2	atACAC TACGT
15007	11	1	2	CGTGC TGCATG
15077	8	1	2	TGTACGTa
15342	11	1	2	ACATGT TGTGT
15466	9	1	2	GTACGTGTa
15567	8	1	2	GTGTACGT
16153	8	1	2	GTACACGT
16290	23	1	2	TGCGTACATG GCATG G GTACATG
16599	8	-	-	tgtatgta
16800	9	1	2	CATGTGTGT
16933	18	-	2	ataTGC TGCAT C Tatgta
17016	12	1	2	ACACACG GTatg
17049	10	1	2	GTACGTACAt
17080	9	1	2	ACATGCACG
17213	8	-	-	tatgtacg
17384	13	1	2	CACGCG ATGTGTG
18405	8	-	-	gcgtatat
18697	9	-	2	taTGCGTAC
18732	9	1	2	taTGCGTGC
19140	10	1	2	CACATGTGTG
19308	13	1	2	ACATGTA GCGTAC
19323	13	1	2	CGCGTG GTGCGTa
19504	12	-	2	aTGCGTGTatac
19929	13	1	2	cacatA G GCACAt
19985	8	-	-	atgcatat
20323	8	1	2	ACACGCGC
20873	16	1	2	ACGTGTACAC T CACAt
21962	18	-	2	aTGCA G ACGTatatgcat
22356	10	1	2	GCGTG GCATG
22392	11	1	2	GCGTG A GTACG
22559	12	-	2	aTGCAC TGTata
22958	11	-	-	atatacatatg
23113	12	1	2	CACATG GTACGT
23399	11	1	-	TGTGTA ACGTG
24092	15	1	2	CGCACGCAC C CGCGC
24224	21	-	2	tataTGCGC T CAtacg atatg
24336	11	1	2	ACGCA G ATGCA
24460	8	-	-	tgtatata
24530	10	-	-	tatgtgtatg
24766	11	-	-	gtaca atatgc
25072	10	1	2	tACGC TGTGT
25338	9	1	2	TGTACACAt
25450	8	1	2	aTGCGCAC
25761	12	1	2	CATGC CACACGC
26260	11	1	2	tACGC TGTGCA
26304	12	1	2	tACAT CACACGT
26412	8	1	2	CATGTACA
26519	11	-	-	atatgtatgtg
26832	15	1	2	ACGCATG GTGTACAt
26865	8	-	-	cgtatatg
27135	9	-	-	tgtatacat
27422	11	1	2	GTGTGT CACGT
27603	8	-	-	tacgtatg
27671	9	1	2	GTACACACG
27724	9	-	-	catatgtgt
28484	10	-	-	atatgtatac
28768	8	-	-	gtatacgc
28856	15	1	2	tgtaTG GTGCGCATG
28952	11	1	2	ataTGC TGCGT
29838	10	-	-	tgcgc tatac
30048	8	1	2	TGCACGTa
30194	10	-	2	ataTG ATGCA
30748	10	1	2	GCATG ATGCA
30911	8	1	2	TGCGTACA
31005	10	-	2	tACATGCGTa
31457	9	1	2	TGCGTACAt
31786	12	1	2	GCGCA G GCATGT
32595	10	-	-	cacgtatatg
32985	10	1	2	CGCAT CATGC
33126	8	1	2	GTACATGC
33331	12	1	2	TGCGTG Atacac
33728	14	1	2	acataTG GTACGCG
33847	12	1	2	TGCGT C CACAta
34055	9	1	2	TGCACACGT
34068	8	1	-	GCATACGC
34109	8	1	2	GCGTGCAC
34137	10	1	2	CACGC TGCGT
34781	9	1	2	tACACGTAC
34905	10	1	2	CGCGCGTatg
34941	9	1	2	aTGCACGTa
35279	11	-	2	atACGC TACAt
35470	8	1	2	aTGTGTAC
35616	8	-	-	atgtatgt
35769	11	-	2	GCGCAC Tatat
35859	15	-	2	taTGTACATG GTACG
36546	9	-	-	atatgtatg
36611	8	1	-	GCGTATGC
37569	11	1	2	CACATGTGCGT
38097	12	1	-	TATGT TACGCGC
38180	12	1	2	TGCGCG A GCGTG
38396	11	1	-	TGCATACGTAC
38491	19	1	2	tACGC TGTGC TGTAC
38731	8	-	-	atgcgtat
39015	13	1	2	ACACATG ATGCGT
39042	9	-	-	tacgcatat
39077	8	1	2	aTGTGTAC
39290	9	-	-	gtatgcata
39334	10	-	-	tatat tgtat
39465	8	1	2	GTACGCAt
39586	10	1	2	TGCAT CACAC

Fig. 7. Complete listing of qAFs in phage T7. Conventions are as in Fig. 5.

λ

457	12	1	2	tACGCGTGCGCA	16217	13	1	2	TGCGCG G GCGTat	33679	9	-	-	catatgcat
1935	11	1	2	aTGCGCGTatg	16481	8	1	2	aTGCGTAC	33717	11	1	2	GCATGT TGCGC
2026	8	1	2	TGCGTGTG	16494	11	1	2	CACACA GCGTa	34035	12	1	2	atACACA GTGCA
2129	8	1	-	TGTGTATG	16540	13	1	2	aTGCGC TGTatgc	34553	10	-	-	catgtatatg
2212	12	1	2	GCATGC CACGTa	16834	11	1	2	GCACAC CGCGC	34641	14	1	2	CATGTGTG GCATGT
2604	11	1	2	tACGC C CGTGC	17169	12	-	2	atACA GCGTatn	34690	8	1	-	ATGTGCAT
2796	10	1	2	GCATG ACACA	17249	10	1	-	GCACGTATGC	34819	8	1	2	TGTGCGCA
2831	8	-	-	tatgtatg	17265	14	1	2	CGTGCA GCATGTGC	34937	8	-	-	gcatatat
2888	13	1	2	TGCGCG Aatatgc	17697	10	1	2	CACGC CGCAt	35081	17	1	2	TGCATGTtaTGC CGCGT
2966	11	1	2	GTGTG G ATGCA	17789	8	1	2	GCACCGT	35111	14	1	2	TGTAC CATGTGCGC
3228	8	1	2	ACGCACGT	18205	10	-	-	cgtat cgcat	35343	14	1	-	CCGTGT T TGTGCAT
3264	9	1	2	CATGCACGC	18398	11	1	-	GTGTG A ATGCC	35482	12	1	2	cgtaT CACACAC
3575	12	1	-	GTGCA A ATGTGT	18795	18	1	2	cacgc t tACGT C CACACG	35868	11	-	-	atgcatatatg
3664	8	-	-	atgtatgc	18833	9	-	-	cgtatacgt	36017	13	-	-	catatac c cutac
3780	12	1	-	CGCGTAT TACGC	18965	17	1	2	atgcgta aTGTGTGTat	36110	8	-	-	tgcgtatg
4493	8	1	-	TATGCGCG	19042	11	1	2	ACATG GTACGT	36665	13	-	-	atgcatatgcata
4803	12	1	2	ataTG G ACACGC	19271	12	1	2	taTGTG A GCGTG	37131	10	-	-	tatgt tatat
4862	15	1	2	CGTGTGCGTG ACAta	19320	10	1	2	GTGCGTACGC	37890	8	1	2	ACGTGCGT
5171	8	-	-	gcatatac	19437	16	1	2	gtgca aTGCG GCAtac	37932	13	1	2	TGTGC T CAtacgt
5293	14	1	2	CATGCGTG ATGCAC	19467	13	1	2	GTGCA G GTataca	37980	9	1	2	CGTGCGTGT
5310	9	1	2	aTGCACGTa	19798	11	1	2	GCATG G ATGTG	38020	9	1	2	TGCATGTAC
5436	10	1	2	GCGCG GCGCA	19877	8	1	-	CGTGTATG	38068	12	1	-	TATGCA ATGCGC
5451	13	1	2	ACGTG A ACGCGCA	19989	13	1	-	GCGCGTA ACGCGT	38111	8	-	-	gcgtatat
5546	10	1	2	GCACGCGTGC	20103	8	1	2	GCACGCGC	38303	10	-	-	atgtatgcat
5619	15	1	2	GTGCACA G GCGCGCA	20652	9	1	2	atACGTGCA	38318	9	-	-	tgcatacat
6141	9	1	2	aTGTACACA	20713	11	1	2	aTGCG G ACACA	38355	8	-	-	tacatatg
6478	9	1	2	ACATGCGTG	21410	10	1	2	CGTGC TGTat	38742	10	1	-	GCGTA ATGTG
6662	8	1	-	CACGTATG	21617	12	1	2	CATGC C CACACA	39364	13	1	2	aTGCCTG A ACGTG
7003	18	1	2	aTGCG G ACGCACA GCGCG	21798	17	1	2	GTGCACATGCGCAtaca	39395	12	1	2	ACATGTGTGTGC
7139	12	1	2	CGTGT C CGTACA	21825	8	1	2	GTGCGCAC	39846	9	1	2	aTGCGCGTa
7490	11	1	2	CACGC CACGCG	21882	13	1	2	CACACG G ACACAC	40033	8	1	2	ACGCGCAC
7666	15	1	2	atACGCG G GTACATG	22058	10	-	-	acgcatatat	40215	8	1	2	CGTGCACA
7775	10	1	2	GCGTGCGCGT	22398	11	1	-	GCATAT TGCGC	40698	11	1	2	CACGCA GTACA
8375	19	1	2	GCATGT CATGCA GCGTGTG	22896	9	-	-	tgtatgtat	40862	8	1	2	GTGTGTGT
8681	10	1	2	atACG GCGTG	23234	15	-	-	catat tgtat catgc	40953	9	1	2	aTGCGTGTa
8954	11	1	2	taTGC C TGTAC	23847	8	-	-	atatgtat	41166	13	1	2	ACGTGCAt tacgt
10015	11	1	2	ACGTG GTGTAC	23997	10	-	-	atgta ataca	41306	11	-	-	tatacg gtatn
10036	9	-	-	gtatgtatg	24168	14	1	2	ACACGT CGTatgca	41477	11	1	2	tACAC CACGTG
10177	11	1	2	GCGTGC CATGC	24369	8	1	2	GTGCATGC	41738	10	1	2	CATGTGCGTG
10266	8	-	-	atatgcac	24857	9	-	-	tgcatatac	41853	10	1	2	ACATG GTACG
10286	13	1	2	ACGTACGC CGTGC	25329	10	-	2	TGCAT CAtac	41947	11	1	2	TGCATG ATGCG
10323	8	-	-	atatgcat	25658	8	1	2	CACATGCA	42320	8	1	2	ACGTGCGC
10745	11	1	2	CATGCGTatgc	25725	9	1	2	TGCGTGCAt	42345	9	1	2	TGCATGTGC
10947	8	1	-	GCGTATGC	26526	11	1	2	ACACACGTGCG	42362	10	1	2	CACGTGTGTG
11164	11	1	2	CTGTG GCGCGT	27054	9	-	-	tacatacgt	42839	14	1	2	aTGCGTG GTGTGCA
11300	9	1	-	TGCGTATGT	27162	17	1	2	CATGT TGCATG GTGCAC	42880	10	1	-	CGTAT CGCGT
11410	12	1	2	ACATG G GCACGC	27206	11	-	-	atgcatatata	43169	8	1	2	aTGCACGC
11531	8	-	-	tgcgtata	27370	10	-	2	gtaTGCATGC	43436	8	1	2	TGTACGTG
11595	12	1	2	GCGCG G GCGCGT	27628	14	1	2	tgcataTGT TGTGT	43756	12	1	2	GCGCG G ACACGT
12685	12	1	2	GCGCAT C GCGGT	28005	12	1	2	CGTGCGCGCAta	43818	8	-	-	atgtatat
13017	12	1	2	gtgtaTG ACATG	28652	10	-	-	atgta acgta	44104	8	1	2	TGCACACA
13059	13	1	2	ACACA G ATGCGTG	28699	11	1	-	TGTGT T TGTGT	44358	16	1	2	GCGCA GCAtat cgcgc
13131	11	1	2	gtaTG G ATGTG	29046	8	1	2	GCGTGCGT	44779	12	-	2	tataT C TGCACA
13226	12	1	2	ACGTACA GCGTG	29096	8	1	2	tACACGCA	44903	9	1	2	aTGCGTACA
13380	13	1	2	gtaTGC TGCGTGT	29163	15	1	2	tACGT C TGCATGTGC	45005	8	1	2	tACACATG
13461	13	1	2	aTGCAC C CGTGCG	29182	11	1	2	TGCGC C CAtat	45069	9	-	-	cgtgcatat
13649	8	1	2	GTACGCGC	29533	11	-	-	tatat c tatac	45333	8	1	2	aTGTGCGC
13679	11	1	2	TGTACG GTatg	29672	8	-	-	atgtatgt	46123	14	1	2	ACATGCA G ATGCGT
13923	11	1	-	ACGTAT CATGC	29881	8	-	-	catatgca	46471	12	1	2	atACACACGCGC
14160	8	1	2	GCGCACGC	30560	10	1	2	TGTGT TGCGC	46789	10	-	-	cgtatgtgta
14174	8	1	2	aTGTGCGC	30951	13	1	2	gcataA GCACACA	46856	8	-	-	tgtatgt
14444	12	1	2	TGTGTGC CGCAt	31190	10	1	-	TGTGCATACA	46960	9	1	2	aTGCACACA
14476	10	1	2	GCACG GTGTG	31469	10	-	-	acata atgca	46976	24	1	2	GCGCGTACACGTat tgcat tatgc
14724	10	-	2	TGCAC CAtat	31492	8	1	-	CATACGCG	47237	8	1	2	aTGCGTGT
14782	10	1	2	ACGCACACAC	31541	17	1	2	CACATGC TGTAC TGCAC	47311	10	1	2	ACACG ATGTG
14810	13	1	2	GCATG GCGCGCAt	31688	10	-	2	GTACGCAtac	47346	10	-	-	atata atgtg
15221	10	1	2	GTGTG GCGCA	31953	12	1	2	TGTGC CACGCAt	47466	11	1	2	CATGTG Atacg
15259	8	-	-	cgtataca	32257	9	1	-	GTGTATGCA	47643	9	-	-	atgtatgta
15358	9	1	2	aTGCGCGTG	32493	9	1	2	GCATGTACA	47839	10	-	2	TGCATGTata
15371	8	1	2	CACGCGTG	33453	18	1	2	gcatatTGCATG GTGTGC	48132	10	-	-	atgtatatgc
15789	11	1	2	ataTG ACACGC	33479	8	-	-	tatacata	48430	9	1	2	CACGCACGT
15925	10	-	2	atACA ACGTa	33667	8	1	2	CACACACA					

Fig. 8. Complete listing of qAFs in phage λ. Conventions are as in Fig. 5.

3) Quasi-alternating fragments (qAFs) are under-represented in every case studied.

4) Potential Z-DNA sites are strongly under-represented in the circular DNAs studied (except the recombinant clone pBR322). A particularly strong under-representation (about 5-fold) is displayed by mitochondrial genomes.

5) Potential Z-DNA sites are under-represented in the linear genomes of bacteriophages T7 and λ but not in the linear Adeno-2 and Adeno-7 DNAs.

6) In all the cases studied potential Z-DNA occurs in non-coding as well as in protein-coding regions.

7) The tendency of short AFs to cluster decreases in the order: random sequences > procaryotic DNAs > mitochondrial and linear eucaryotic viral DNA > circular eucaryotic viral DNAs.

Comparison of predicted with experimentally mapped Z-DNA sites. The studies of anti Z-DNA Ig binding to plasmid pBR322 DNA (5, 8, 19) provide evidence for 3 major and other minor immunoglobulin binding sites mapped by electron microscopy techniques (8) with a resolution of about 100 bases (major sites) and 300 bases (minor sites). There is good correspondence (with the exception of site B at position 960 ± 80; ref. 8) with the potential sites we have identified at positions (Figures 2 and 5): 237, 258, 1410, 1452, 2107, 2290, 2315, 2785 and 3099.

The anti Z-DNA Ig binding sites detected in the SV40 genome by filter-binding studies (18) and by immuno-electron microscopy (27) show the existence of 3 major antibody binding sites in the nucleotide sequences associated with the transcriptional enhancers within the nucleosome-free "gap" region of the papovaviral chromatin. These sites occur at positions 126, 198 and 258 and are predicted in this paper (Figures 2 and 5). Three other potential Z-DNA regions predicted by our algorithm (positions 1056, 3218 and 3575 of the SV40 genome) may correspond to minor antibody binding sites observed in the electron microscopy studies. Thus, among 10 predicted potential Z-DNA sites in SV40, 3 and possibly 6 have been experimentally detected, at least within the resolution currently available.

Mapping of anti Z-DNA Ig binding sites in ϕX174 DNA provides further experimental evidence supporting the predictions of potential Z-DNA sites made in this paper. According to the listing (Fig. 5) and plot (Fig. 2) there are 13 potential Z-DNA sites in ϕX174 DNA. Nine of them (positions 763, 811, 826, 1027, 2146, 3555, 4161, 4911 and 5345) correspond well with antibody binding sites identified by high resolution darkfield electron microscopy (23). Revet et al. (23) identify a site (no. 8) at position 3542 ± 62 and consider its possible relationship to the sequence starting at nucleotide 3504. By our criteria, this fragment is rejected due to its high alternating A-T content (Fig. 5). However, we note as a potential site the qAF at 3555 (which meets both definitions) and which is within the resolution limits of the site identified by e.m.

Studies of the PM2 bacteriophage genome also provides evidence for the correspondence of anti Z-DNA Ig binding sites to tracts of purine-pyrimidine repetitions (24, 25). The immunoelectron microscopy mapping of anti-Z-DNA Ig binding sites in the purine-pyrimidine rich region of this phage DNA (26) shows the existence of Z-DNA within a protein coding region. There are 13 potential Z-DNA sites predicted by our algorithm (Fig. 5). Ten of them (positions 129, 212, 345, 483, 528, 699, 812, 1194, 1205, 1494) correspond well with antibody binding sites identified by immunoelectron microscopy (26).

The correspondence between experimentally mapped Z-DNA sites in supercoiled circular DNAs and those predicted by the criteria we have defined is satisfactory but not perfect [some experimental positions we cannot account for and others we identify have not (yet) been observed]. As additional data emerge, the specific values of the empirical parameters (a, b, c in definitions 1, 2) will require adjustment. In any event, we expect that they will depend on superhelix density and, to a degree, each other. For example, alternating fragments exclusively composed of G and C are under-represented (Table 1, below) but where they do occur (23, 26) the Z conformation may be expressed even for lengths smaller than the value 8 used in this work. In addition, we do not address the means for defining a *hierarchy* in Z-forming potential, for which the experimental data provide some indications. It is obvious that the ultimate but as yet unattainable goal will be to replace the empirical criteria employed here with rigorous thermodynamically defined relationships.

The under-representation of potential Z-DNA It has already been suggested that Z-DNA could play a role in the control of transcription (22). In the circular DNA molecules, such processes would be coupled to changes in the free energy of supercoiling. (Since the B to Z transition lowers the negative superhelix density, one Z-forming tract may affect the potential of another; 23). Thus, it would seem reasonable that in such genomes the number of sites allowing a B to Z transition would be limited and highly regulated. Furthermore, the genomes examined in this work are almost fully transcribed. For these various reasons, the observed under-representation of potential Z-DNA forming sequences is not unexpected. In this connection, it is noteworthy that $[d(G-C)]_n$ and $[d(C-G)]_n$ tracts are avoided in circular DNA genomes, whereas these are the sequences which undergo the B-Z transition most readily *in vitro*. One can envisage positive as well as negative selection processes accounting for this phenomenon. Clearly, the intervention of proteins with specificity for different helical conformations as well as other factors determining higher order structure of DNA *in vivo* will determine which of the sites we and others have identified actually undergo the B→Z transition and if so, whether functional roles are involved.

ACKNOWLEDGEMENTS

We thank Dr. E. Trifonov for discussions concerning the definition of potential Z-DNA sequences, and Drs. G. Hamm and K. Stüber for discussion and help in exploitation of EMBL Nucleotide Sequence Library. We are indebted to Dr. J.H. van de Sande for providing manuscripts prior to publication. Ms. Melanie Harvey is acknowledged for typing the manuscript.

*Present address: Stanford University, Department of Chemistry, Stanford, CA 94305, USA

+Present address: University of California-Berkeley, Naval Biosciences Laboratory, Oakland, CA 94625, USA

REFERENCES

1. Pohl, F.M. and Jovin, T.M., (1972) J. Mol. Biol. 67, 375-379.
2. Wang, A.H-J., Quigley, G.J., Kolpak, F.J., van der Marel, G., van Boom, J.H., and Rich, A., (1981) Science 211, 171-176.
3. Drew, H.R. and Dickerson, R.E., (1981) J. Mol. Biol. 151, 535.
4. Jovin, T.M., McIntosh, L.P., Arndt-Jovin, D.J., Zarling, D.A., Robert-Nicoud, M., van de Sande, J.H., Jorgensen, K.F. and Eckstein, F., (1983) J. Biomol. Struct. Dynam. 1,21-57.
5. Nordheim, A., Lafer, E.M., Peck, L.J., Wang, J.C., Stollar, B.D., and Rich, A., (1982) Cell 31, 309-318.
6. Singleton, C.K., Klysik, J., Stirdivant, S.M., and Wells, R.D., (1982) Nature 299, 312-316.
7. Pohl, F.M., Thomae, R., and DiCapua, E., (1982) Nature, 300, 545-546.
8. DiCapua, E., Stasiak, A., Koller, T., Brahms, S., Thomae, R., and Pohl, F.M., (1983) EMBO J. 2, 1531-1535.
9. Wang, J.C., Peck, L.J., and Becherer, K., (1983) Cold Spring Harbor Symp. Quant. Biol. 47, 85-92.
10. Peck, L.J. and Wang, J.C., Proc. Natl. Acad. Sci. USA (1983), 80, 6206-6210.
11. Zarling, D.A., Arndt-Jovin, D.J., McIntosh, L.P., Robert-Nicoud, M., and Jovin, T.M., (1984a) J. Biomol. Struct. Dynam. 1, 1081-1107.
12. Zarling, D.A., Arndt-Jovin, D.J., Robert-Nicoud, M., McIntosh, L.P., Thomae, R., and Jovin, T.M., (1984b) J. Mol. Biol., 176, 369-415.
13. Nordheim, A., Pardue, M.L., Lafer, E.M., Moller, A., Stollar, B.D., and Rich, A., (1981) Nature, 294, 417-422.
14. Lemeunier, F., Derbin, C., Malfoy, B., Leng, M., and Taillandier, E. (1982) Exp. Cell Res. 141, 508-513.
15. Arndt-Jovin, D.J., Robert-Nicoud, M., Zarling, D.A., Greider, C., Weimer, E., and Jovin, T.M., (1983) Proc. Natl. Acad. Sci. USA 80, 4344-4348.
16. Robert-Nicoud, M., Arndt-Jovin, D.J., Zarling, D.A., Jovin, T.M., (1984) EMBO J. 3, 721-731.
17. Russel, W.C., Precious, B., Martin, S.R. and Bayley, P.M., (1983) EMBO J. 2, 1647-1653.
18. Nordheim, A. and Rich, A., (1983) Nature 303, 674-679.
19. Azorin, F., Nordheim, A., Rich, A., (1983) EMBO J. 2,649-655.
20. Quadrifoglio, F., Mannzini, G., Yathindra, N., and Crea, A., (1983) Nucleic Acids: The Vectors of Life. Pullman, B. and Jortner, J. (eds.) pp. 61-74. D.Reidel, Dordrecht, Holland.
21. Rich, A., Nordheim, A., and Azorin, F., (1983) J. Biomol. Struct. Dynam. 1, 1-19.
23. Revet, B., Zarling, D.A., Jovin, T.M. and Delain, E., (1984) The EMBO J. in press.
22. Rich, A., (1983) Cold Spring Harbor Symp. Quant. Biol. 47, 1-13.
24. Stockton, J.F., Miller, F.D., Jorgenson, K.F., Zarling, D.A., Morgan, A.R., Rattner, J.B. and van de Sande, J.H., (1983) EMBO J. 2, 2123-2128.
25. Miller, F.D., Jorgenson, K.F., Winkfein, R.J., van de Sande, J.H., Zarling, D.A., Stockton, J. and Rattner, J.B., (1983) J. Biomol. Struct. Dynam. 1, 611-620.
26. Miller, F.D., Winkfein, R.J., Rattner, J.B., and van de Sande, J.H., (1984) Bioscience Reports, in press.
27. Hagen, F.K., Zarling, D.A., and Jovin, T.M., (1985) EMBO J., in press.