Unusual frequencies of certain alternating purine-pyrimidine runs in natural DNA sequences: relation to Z-DNA

Edward N. Trifonov*, Andrzej K. Konopka and Thomas M. Jovin

*Department of Polymer Research, The Weizmann Institute of Science, Rehovot 76100, Israel, and Max-Planck-Institut für biophysikalische Chemie, Abteilung Molekulare Biologie, Postfach 2841, D-3400 Göttingen, FRG

Received 1 April 1985

Prokaryotic, eukaryotic and mitochondrial DNA sequences of total Length 300000 nucleotides have been analyzed to find out whether stretches of alternating purines and pyrimidines are unusual in terms of occurrence, composition and base sequence. Alternating runs longer than 5 nucleotides are significantly underrepresented in the natural sequences as compared to random ones. Octanucleotides are the most deficient, occurring at only 60% of the frequency expected in random sequences. An unexpectedly high proportion of these octamers consists of alternating tetramers with the repeat structure $(PuPyPuPy)_2$ or $(PyPuPyPu)_2$. DNA stretches containing such sequences can potentially form a S1 nuclease sensitive slippage (staggered loop) structure, which might serve as a locally unstacked intermediate in the B- to Z-DNA conformational transition.

Left-handed DNA *B-Z transition* *DNA loop* *S1 nuclease*

1. INTRODUCTION

Electron microscopy and biochemical techniques have been used to map anti-Z-DNA immunoglobulin binding sites in the natural genomes of phages PM2 [1], ϕX174 [2] and eukaryotic virus SV40 [3,4], as well as in the cloning vector pBR322 [5]. The results are consistent with the observations [3,6–8] that the alternation of purines and pyrimidines is one of the most important factors potentiating the B-Z transition in the handedness of DNA. This finding applies both to synthetic polynucleotides as well as to natural sequences under topological stress in vitro, i.e., in the form of negatively supercoiled closed circular molecules.

The minimum length of an alternating mixed sequence purine-pyrimidine tract which readily undergoes the B-Z transition has been estimated as 8 base pairs ([3,4]; Konopka et al., submitted). The base composition is also an important factor in determining the relative stabilities of the right- and left-handed conformations, as determined with defined polynucleotides in solution [6,9] and in natural DNA genomes ([1–6]; Konopka et al., submitted). Thus, alternating pyrine-pyrimidine stretches rich in A and T are less stable in the Z conformation compared to their G·C-rich counterparts.

Here we analyzed the occurrence of alternating repetitions of purines and pyrimidines in different DNA sequences taken from the EMBL Nucleotide Sequence Data Library. Deviations from a random distribution were observed which may indicate that certain DNA stretches with the potential for adopting the left-handed Z conformation are avoided or favored in natural DNA genomes.

2. METHODS AND RESULTS

To estimate the frequencies with which the alternating PuPy (and PyPu) stretches of different lengths (2–16 nucleotides) are encountered, the following sequences were computer-searched: genomes of bacteriophages λ and T7, plasmid

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/85/$3.30

pBR322, transposon Tn903, *E. coli* gene coding for alanyl-tRNA synthetase, genes *atp*B, *atp*E, *atp*F, *ace*A, *ace*E, *atp*H, *atp*A, *pol*A, *dna*G, *ndh*, *omp*A, *pap*A, *pap*B, *pap*C, *thr*A, operons *trp* and *rpo*BC; complete or partial genomes of the viruses SV40, BKV, polyoma, adenoviruses 2 and 7, cauliflower mosaic virus; *Drosophila* cluster of 3 cuticle genes and genes coding for tRNAs and vitellogenin; chicken ovomucoid and ovalbumin genes; human fetal γ-hemoglobin genes, interferon genes and *ras* oncogene; human, bovine and mouse mitochondrial genomes. All these sequences were entries in the EMBL Nucleotide Sequence Data Library, Release 3.1, and had total lengths (nucleotides) of: 138811 (prokaryotes), 108791 (eukaryotes) and 49202 (mitochondria). We considered isolated runs with the structure (Pu)PuPy...PuPy(Py), (Pu)PuPy...PyPu(Pu), (Py)PyPu...PyPu(Pu) and (Py)PyPu...PuPy(Py), discarding overlapping shorter runs which were part of longer ones. Alternating runs of lengths more than 16 nucleotides are not considered in this calculation, being not frequent enough for the statistical treatment. Some striking cases of long alternating runs are discussed below. A similar analysis was applied to random sequences with the same lengths and base compositions, in which case the expected frequencies can also be calculated from the a priori probabilities.

The results of this search are presented separately for prokaryotes, eukaryotes and mitochondria in fig.1A,B and C, respectively, as well as for the whole ensemble of the sequences (fig.1D). Displayed are histograms of the values for alternating runs of length 2–16 nucleotides related to the number in the corresponding random sequences (horizontal lines at ordinate 1).

It is evident from fig.1 that the alternating stretches of lengths longer than 5 bases are under-represented in all 3 groups of the natural sequences. This is seen clearly in the total distribution (fig.1D) in which the stretches 8 nucleotides long are the most deficient. In case of eukaryotes (fig.1B) the distribution shows a broad minimum at about 7–10 bases rather than a sharp drop at the length 8. A total of 347 such alternating tracks were found representing only 60.5% of the number expected from the analysis of the random sequences. The deficit amounts to 9.5 units of standard deviation (SD) (estimated as the square root of the expected

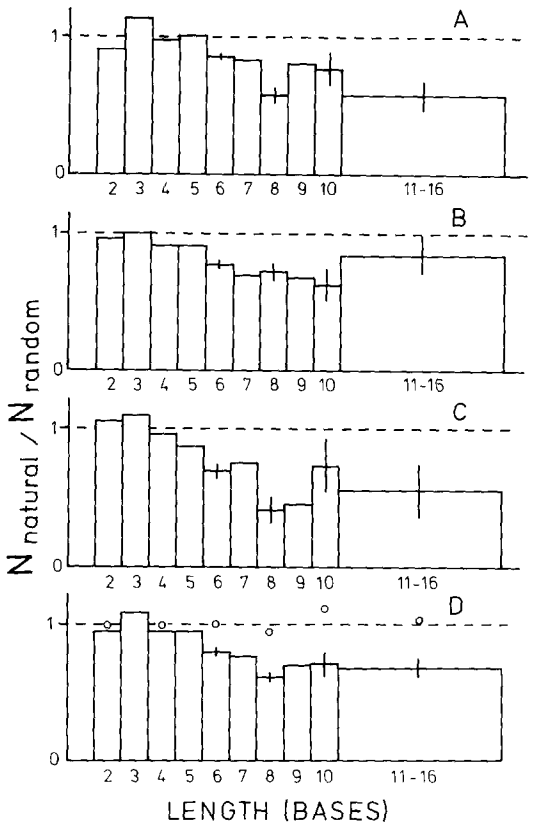

Fig.1. Relative occurrences of alternating PuPy and PyPu runs in (A) prokaryotic, (B) eukaryotic, and (C) mitochondrial nucleotide sequences of length 138811, 108791 and 49202 bases, respectively. (D) Combined distribution. Shown is the ratio of the total number $N_{natural}$ of the isolated alternating runs of specified length k found in the natural sequences to the number N_{random} of runs in the corresponding random sequences. $N_{random,k}$ is of the order $N \cdot 2^{-(k+1)}$, where N is the total length of the sequence analyzed. This value slightly depends on the Pu/Py ratio. The indicated error bars correspond to $\pm (N_{random})^{-1/2}$. Stretches of length 11–16 bases were combined. The results of analysis by direct examination of alternating stretches in random sequences are indicated in panel D (circles).

number). As a control, it was established that the predicted and observed frequencies of alternating stretches in the random sequences are identical within the error limits (e.g., fig.1D). The behavior of the distributions for longer runs is not well defined because of the limited size of the ensemble chosen for the analysis. The frequencies of the longer alternating stretches tend to approach the

levels corresponding to the random sequences. These levels should be exceeded at certain lengths since very long alternating runs (up to 76 nucleotides) have been found in natural sequences [10–14], occurrences of which are extremely improbable in random ensembles of any reasonable size. Long tracks of $(CA)_n$ *or* $(TG)_n$ (n up to 33) are frequently found in intergenic regions and intervening sequences of eukaryotic genes [10–13].

We nota an unusually high occurrence of alternating triplets (8% more than expected on a random basis; fig.1D). This circumstance is apparently a consequence of the 3 nucleotide code for proteins. The distribution of bases within the 'average' triplet is known to be quite nonuniform [15,16], and leads to the consensus sequence PuNPy [17,18]. This would results in the frequent occurrence of PyPuPu and PyPuPy runs linking neighboring codons in the protein coding sequences. The triplet PyPuPy is probably responsible for the unusually high frequency we have identified.

One obvious question is whether the octamers which survived the negative selection have any unusual features in terms of composition or nucleotide sequence. Indeed, the perfect repeat structures $(PuPyPuPy)_2$ and $(PyPuPyPu)_2$ including a degenerate case of repeating dimers are significantly over-represented, particularly, in the eukaryotic and mitochondrial sequences (see table 1). Of 1539 octamers found in all the sequences analyzed (in this case we considered both isolated and overlapping octamers, within runs up to 16 bases), 135 are repeated tetramers, a value which exceeds the expected number (96) by 4 SD units. All the repeated tetramers identified in the eukaryotic and mitochondrial sequences are listed in fig.2. Also shown (in parentheses) are the results of a search expanded to cover the entire ensemble of 1280 eukaryotic and mitochondrial sequences in the EMBL library (a total of ~970 kb of non-prokaryotic DNA), using the NAQ programs of the National Biomedical Research Foundation.

The listing in fig.2 indicates that some of the repeat motifs appear to be very frequent, e.g., $(ACAC)_2$, $(CACA)_2$, $(ACAT)_2$ (and permutations thereof), and their complementary counterparts; while others, like $(ACGT)_2$ and its permutations are not found at all. One striking example of naturally occurring tetramer repeats is a cluster located in the 5-non-coding region of the mouse β-major globin gene [14]:

$$(ACAT)_2...(CATA)_2...(TACA)_7...(ACAC)_2...(ACAC)_2...(ACAC)_2...(CACA)_3$$

Table 1

Occurrence of alternating octamers (isolated and overlapping[a]) and repeating tetramers

	Octamers	Repeating tetramers	Expected number of tetramers[b]
Prokaryotes	723	46	45.2
Eukaryotes	615	65	38.4
Mitochondria	201	24	12.6
Total	1539	135	96.2

[a] Overlapping is considered only within runs up to 16 bases

[b] Of 512 different alternating octamers 32 are repeats. Therefore, the expected fraction of repeating tetramers is 1/16. These include also degenerate cases of dimers repeating 4 times

3. DISCUSSION

The analysis of the natural nucleotide sequences described above results in two observations:

(i) Alternating PuPy and PyPu sequences 8 nucleotides long are deficient, but

(ii) The repeating tetramers $(PuPyPuPy)_2$ and $(PyPuPyPu)_2$ are over-represented.

In particular, the repeated tetramers $(ACAC)_2$ and $(GTGT)_2$ (and their derivatives) are very frequent, whereas the fragments $(ACGC)_2$ and $(ACGT)_2$ are not present at all in non-prokaryotic genomes. A corresponding analysis of prokaryotic DNA sequences (not shown) indicates that the fragments $(ACAC)_2$ and $(GTGT)_2$ occur with the same frequency as the fragment $(ACGC)_2$ avoided in eukaryotic genomes [although the repeated tetramer $(ACGT)_2$ is absent in prokaryotic as well as in eukaryotic genomes]. Furthermore, long

ACACACAC	5	(40)	GTGTGTGT	7	(74)
CACACACA	5	(42)	TGTGTGTG	9	(68)
ACATACAT	6	(30)	ATGTATGT	2	(32)
CATACATA	4	(16)	TATGTATG	2	(30)
ATACATAC	9	(17)	GTATGTAT	1	(25)
TACATACA	6	(25)	TGTATGTA	0	(23)
ATATATAT	5	(~250)			
TATATATA	6	(~250)			
GCGCGCGC	4	(32)			
CGCGCGCG	6	(32)			
ATGCATGC	4	(14)			
TGCATGCA	0	(17)			
GCATGCAT	1	(8)			
CATGCATG	0	(11)			
GCACGCAC	3	(9)	GTGCGTGC	2	(11)
CACGCACG	2	(4)	CGTGCGTG	0	(8)
ACGCACGC	0	(0)	GCGTGCGT	0	(5)
CGCACGCA	0	(1)	TGCGTGCG	0	(5)
ACGTACGT	0	(0)			
CGTACGTA	0	(1)			
GTACGTAC	0	(3)			
TACGTACG	0	(2)			

Fig.2. List of repeating alternating tetramers found in the eukaryotic and mitochondrial sequences. For convenience, the repeating tetramers are given in permutation groups, and such that the sequences on the left and right are complementary to each other. The amounts indicated correspond to only one strand of DNA and should be augmented by the values for the appropriate complementary sequence in order to account for the second strand, if desired. Figures in parentheses correspond to the entire ensemble of non-prokaryotic sequences in the EMBL library. In this case alternating runs longer than 16 bases are taken into account as well.

$(ATAT)_2$ tracts are frequent in eukaryotic DNAs, whereas in prokaryotes they occur at the expected frequency. Taken as a group, repeated tetramers in prokaryotic genomes are present at the statistically predicted frequencies (table 1). It is tempting to speculate that these facts are correlated with the existence and function of chromatin.

Solely on the basis of sequence analysis we obviously cannot determine whether the underrepresentation of alternating octamers and high proportion of repeated tetramers have anything to do with the formation of left-handed Z-DNA. However, it is of interest that the well characterized (in vitro!) Z-sites in the SV40 genome are both repeating tetramers $(ATGC)_2$ [3,4]. Might it be that the repeat structure of alternating PuPy and PyPu runs somehow favors the B-Z transition? One possible mechanism would involve the formation of a slippage structure (or staggered loop; fig.3) as an intermediate in the transition, especially since the latter must proceed via some unstacking of base-pairs in order to enable *anti*-to-*syn* rotation of the bases [19]. The slippage structure is known to be a typical element in the helix-coil transition in homopolymers and copolymers [20]. Due to the intermediates of the kind shown in fig.3, the double-stranded homopolymers differ significantly from mixed sequence molecules in their melting behavior [21]. The B-Z transition could also be favored by other locally unstacked structures such

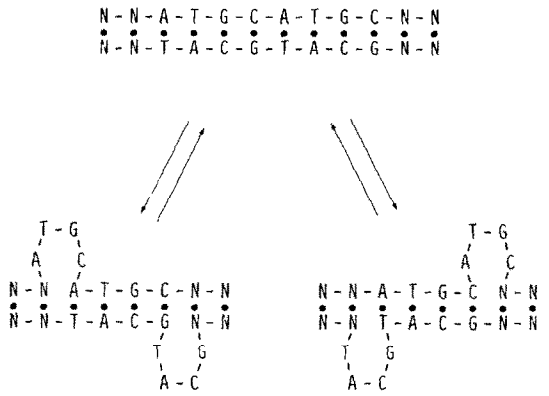

Fig.3. Staggered loop structures which can be formed in DNA with the PuPy alternating repeat $(ATGC)_2$.

as cruciforms [22,23] or local unwinding in regions of low stability.

In all the above cases the (transiently or permanently) unstacked bases constitute targets for the single-strand DNA specific nuclease S1 [24,25]. The Z-sites are also known to be S1 nuclease sensitive [26], a feature usually attributed to a special structure at the junctions between regions in the B and Z conformations [27]. However, the S1 nuclease digestion data are equally compatible with the occurrence of local unstacking prior to the B-to-Z transition as a first step, and/or the expression of an equilibrium between an established Z helical region with an intermediate locally unstacked structure. In this connection, the formation of a slippage structure such as the one shown in fig.3 is an attractive possibility. It is also of possible relevance that the sequences of most known S1 sensitive sites have a repeat structure, whether this is just homopolymeric stretch [28] or a more elaborate repeating sequence [29–34]. In many of these cases the slippage structure is the only alternative to a completely unwound state of the S1 sensitive region of the molecule, and may constitute the energetically most efficient way for (transiently) exposing single strands of DNA to the nuclease.

ACKNOWLEDGEMENT

E.N.T. is grateful to the European Molecular Biology Organization for a short-term fellowship.

REFERENCES

[1] Miller, F.D., Winkfein, R.J., Rattner, J.B. and Van de Sande, J.H. (1984) Biosci. Rep. 4, 885–895.
[2] Revet, B., Zarling, D.A., Jovin, T.M. and Delain, E. (1984) EMBO J. 3, 3353–3358.
[3] Nordheim, A. and Rich, A. (1983) Nature 303, 674–679.
[4] Hagen, F., Zarling, D.A. and Jovin, T.M. (1985) EMBO J., in press.
[5] DiCapua, E., Stasiak, A., Koller, T., Brahms, S., Thomae, R. and Pohl, F.M. (1983) EMBO J. 2, 1531–1535.
[6] 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. Dyn. 1, 21–57.
[7] 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.
[8] 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. Dyn. 1, 611–620.
[9] Quadrifoglio, F., Mannzini, G., Yathindra, N. and Crea, A. (1983) in: Nucleic Acids: The Vectors of Life (Pullman, B. and Jortner, J. eds) pp. 61–74, Reidel, Dordrecht.
[10] Nishioka, J. and Leder, P. (1980) J. Biol. Chem. 255, 3691–3694.
[11] Kim, S., Davis, M., Sinn, E., Patten, P. and Hood, L. (1981) Cell 27, 573–581.
[12] Hamada, H., Petrino, M.G. and Kakunaga, T. (1982) Proc. Natl. Acad. Sci. USA 79, 6455–6469.
[13] Hamada, H., Petrino, M.G., Kakunaga, T., Seidman, M. and Stollar, B.D. (1984) Mol. Cell. Biol. 4, 2610–2621.
[14] Gilmour, R.S., Spandidos, D.A., Vass, J.K., Gow, J.W. and Paul, J. (1984) EMBO J. 3, 1263–1272.
[15] Shulman, M.J., Steinberg, C.M. and Westmoreland, N. (1981) J. Theor. Biol. 88, 409–420.
[16] Trifonov, E.N. (1984) CODATA Bull. 56, 21–26.
[17] Eigen, M. (1978) Naturwissenschaften 65, 341–369.
[18] Shepherd, J.C.W. (1981) Proc. Natl. Acad. Sci. USA 78, 1596–1600.
[19] Wang, A.H.-J., Quigley, G.J., Kolpak, F.J., Crawford, J.L., Van Boom, J.H., Van der Marel, G and Rich, A. (1979) Nature 282, 680–686.
[20] Crothers, D.M. and Zimm, B.H. (1964) J. Mol. Biol. 9, 1–9.
[21] Lazurkin, Y.S., Frank-Kamenetskii, M.D. and Trifonov, E.N. (1970) Biopolymers 9, 1253–1307.
[22] Beerman, T.A. and Lebowitz, J. (1973) J. Mol. Biol. 79, 451–470.

[23] Lilley, D.M.J. (1980) Proc. Natl. Acad. Sci. USA 77, 6468–6472.
[24] Ando, T. (1966) Biochim. Biophys. Acta 144, 158–168.
[25] Drew, H.R. (1984) J. Mol. Biol. 176, 535–557.
[26] Singleton, C.K., Klysik, J., Stirdivant, S.M. and Wells, R.D. (1982) Nature 299, 312–316.
[27] Singleton, C.K., Kilpatrick, M.W. and Wells, R.D. (1984) J. Biol. Chem. 259, 1963–1967.
[28] Nickol, J.M. and Felsenfeld, G. (1983) Cell 35, 467–477.
[29] Henthschel, C.C. (1982) Nature 295, 714–716.
[30] Dybvig, K., Clark, C.D., Aliperti, G. and Schlesinger, M.J. (1983) Nucleic Acids Res. 11, 8495–8508.
[31] Bock, S.C. and Levitan, D.J. (1983) Nucleic Acids Res. 11, 8569–8582.
[32] Mace, H.A.F., Pelham, H.R.B. and Traver, A.A. (1983) Nature 304, 555–557.
[33] McKeon, C., Schmidt, A. and De Crombrugghe, B. (1984) J. Biol. Chem. 259, 6636–6640.
[34] Kilpatrick, M.W., Klysik, J., Singleton, C.K., Zarling, D.A., Jovin, T.M., Hanau, L.H., Erlanger, B.F. and Wells, R.D. (1984) J. Biol. Chem. 259, 7268–7274.