FEBS Letters 393 (1996) 124–130 FEBS 17444

Distribution of bending propensity in DNA sequences

Andrei Gabrielian**, András Simoncsits, Sándor Pongor*

International Centre for Genetic Engineering and Biotechnology (ICGEB), Area Science Park, Padriciano 99, 34012 Trieste, Italy

Received 26 June 1996; revised version received 12 July 1996

Abstract Local bending propensity and curvature of DNA can be characterized using a vector description of DNA bendability, based on a set of parameters derived from deoxyribonuclease I (DNase I) cleavage experiments. Two characteristics – arithmetic and vector averages of bendability – were successfully used to predict experimentally known bendable, rigid and curved segments in DNA. A characteristic distribution of bendability is conserved in evolutionarily related kinetoplast sequences. An analysis of the *M. genitalium* and *H. influenzae* genomes as well as fragments of human and yeast genomes shows, on the other hand, that highly curved segments – similar to artificially designed curved oligonucleotides – are extremely rare in natural DNA.

Key words: Bendability; DNase I cleavage

1. Introduction

The ability of DNA to bend is thought to play important roles in processes such as gene regulation, packaging and DNA replication. Recently, trinucleotide bending propensity parameters were deduced from deoxyribonuclease I (DNase I) digestion data [1,2]. DNase I, an enzyme with no pronounced sequence specificity, bends DNA towards the major groove [3,4]. Since the cleavage rate is thought to primarily depend on the bendability of DNA in this direction [5], we consider these trinucleotide parameters as indicators of bendability, i.e. flexibility in the direction of the major groove. This is an approximation since these DNase I-derived parameters do not in principle discriminate between bendability and inherent bending towards the major groove [1,2].

Here we attempt to classify DNA sequence segments into groups based on their bendability properties predicted from the sequence. We show that a simple vectorial representation can be used to compute numeric indices that correlate well with DNA curvature, and a 2D plot of these allows one to identify stiff, bendable and curved segments in long DNA sequences.

2. Materials and methods

The DNA sequences were taken from the EMBL and the GenBank nucleotide sequence databanks and are indicated in the text by GenBank locus names. The bendability parameters were determined by DNase I digestion experiments [1]. For the purposes of the present calculations the values were recalculated to a relative scale between 0 and 10, so that 0 corresponds to the most rigid and 10 to the most bendable segments [2]. Bendability plots were drawn by first dividing a DNA sequence into overlapping trinucleotides, then assigning a bendability value given in Table 1 to the center of each trinucleotide. (In this way the first and last nucleotides will have no values so a sequence of 32 residues will have 30 values.) Average bendability and helical asymmetry (see below at Eq. 1) were calculated for segments of 32 bp, i.e. approximately three helical turns. The calculated profiles do not significantly depend on this window length. Random sequences were generated by random shuffling the sequence of entire genomes. The shuffled sequences were then divided into overlapping segments of 32 residues for the calculation of the values given in Table 1 which are averages and standard deviations calculated from 10 runs of randomization.

3. Results and discussion

3.1. A vector representation of DNA bendability

Helical circle diagrams offer a simple visual way to represent the angular distribution of bendability within DNA segments. In this representation, the bendability parameters are drawn as vectors pointing towards the major groove, and plotted along idealized B-DNA (Fig. 1). Sequence motifs that are curved according to experimental data show a strongly asymmetric vector distribution. This is shown by two examples in Fig. 1, a kinetoplast minicircle sequence

Table 1
DNase I-based relative bendability parameters [2]

Trinucleotide	Relative bendability (a.u.)
AAA/TTT	0.1
AAC/GTT	1.6
AAG/CTT	4.2
AAT/ATT	0.0
ACA/TGT	5.8
ACC/GGT	5.2
ACG/CGT	5.2
ACT/AGT	2.0
AGA/TCT	6.5
AGC/GCT	6.3
AGG/CCT	4.7
ATA/TAT	9.7
ATC/GAT	3.6
ATG/CAT	8.7
CAA/TTG	6.2
CAC/GTG	6.8
CAG/CTG	9.6
CCA/TGG	0.7
CCC/GGG	5.7
CCG/CGG	3.0
CGA/TCG	5.8
CGC/GCG	4.3
CTA/TAG	7.8
CTC/GAG	6.6
GAA/TTC	5.1
GAC/GTC	5.6
GCA/TGC	7.5
GCC/GGC	8.2
GGA/TCC	6.2
GTA/TAC	6.4
TAA/TTA	7.3
TCA/TGA	10.0

---

*Corresponding author.
E-mail: pongor@icgeb.trieste.it

***Permanent address:* Institute of Molecular Biology, Russian Academy of Sciences, Vavilov St. 32, 117984 Moscow, Russia.

PII S0014-5793(96)00837-X

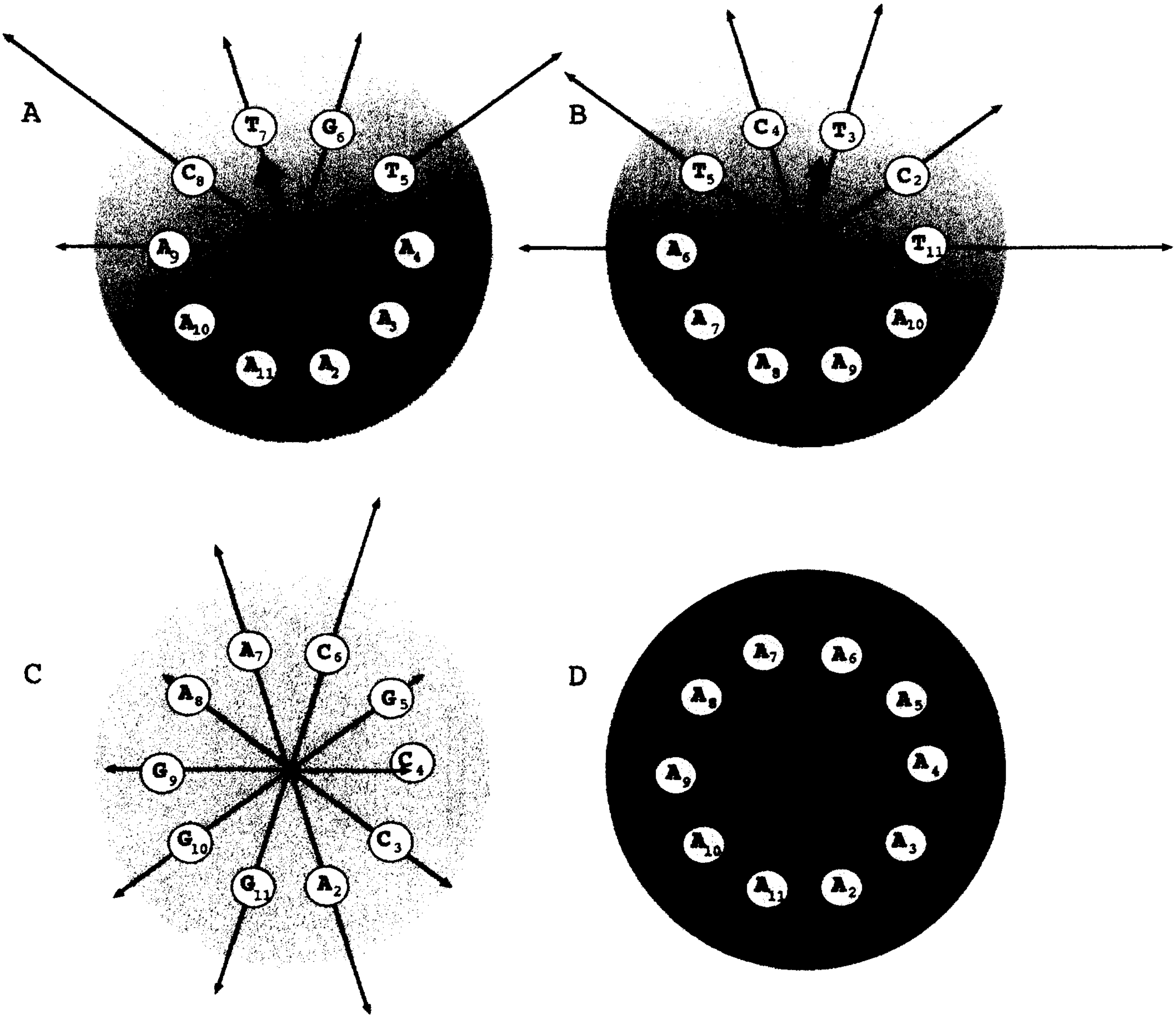

Fig. 1. Vectorial representation of bendability (helical circle diagrams). A: A curved sequence motif (A)AAATGTCAAA(A) from a *Leishmania tarentolae* class II minicircle [29]. B: A curved sequence motif (A)CTCTAAAAAT(A) designed by Ulanovsky et al. [6]. C: A straight sequence from the lambda phage OR3 operator region [8,36]. (C)ACCGCAAGGG(A). D: Poly-A sequence. The length of black arrows is proportional with that of the bendability parameter at the given sequence position. The red arrow is the vectorial average of the bendability vectors (given in Eq. 1. The length of the average vector is negligible in C and D, so it is denoted by a dot only). The radius of the shaded circle indicates the average bendability of genomic sequences (about 5.3). Stiff and flexible parts of the helix are indicated by dark and light shading, respectively.

from *Leishmania tarentolae* (Fig. 1A) and an artificially designed highly curved sequence motif of Ulanovsky and coworkers [6] (Fig. 1B).

On the other hand, the majority of DNA sequences do not provide asymmetric bendability distributions, i.e. in the general case bendable and stiff trinucleotides are placed seemingly randomly on any side of the helical circle. In principle, symmetric distributions can be produced in many ways. Two of these, corresponding to the extremes of the bendability range, are of particular interest (Fig. 1C,D). (i) If the bendability is uniformly high along the sequence, the segment is expected to be bendable, i.e. flexible toward the major groove. Curvature is not expected since the vectors cancel each other. An interesting example of this case is the lambda phage OR3 operator (Fig. 1C). One notices that the bendability distribution is quite symmetric. In fact, this region is straight in solution [7] but its bendability is known from experiment, since it was shown to strongly bend when in contact with the Cro protein [8]. (ii) If the bendability is uniformly low along the sequence, the segment is expected to be rigid. An extreme example is the poly-(dA) sequence (Fig. 1D). Poly(dA) is in fact known to be rigid [9], and, as a homopolymer, it is devoid of static, macroscopic curvature.

3.2. Asymmetric bendability reflects propensity to curvature

Based on the observation that average bendability and the asymmetry of the vector distribution can be important in characterizing individual DNA segments, we introduce two

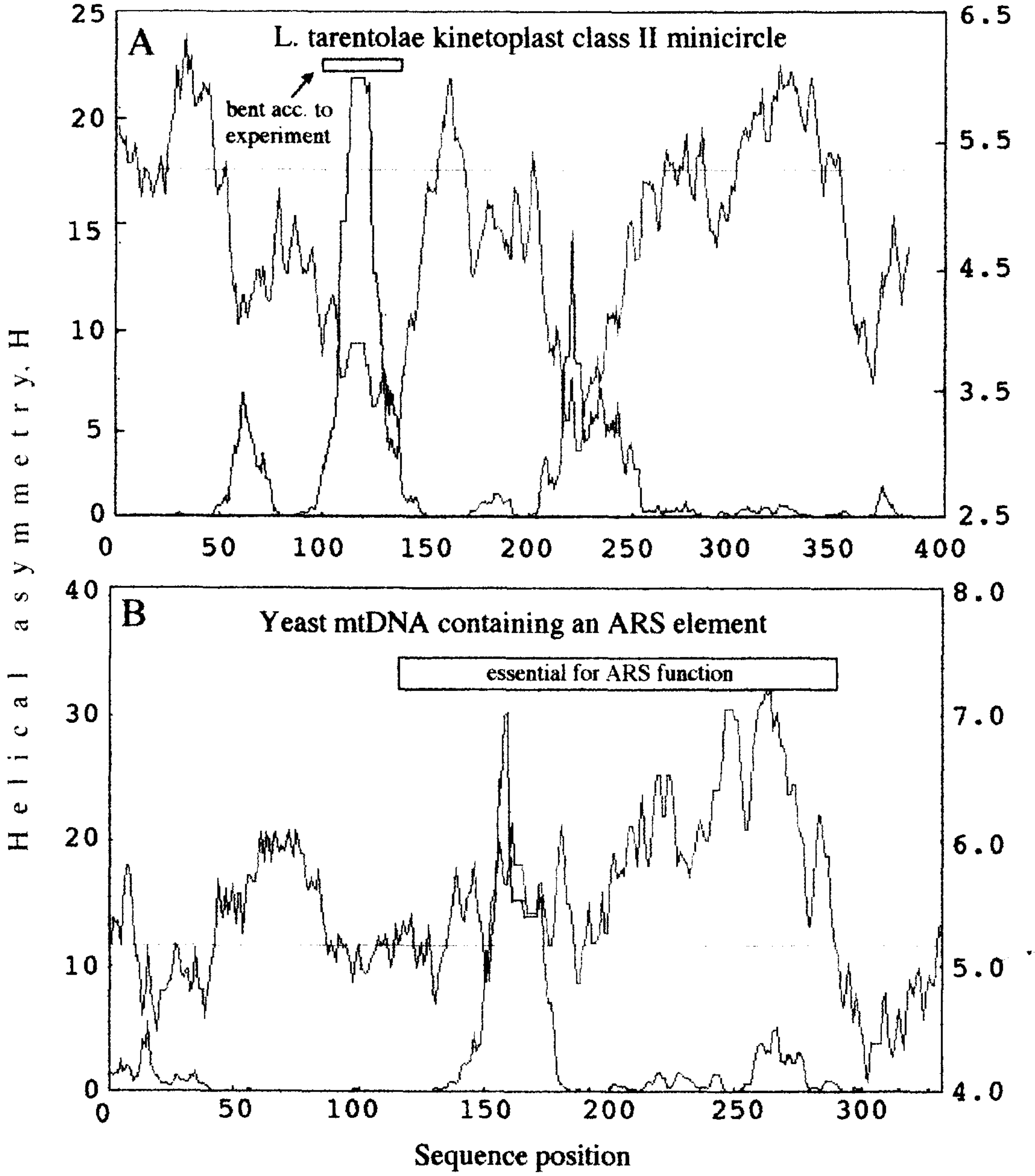

Fig. 2. Helical asymmetry (red) and bendability (blue) versus sequence plots. A: *Leishmania tarentolae* class II minicircle (Genbank LEIKPMNC2). B: *Sacharomyces cerevisiae* mitochondrial autonomous replicating sequence element (ARS, Genbank MISCARS) The blue dotted line indicates the average bendability value of DNA. The average helical asymmetry value (which is close to 0) is not indicated.

numeric indices. H, the helical asymmetry index, proportional to the vectorial average of the bendability parameters as shown in the helical wheel diagrams, is designed to characterize curvature. The vector sum (red arrow in Fig. 1) is calculated with the approximation that the bendability vectors lie in the same plane and that the DNA segment is an ideally straight B-DNA:

$$L = \frac{1}{n}\left[\left(\sum_{i=1}^{n} f_i\cos(i\omega)\right)^2 + \left(\sum_{i=1}^{n} f_i\sin(i\omega)\right)^2\right]^{1/2} \quad (1)$$

where f_i is the bendability parameter (taken from Table 1) for position i, ω is the twist angle (36° for ideal B-DNA) and n is the number of vectors in the segment. (In this study a segment length of 32 residues, i.e. approximately three helical turns, was used.) We define the helical asymmetry index H as L^4 (the power was chosen only to improve the presentation of the data). As a first point of reference, we calculated average H and bendability values for various complete genomes (Table 2). The average bendability values are all around 5, i.e. in the middle of the bendability range. The average of helical asymmetry index, on the other hand, is not substantially different from 0 (Table 2). Random shuffling of the sequences does not appreciably influence the average values. We found that the vast majority of segments in genomic DNA have flexibility values close to the average genome values. For example, in the sequence of yeast chromosome III only a minor propor-

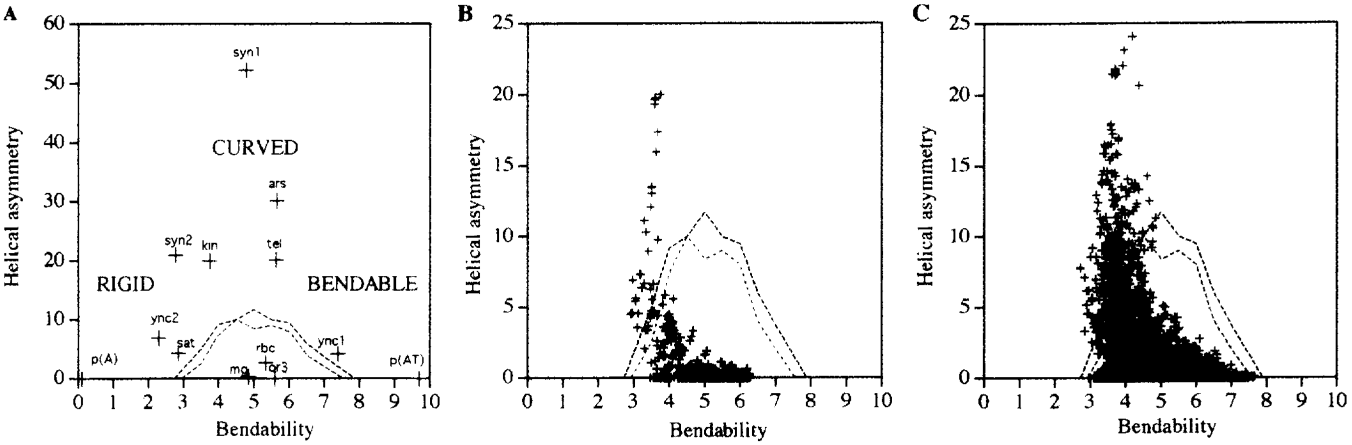

Fig. 3. Helical asymmetry index vs. bendability plots in selected sequences. A: Curved and straight sequences from Table 1. B: *Leishmania tarentolae* class II minicircle. C: 10 different kinetoplast sequences from various flagellatae, not excluding the example in B (Genbank locus name, species, length: CFU12625_S2, *C. fasciculata*, 1371 bp; KILTB4MC_S2. *L. tarentolae*, 887 bp; LDKMPL13_S7, *L.* sp., 821 bp; LGKINET11_S2, *L. guliki*, 860 bp; LIKINET10_S3, *L. infantum*, 793 bp; LMKINET4_S3, *L. major*, 692 bp; LTKINET3_S2, *L. tropica*, 854 bp; PBKINET6_S2, *P. brevicola*, 1477 bp; TBU03908_S2, *Trypanosoma brucei*, 1014 bp; TRBKINC5BA_S2, *Trypanosoma rangeli*, 1764 bp).

Table 2
Average bendability and helical asymmetry (curvature) values for selected genomic sequences[a]

No.	Sequence	Bendability[b]			H (helical asymmetry index)[b]	
		average (S.D.)	Min.	Max.	average (S.D.)	Max.
1	*H. influenzae* complete genome (1860 kb, 86% coding) [18]	5.02 (0.59)	1.82	7.33	0.35 (0.83)	38.1
2	Randomized[c] *H. influenzae* complete genome	5.34 (0.58)	2.09	7.84	0.25 (0.61)	20.3
3	*M. genitalium* complete genome (590 kb, 88% coding) [24]	4.85 (0.64)	1.66	7.56	0.35 (0.81)	28.5
4	Randomized[c] *M. genitalium* complete genome	5.25 (0.61)	2.43	7.76	0.34 (0.80)	10.3
5	*S. cerevisiae* chromosome III (SCCHRIII, 315 kb, 35% coding) [25]	5.21 (0.65)	0.78	9.47	0.29 (0.74)	41.6
6	Randomized[c] sequence of yeast chromosome III	5.20 (0.67)	2.37	7.58	0.32 (0.79)	21.7
7	Adenovirus type II complete genome (ADRCG,36 kb, 53% coding)	5.46 (0.54)	2.30	7.36	0.14 (0.38)	9.74
8	Randomized[c] adenovirus genome	5.56 (0.44)	3.85	7.18	0.09 (0.20)	4.95
9	Opossum mitochondrial genome (17 kb, 64% coding) [26]	5.33 (0.67)	1.70	7.73	0.42 (1.29)	22.5
10	Randomized[c] opossum mitochondrial genome	5.30 (0.58)	3.1	7.34	0.34 (0.70)	10.1
11	Human growth hormone gene (HUMGH)	5.66 (0.44)	4.49	6.97	0.13 (0.24)	2.14
	Average[d]	5.26 (0.59)			0.29 (0.73)	

[a] Helical asymmetry and average bendability values calculated for a window of 32 residues.

[b] Data given as average, standard deviation, lowest (Min.) and highest (Max.) values. The minimal value of the helical asymmetry is not given, since this quantity is close to 0 throughout most of the DNA sequences. Overlapping coding segments were considered only once in the calculations.

[c] Calculated from sequences 1, 3, 5, 7, 9, 11 (randomized sequences were not included).

[d] Average and standard deviation were calculated from 10 runs of randomization.

Table 3
Helical asymmetry (curvature) and bendability values for selected sequence segments[a]

No. and abbreviation	Origin (reference)	Sequence	Bendability index[b] (F)	Helical asymmetry index[b] (H)
	Curved DNA			
1 syn1	Synthetic [6]	(tctctaaaaaatatataaaaa)$_n$	4.79	52.3
2 syn2	Synthetic [10]	(aaaattttgc)$_n$	2.78	21.0
3 syn3	Synthetic [10]	(aaaattttcg)$_n$	2.22	8.18
4 syn4	Synthetic [27]	tctcaaaaaacgcgaaaaaaccggaaaaaagc	3.14	7.94
5 syn5	Synthetic [28]	(ccgaaaaagg)$_n$	3.86	14.0
6 syn6	Synthetic [28]	(ggcaaaaaac)$_n$	3.20	14.6
7 kin	*L. tarentolae* kinetoplast [29]	ccaaaaatgtcaaaaaataggcaaaaaatgcc	3.76	20.0
8 ars	Yeast mtDNA ARS element [30]	aaaatatatataatttataattttcatataat	5.66	30.2
9 rbc	Pea *rbcS* gene, regulatory region [19]	tggctgcaaactttatcattttcactatctaac	5.34	2.72
10 sat	*C. risortia* bent satellite DNA (Genbank CRBENSAT)	agaattgggacaaaaattggaaattttaaggg	2.86	4.36
11 tel	*T. thermophila* mitochondrium telomeric repeat [31]	cttagaggtatgttagctattagtgttgttta	5.64	20.2
	Straight DNA			
12 syn6	Synthetic [32]	(tccccgggga)$_n$	4.84	0.0009
13 syn7	Synthetic [33]	(atctaatctaacacaacaca)$_n$	5.14	0.0008
14 syn8	Synthetic [10]	(ttttaaaacg)$_n$	2.86	0.0025
15 syn9	Synthetic [10]	(ttttaaaagc)$_n$	3.60	0.0072
16 or3	OR3 operator region [37]	actacgttaaatctatcaccgcaagggataaa	5.60	0.0100
17 or3′	OR3 region, mutated [8]	actacgttaaatctatcacc$\underline{a}$caagggataaa	5.63	0.0070
18 p(A)	poly-A [20,21]	(a)$_n$	0.10	0
	Selected other sequences[a]			
19 tcr	Human T-cell receptor locus, putative microsatellite domain	(at)$_n$	9.70	0
20 ync1	Yeast chromosome III, segment in non-coding region	aatatatataaatatataaagcatcatatgat	7.40	4.22
21 ync2	Yeast chromosome III, segment in non-coding region	agttaaaaattttcaatttttttttcacttttt	2.30	6.94

[a] Helical asymmetry and average bendability values calculated for a window of 32 residues. In cases where H varied with the sequence position, (H,F) pairs corresponding to the position with the maximal H value are given.

[b] Data are given as average and standard deviation.

tion of segments have bendabilities exceeding the average plus one standard deviation (17%), and a negligible number of segments exceed the thresholds of average plus two or three standard deviations (1.7% and 0.24%, respectively). Similar percentage values (18.6%, 1.8% and 0.06%, respectively) were obtained for the *H. influenzae* complete genome.

Motifs that are known to be curved by experiment show a quite different picture (nos. 1–11 in Table 3). The H helical asymmetry index values are high, typically greater than 2. In contrast, the H index is quite low for straight DNA, typically below 0.1 (nos. 12–18 in Table 3). H in fact seems to be a quite sensitive indicator of curvature. All curved motifs tested by us so far give helical asymmetry values much higher than the genomic averages, and motifs known to be straight give values close to 0. As an example, the highly curved (AAAATTTTNN)$_n$ motifs [10] have H values above 8 while their mirror image (TTTTAAAANN)$_n$ motifs, which are not curved [10], have values below 0.01. We note that many of the curved motifs have relatively low bendability values. This is in keeping with the notion that curved conformations need to be rigid, in order to be sufficiently populated and so detectable by physical methods [11]. Bendability variations along the DNA sequence can be conveniently followed by plotting H and F as a function of the sequence position (Fig. 2). For example, the plots of H and F for kinetoplast minicircle from *L. tarentolae* and yeast mitochondrial autonomously replicating sequence (ARS) are quite different in this respect, even though both coincide with experimentally tested regions of curvature. In the kinetoplast sequence the dominant H peak corresponds to a deep minimum of bendability (Fig. 2A). In contrast, there is no apparent drop of bendability in the highly curved region of ARS sequence (Fig. 2B).

In Table 3, we present additional examples that have extreme values of helical asymmetry. Most of these sequences were investigated experimentally for the presence of curvature. (AT)$_n$ type sequences such as those located in the human T-cell receptor locus microsatellite region possess the highest bendability value found among all of the analyzed sequences. The bendability of (AT)$_n$ repeats is in fact well known from experiment [12]. Also, protein-induced bending at TATA sequences seems to be especially pronounced [13–17]. The lowest bendability value, on the other hand, is associated with poly(dA) motifs, which is also in agreement with experimental data, obtained by nucleosome studies [9,17] and X-ray crystallography [20,21]. It is interesting to note that the motifs of highest helical asymmetry index are in fact the properly positioned combinations of these two type of segments (i.e. $(AT)_n$ and poly(dA)). An example is the central 'aaaaatatata' motif in *synl* (Table 2 [6,22]). Comparing the extreme values in Tables 2 and 3 one can see that the high helical asymmetry values characteristic of strong intrinsic curvature are conspicuously missing from the genomes analyzed so far.

3.3. Patterns of bendability in DNA sequences

The asymmetry versus bendability (H vs. F) distributions in long sequences can be conveniently analyzed by dividing the sequence into overlapping segments and plotting its asymmetry value H against the average bendability F. On this plot, every DNA segment is represented as a point with co-ordinates corresponding to its calculated bendability and helical asymmetry, which makes it easy to locate segments of extreme characteristics. The properties of this 2D plot are illustrated on examples taken from Table 3. Fig. 3A shows that the segments corresponding to curved DNA (like kinetoplast sequences, yeast ARSs) fall far outside the region corresponding to 'average' DNA segments or random sequences. It is plausible to suppose that the flexible and the 'average' molecules have little pronounced conformation in solution, while the rigid and the curved ones are likely to have a determined shape.

If patterns of bendability describe conformational signals of functional significance, they are expected to be associated only with certain DNA regions. Furthermore, they have to be conserved in functionally analogous and evolutionarily related sequences but rare or absent in other genomic regions or in random-shuffled sequences. Here we attempt to show that bendability distributions are characteristic and conserved within a group of related DNA sequences, the flagellate kinetoplasts. Two-dimensional plot of kinetoplast minicircle of *L. tarentolae* shows a characteristic distribution with several curved segments (Fig. 3B). When the sequence is reshuffled, the random sequence does not contain pronounced asymmetric 'outliers' any more. Fig. 3C shows the bendability distribution of 10 other kinetoplast sequences. The conservation of the distribution is quite apparent. So the asymmetry versus bendability distributions meet the basic criteria of conservation and uniqueness expected from sequence patterns.

Summarizing, we can conclude that the distribution of bendability in curved DNA is highly asymmetrical, and this makes it possible to discriminate bendable, curved and rigid DNA segments using simple graphic representations. The mathematical tools used for this purpose are analogous to those originally developed for protein α-helices, such as the helical wheel diagrams of Schiffer and Edmundson [33] and the hydrophobic moment of Eisenberg and co-workers [34,35]. Applying this formalism to the bendability distribution in DNA we found that intrinsically curved sites seem to have an asymmetric, helically phased bendability pattern. On the other hand, sites bendable by proteins (like the lambda phage operator region) may be characterized by a high overall bendability. It thus appears that the bendability of DNA can be a major factor underlying both intrinsic and induced curvature. A similar conclusion was reached recently by Young and co-workers [21], based on a comprehensive analysis of DNA crystal structures. The second finding is that the bendability distribution in some DNA sequences, such as the kinetoplast DNA, is characteristically non-random, and is conserved in evolutionarily related examples. The method applied is simple and can be used for the analysis of full genomes. Our preliminary data on complete genomes suggest that high curvature, found in artificially designed oligonucleotides, seems to be absent in natural DNA sequences.

References

[1] Brukner, I., Sanchez, R., Suck, D. and Pongor, S. (1995) EMBO J. 14, 1812–1818.

[2] Brukner, I., Sanchez, R., Suck, D. and Pongor, S. (1995) J. Biomol. Struct. Dynam. 13, 309–317.

[3] Lahm, A. and Suck, D. (1991) J. Mol. Biol. 222, 645–667.

[4] Weston, S.A., Lahm, A. and Suck, D. (1992) J. Mol. Biol. 226, 1237–1256.

[5] Hogan, M.E., Roberson, M.W. and Austin, R.H. (1989) Proc. Natl. Acad. Sci. USA 86, 9273–9277.

[6] Ulanovsky, L., Bodner, M., Trifonov, E. and Choder, M. (1986) Proc. Natl. Acad. Sci. USA 83, 862–866.

[7] Baleja, J.D., Pon, R.T. and Sykes, B.D. (1990) Biochemistry 29, 4828–4239.

[8] Lyubchenko, Y., Shlyakhtenko, L.S., Appella, E. and Harrington, R.E. (1993) Biochemistry 32, 4121–4127.

[9] Rhodes, D. (1979) Nucleic Acids Res. 6, 1805–1816.

[10] Hagerman, P.J. (1986) Nature 321, 449–450.

[11] Travers, A.A. (1995) In DNA-Protein: Structural Interactions (D.M.J. Lilley, Ed.), pp. 49–75. IRL Press, Oxford.

[12] Calladine, C.R. and Drew, H.R. (1992) Understanding DNA – The Molecule and How It Works. Academic Press, San Diego, CA.

[13] Kim, J.L. and Burley, S.K. (1994) Nature Struct. Biol. 1, 638–653.

[14] Kim, J.L., Nikolov, D.B. and Burley, S.K. (1993) Nature 365, 512–520.

[15] Shakked, Z., Guzikevich-Guerstein, G., Frolow, F., Rabinovich, D., Joachimiak, A. and Sigler, P.B. (1994) Nature 368, 469–473.

[16] Starr, D.B., Hoopes, B.C. and Hawley, D.K. (1995) J. Mol. Biol. 250, 434–446.

[17] Simpson, R.T. and Kunzler, P. (1979) Nucleic Acids Res. 33, 1387–1415.

[18] Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M. et al. (1995) Science 269, 496–512.

[19] Cacchione, S., Cerone, M.A., De Santis, P. and Savino, M. (1995) Biophys. Chem. 53, 267–281.

[20] Dickerson, R.E., Goodsell, D.S. and Neidle, S. (1994) Proc. Natl. Acad. Sci. USA 91, 3579–3583.

[21] Young, M.A., Ravishanker, G., Beveridge, D.L. and Berman, H.M. (1995) Biophys. J. 68, 2454–2468.

[22] Bolshoy, A., McNamara, P., Harrington, R.E. and Trifonov, E.N. (1991) Proc. Natl. Acad. Sci. USA 88, 2312–2316.

[23] Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M. et al. (1995) Science 270, 397–403.

[24] Oliver, S.G. et al. (1992) Nature 357, 38–46.

[25] Janke, A., Feldmaier-Fuchs, G., Thomas, W.K., von Haeseler, A. and Paabo, S. (1994) Genetics 137, 243–256.

[26] Calladine, C.R., Drew, H.R. and McCall, M.J. (1988) J. Mol. Biol. 201, 127–137.

[27] Koo, H.S., Wu, H.M. and Crothers, D. (1986) Nature 320, 501–506.

[28] Marini, J.C., Levene, S.D., Crothers, D.M. and Englund, P.T. (1983) Proc. Natl. Acad. Sci. USA 80, 7678–7678.

[29] Mabuchi, T. and Wakabayashi, K. (1984) J. Biochem. 95, 589–592.

[30] Morin, G.B. and Cech, T.R. (1986) Cell 46, 873–883.

[31] Cacchione, S., DeSantis, P., Foti, D., Palleschi, A. and Savino, M. (1989) Biochemistry 28, 8706–8713.

[32] Bednar, J., Furrer, P., Katritch, V., Stasiak, A., Dubochet, J. and Stasiak, A. (1995) J. Mol. Biol. 254, 579–594.

[33] Schiffer, M. and Edmundson, A.B. (1967) Biophys. J. 7, 121–135.
[34] Eisenberg, D., Weiss, R.M. and Terwilliger, T.C. (1982) Nature 299, 371–374.
[35] Eisenberg, D., Schwarz E., Komaromy, M. and Wall, R. (1984) J. Mol. Biol. 179, 125–142.
[36] Lyubchenko, Y., Shlyakhtenko, L., Chernov, B. and Harrington, R.E. (1991) Proc. Natl. Acad. Sci. USA 88, 5331–5334.