Prediction of Bendability and Curvature in Genomic DNA

ANDREI GABRIELIAN
KRISTIAN VLAHOVIČEK*
MIRCEA G. MUNTEANU
M. MICHAEL GROMIHA
IVAN BRUKNER
ROBERTO SANCHEZ
SÁNDOR PONGOR

*Permanent address: Department of General and Inorganic Chemistry
Faculty of Science
University of Zagreb
Kralja Zvonimira 8
10000 Zagreb
Croatia

Correspondence:
Sándor Pongor
International Centre for Genetic Engineering and Biotechnology (ICGEB)
Area Science Park
Padriciano 99
34012 Trieste
Italy

Introduction

The behaviour of DNA chain can be studied with a wide variety of models of varying complexity ranging from full scale atomic models to simple elastic rod models *(28, 32, 34, 41, 51)*. As very large amounts of genomic sequence data are being produced, there is a growing need for simple methods that can help experimenters to find regions that are conspicuous in terms of flexibility or intrinsic curvature. Typically, proteins induce bending in short segments of up to 50 basepairs, and regions of static curvature are also quite short *(36, 46)*. Over the past years we have been interested in developing simple mechanic models that can describe the local behaviour of DNA in such short segments, in a sequence-dependent fashion *(8, 10-12, 16-18, 21, 22)*. This goal is thus slightly different from that of the modelling studies undertaken to investigate the static or dynamic behaviour of long chains, such as plasmids *(28, 34, 41)*. The approach thus targets an intermediate range of DNA segment length, between full atomic scale studies of short DNA segments and dynamic modelling of long DNA chains. We are particularly interested in how the bendability of DNA can explain the anisotropic bending and behaviour in such short segments.

Description of the *Sequence-Dependent Anisotropic DNA Bendability* Model

As a first approximation, let's consider a B-DNA segment as a cylindrical, segmented, elastic rod, each segment of length l corresponding to, for example, one basepair or dinucleotide step. If we consider each segment of the rod to have a different (for the moment isotropic) Young's modulus E(i), the average rigidity of a rod of n segments can be calculated as

$$\frac{1}{\langle E \rangle} = \frac{1}{N} \sum_i \frac{1}{E_i} \qquad [1]$$

The bending energy of this elastic rod of length n subjected to pure bending can be given as

$$\Delta G = \frac{1}{2} Bnl\kappa^2 \qquad [2]$$

where κ is the curvature and B is the bending rigidity given as

$$B = EI \qquad [3]$$

Here I is the moment of inertia which for a rod of circular cross section is given as $I = \pi r^4 / 4$. r is the radius of the cylinder which is taken as 10 Å.

The shape of such a rod can be accurately described in terms of Rise, Twist, Tilt and Roll parameters. All the parameters can have a static component, plus a dynamic one, describing the fluctuation around the static average value. In the general case, all these parameters will have different values for all the segments. The form of the dynamic component is also crucial. In the simplest case we can consider one spring constant per parameter, which corresponds to a harmonic oscillator with a symmetrical, parabola-like energy distribution. This would amount to saying that DNA has the same resistance to compression and stretching, or to bending towards the major and the minor grooves, respectively. In order to allow anisotropy, one has to use a more complex description, and the simplest of these is to allow different spring constants acting in both directions along each parameter, respectively. This is a non-linear elastic model simplified to a "bilinear" description, in which displacement in the direction of each parameter is linearly dependent on the force applied. This of practical importance with methods like finite element modelling *(2, 3, 52, 55)*. In this general, anisotropic model each segment of the rod will have 12 parameters (a static average plus two spring constants for each Rise, Twist, Roll and Tilt). For prediction we need reliable values for all these parameters and the currently available structural data are not sufficient for this purpose. Also, there are various theoretical possibilities to introduce sequence-dependence into the model and one can show that the current geometric models are special sub-cases of this general model. For example, static dinucleotide models assign static torsional angles to each of the 10 ds dinucleotides but neglect the dynamic parameters (spring constants) and use a uniform Rise value.

In order to incorporate sequence-dependent bendability information we use an approach of simplification that is different from that of the static models. Namely, we consider B-DNA as a more-or-less straight rod that has different flexibility parameters in each of the segment, but use uniform, ideal B-DNA values for all static parameters. And since we are primarily interested in bending, as opposed to torsion (supercoiling) or stretching), we allow only bending, but in an anisotropic way. This results in a simplified segmented rod model, shown in Figure 1. In this picture, each disk corresponds to one basepair and the arrow indicates the direction of facilitated flexibility (i.e. that of the major groove). In principle, such an anisotropic bendability model will have 4 parameters for each segment (two spring constants for each Roll and Tilt). In the present form of the *sequence-dependent DNA bendability* (SDAB) model one further simplification is applied: we take bendability towards the major groove as the principal parameter, and consider DNA 10-times stiffer in all other directions *(21)* In this way, we have one bendability parameter (Young's modulus) per segment, plus one general parameter, the anisotropy ratio (taken as 10 in this case) which is the ratio of the Young's moduli (i.e. ratio of the flexibility along the half axes arrows as shown in Figure 1). Concepts of anisotropy are extensively reviewed in *(36)*.

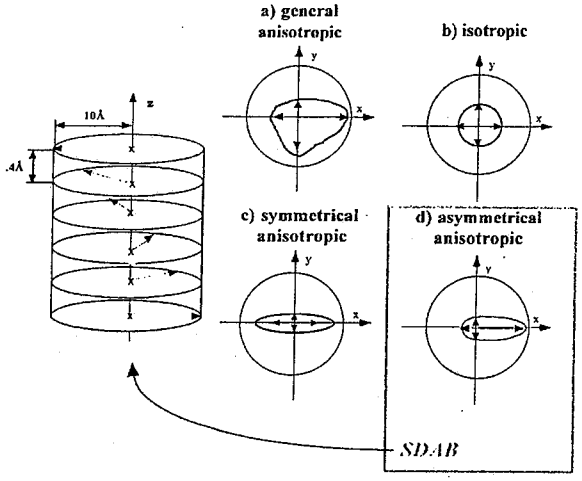


FIGURE 1. *The sequence-dependent anisotropic bendability model of DNA. Schematic outline of the sequence-dependent bending models. Each element corresponds to one basepair. The arrows in each basepair schematically indicate the direction pointing towards the major groove. The arrows are proportional to the flexibility (the inverse of stiffness) in a given direction (+x, -x, +y, -y). In the general case (a), all of these are different. In the isotropic case (b) all of them are equal. In the symmetrical anisotropic case (c) two of them pointing to the grooves are equal and the two other directions are stiffer. In the asymmetrical anisotropic case (d) which corresponds to the sequence-dependent anisotropic DNA bendability (SDAB) model described here, one direction (that of the major groove) is more flexible, the other three are more rigid. The proportion of the stiffness values is the anisotropy ratio, taken as 10.*

Parametrization of the *Sequence-Dependent Anisotropic DNA Bendability* Model

The rationale behind building bendability based models as opposed to static models is that bendability data can be deduced from protein/DNA binding studies and such data are more easily available than 3D structures. In particular, bovine pancreatic deoxyribonuclease I (DNAseI) can be considered to be a good molecular probe of DNA bendability since all DNAseI/DNA complexes solved so far show that productive binding of DNAseI requires DNA to be bent towards the major groove *(27, 49)*. Also, nucleosome positioning data can be used, since nucleosomes bend DNA in the same direction *(39)*. In principle, both of these can be used to deduce nearest-neighbor (dinucleotide-, trinucleotide, tetranucleotide) parameters for bendability. The first question is the complexity of the model, i.e. if one should use dinucleotide, trinucleotide or tetranucleotide parameters. This clearly depends on the number and quality of the experimental data, as we illustrate it here on the example of the DNAseI parameters.

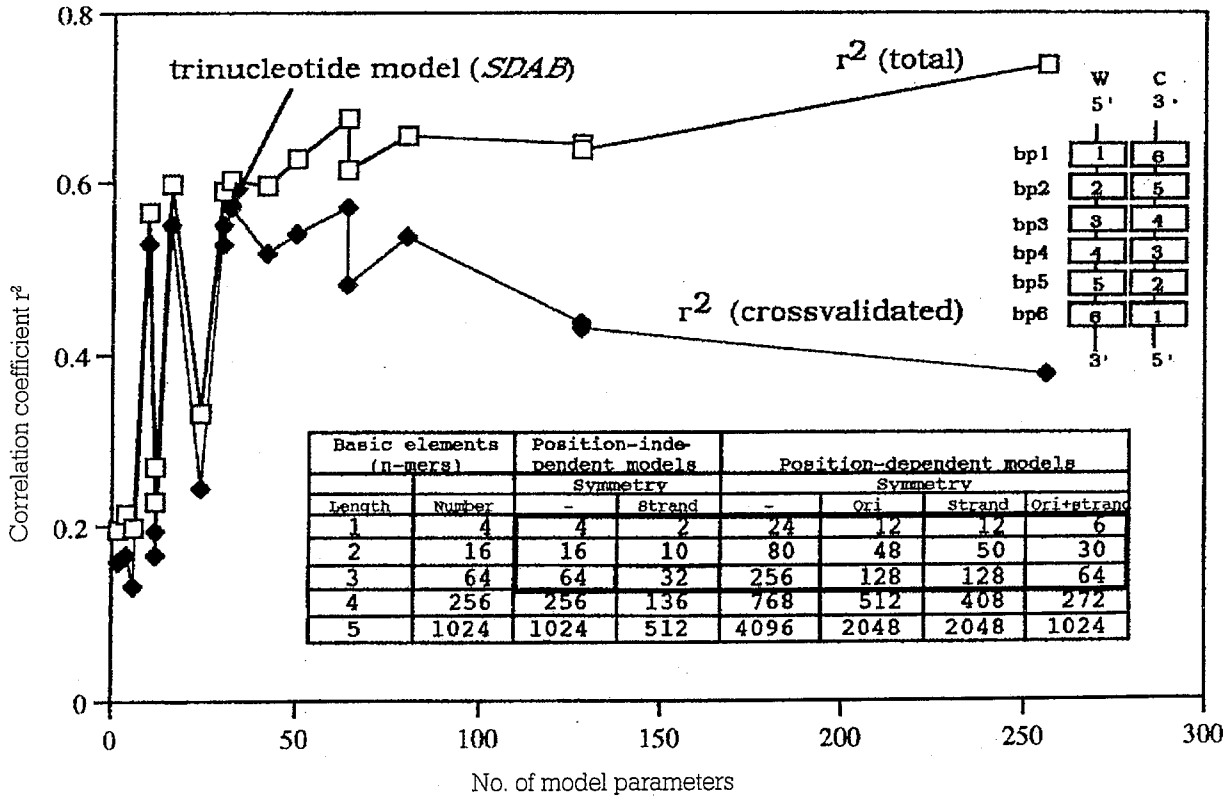
DNAseI binds to a sequence segment of 6 nucleotides, and the simplest is to suppose that the P_w probability of DNAseI cutting within a particular segment will depend on independent $p(a)_i$ contributions of subsequences (dinucleotides, trinucleotide, tetranucleotides)

$$P_w = \prod_{i=1}^{n} p(a)_i \qquad [4]$$

Equating P_w with the experimentally determined frequencies of cleavage, F_w, leads to a linear system of equations of the form

$$\ln = F_W = \sum_{i=1}^{n} \ln p(a)_i \qquad [5]$$

There are many ways how the "subsequences" can be selected and this will determine the complexity of the model. In the segment in contact with the enzyme (Figure 2, inset) there are 6 basepairs , 5 dinucleotide steps 4 tetranucleotide steps etc. In principle we can consider each of these independent, so we will have to determine separate parameter values for each position. There are plausible ways to decrease the number of parameters: For example, one can introduce strand symmetry, since in fact AA is supposed to have the same effect as TT. In addition, one can introduce "orientation symmetry" by assigning the same value for a dinucleotide in position 1 on either the W or the C strands. Finally, one can completely neglect the position effect, and this leads to position-independent models in which only the type of the dinucleotides or trinucleotides matter. These possibilities all lead to different models with different number of parameters, as shown in the table of Figure 2. The data set of Brukner et al. *(10)* contained 709 DNAse I cleavage data obtained in triplicate. In order to determine which of the theoretically possible models can be fit to the data we first selected those which have substantially less parameters than the data. As expected models with a higher number of parameters pro-



Basic elements (n-mers)		Position-independent models		Position-dependent models			
		Symmetry			Symmetry		
Length	Number	-	Strand	-	Ori	Strand	Ori+strand
1	4	4	2	24	12	12	6
2	16	16	10	80	48	50	30
3	64	64	32	256	128	128	64
4	256	256	136	768	512	408	272
5	1024	1024	512	4096	2048	2048	1024

FIGURE 2. *Selection of model complexity. DNAse I contacts a window of 6 basepairs of DNA (right inset). The number of parameters (table) is given as number of positions times the number of elements. The number of elements and positions can be reduced by allowing strand symmetry (i.e. AA=TT) or 5'-3' orientational symmetry, respectively. The models indicated by the rectangle were fit to the 709 quantitative DNAseI digestion data by least squares fitting as described* (10, 11). *A randomly chosen 80% of the data was used to fit a model used to calculate the cross-correlated correlation coefficient on the remaining 20% of the data. The model indicated with the arrow ("trinucleotide model") was used in Brukner et al, 1995* (10, 11) *and in this work.*

duce better fits according to equation [5] (r^2(total) in Figure 2). A simple strategy for choosing and optimal model is cross-validation (complexity regulation) by which one fits the model to a randomly chosen of the data ("learning-set") and determines the correlation coefficient on the remaining data ("test set") not included into the fit. This cross-validated correlation coefficient will not increase monotonously with the number of parameters but will have an optimum value or optimum range. We used a learning set/test set ratio of 80:20 and found that in fact the cross-validated correlation coefficient has an optimum in the range of 30-40 parameters. Therefore we chose the simplest trinucleotide model, which happened to have highest cross-validated correlation coefficient value (indicated by an arrow in Figure 2). This is a position-independent model so one can interpret the parameters as a bendability values characteristic of a ds trinucleotide.

The second question is how to transform the bendability parameter into a spring constant or Young's modulus directly usable in the physical model. The DNAseI -derived parameters correspond to bending flexibility measured on an arbitrary relative scale. Since flexibility proportional to the inverse of rigidity, for a simple linear system one can take

Relative flexibility $\sim 1/B \sim 1/E$

Using this assumption one can calculate sequence-dependent E-values for an ideal, flexible rod model, on the condition that the average Young's modulus values should remain equal to the experimentally determined value, i.e. 3.4×10^8 N/m^2 *(43)*. The values are shown in Table 1.

TABLE 1

ds Tri-nucleotide	DNAseI bendability		Consensus-bendability	
	[au]	Stiffness [10^8N/m^2]	[au]	Stiffness [10^8N/m^2]
AAA/TTT	0.1	7.176	0.05	6.947
AAC/GTT	1.6	6.272	2.65	5.323
AAG/CTT	4.2	4.736	4.70	4.047
AAT/ATT	0.0	7.237	0.35	6.763
ACA/TGT	5.8	3.810	5.50	3.562
ACC/GGT	5.2	4.156	5.30	3.666
ACG/CGT	5.2	4.156	5.30	3.672
ACT/AGT	2.0	6.033	3.90	4.528
AGA/TCT	6.5	3.410	4.90	3.930
AGC/GCT	6.3	3.524	6.90	2.715
AGG/CCT	4.7	4.445	5.05	3.832
ATA/TAT	9.7	1.613	6.25	3.085
ATC/GAT	3.6	5.087	4.45	4.207
ATG/CAT	8.7	2.169	7.70	2.206
CAA/TTG	6.2	3.581	4.75	4.010
CAC/GTG	6.8	3.239	6.65	2.866
CAG/CTG	9.6	1.668	6.90	2.702
CCA/TGG	0.7	6.813	3.05	5.062
CCC/GGG	5.7	3.868	5.85	3.354
CCG/CGG	3.0	5.440	3.85	4.559
CGA/TCG	5.8	3.810	7.05	2.599
CGC/GCG	4.3	4.678	5.90	3.317
CTA/TAG	7.8	2.673	5.00	3.862
CTC/GAG	6.6	3.353	6.00	3.262
GAA/TTC	5.1	4.214	4.05	4.442
GAC/GTC	5.6	3.925	5.50	3.544
GCA/TGC	7.5	2.842	6.75	2.787
GCC/GGC	8.2	2.448	9.10	1.389
GGA/TCC	6.2	3.581	5.00	3.869
GTA/TAC	6.4	3.467	5.05	3.819
TAA/TTA	7.3	2.955	4.65	4.065
TCA/TGA	10.0	1.447	7.70	2.218

Modelling of Minicircles of Curved and Straight DNA

In order to illustrate that the macroscopic bendability of a DNA model is anisotropic, we designed a simple experiment outlined in Figure 3. A rod model is circularized into a minicircle (a) which is then writhed around (b) and the energy of the model is determined by finite element methods as a function of the rotation angle (c). We find that sequences that are repeats of curved DNA motifs in fact will have a rotational preference, i.e. there will be one stable energy minimum (Figure 3). This means, at the same time, that such a rod-model has a preferred direction of bending, so as a result of thermal fluctuations it will preferentially bend into one direction. In other terms, the physically measurable average conformation of such a model will be curved.

Another consequence of the energy minimum found in the circles is that, in the minimum energy conformation of helically phased repeats, certain motifs will face inwards and the others outwards. For example, in repeats of AAAAGGGCCC, the GGGCCC motif faces inwards and the AAAAA are on the outer side of the circle. The roll values at the central GC are the highest while there are slight negative rolls in the AAAAA tract. Since the model does not contain any static component (moreover the AAA tracts are quite stiff, especially to the negative roll direction) this finding may seem unexpected. However, on a purely geometric basis it is quite plausible that the motifs on the outside of the circle will have negative rolls, irrespective of the sequence.

Sequences that are known to be straight, on the other hand have bi-stable energy profiles i.e. a given motif is either on the inside or on the outside of the circle (Figure 3b). This behaviour can however be easily explained by the fact that the straight motifs are designed in such a way that they contain motifs in antihelical phasing. Since both the curved and the straight sequence motifs are correctly predicted by the static dinucleotide models such as that of Bolshoy et al *(6)* or of Olson et al *(35)*, mans that the anisotropic bendability model is in good qualitative agreement with the static models. In other terms, helical phasing of anisotropic bendability seems to be a sufficient basis for curvature. This observation points to the problem that "static' and "dynamic" curvature which are so clearly distinguished in theoretical models, can not be clearly separated in experimental measurements, i.e. it is not possible to decide if DNAseI or nucleosome parameters represent bendability or static bending properties.

The same modelling experiment can be used to illustrate how the anisotropy of basepair elements influ-

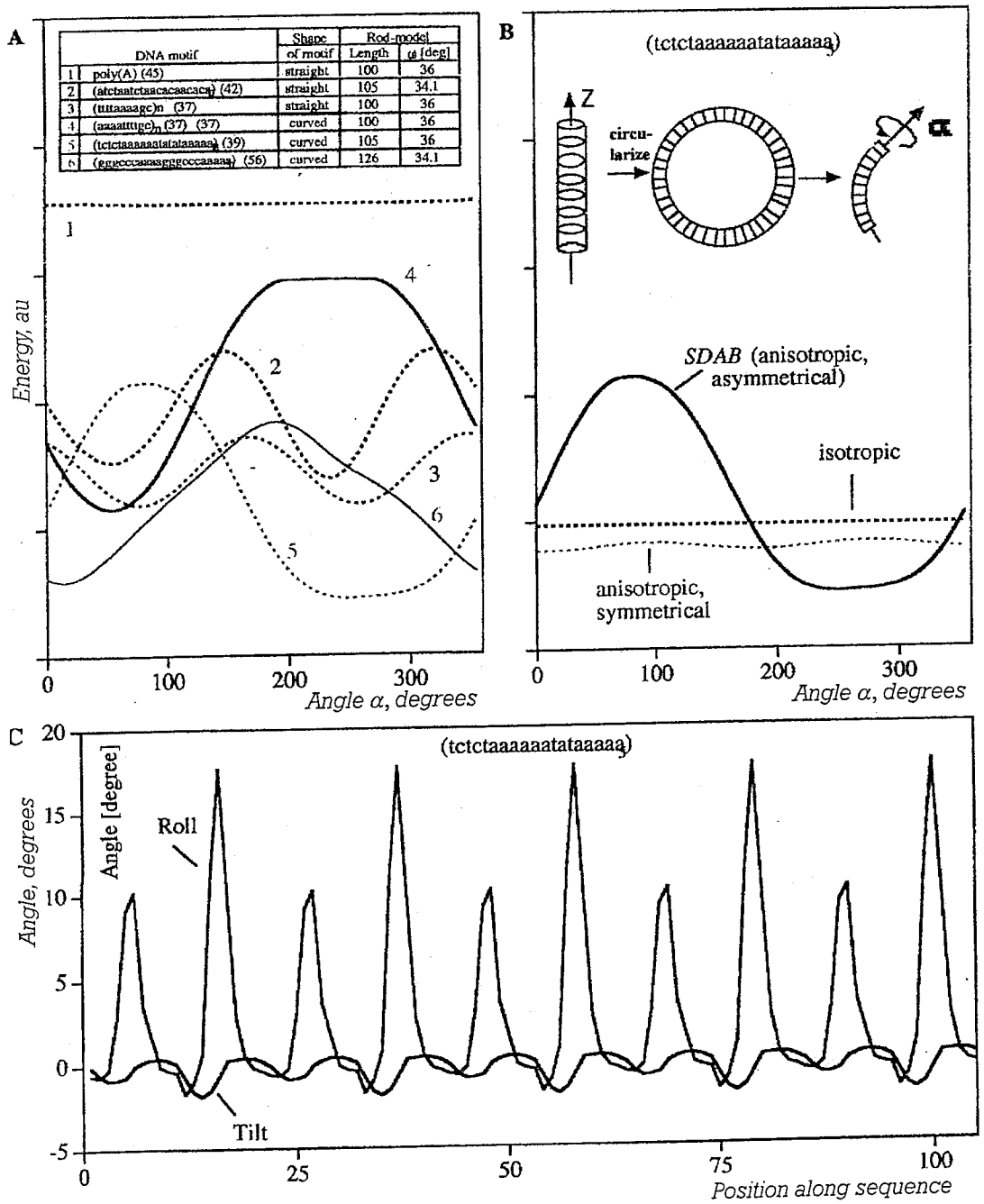


FIGURE 3. *Testing bending anisotropy in DNA by finite element methods. A DNA model was built (as outlined in Figure 1) from each repeat motif (inset in A, top left). The model was bent into a circle and the circle was twisted around the now circular z-axis as shown (inset, in B, top right) The energy of the model was calculated by finite element methods plotted as a function of the angle of twisting (A,B). Curved motifs exhibited single energy minima i.e. clear bending preferences in one direction. Straight (i.e. antihelically phased) repeat motifs have two minima at opposite bending directions (A). However, curved motifs exhibit single minima only with the asymmetrical, aniso-tropic* SDAB *model (B); the other models gave double or no minima. The roll and twist angles of the models show the expected periodicities, and negative values at the outside of the circle (C) (Note that the average energy of the models depends on the sequence length. cf. M.G. Munteanu et al., to be published elsewhere)*

ences anisotropy of the circle model built from a curved sequence motif (Figure 3d). Isotropic elements have no rotational preferences. If the elements are equally bendable to the major and minor grooves, the model will have two bending-energy minima. One single energy minimum was obtained only if the elements are anisotropic.

Protein/DNA Binding: DNA Rigidity vs. Complex Stability

Repressor proteins bind to short DNA motifs with high specificity and the DNA is often bent in the resulting protein/DNA complex. While the conformation of the operator DNA within the chromosome can not be easily determined, it is known that oligonucleotides corresponding to many of the operator DNA sequences are not intrinsically curved so bending is induced by the binding of the protein. If this is the case, the rigidity of the operator DNA can be expected to play a role in the binding. Takeda et al. *(45)* determined a free energy values for the binding of the Cro protein to various short oligonucleotides that contained operator sequence as well as a number of others devoid of such motifs ("non-operators"). By plotting the free energy values against the rigidity of the oligonucleotides (Figure 4) we find that cognate (operator) and non-cognate (non-operator) DNA follow two adverse, quasi-linear relationships. In the operator sequences, ΔG is higher for stiffer molecules (R=0.95) , i.e. the stiffer the molecule, the weaker the binding. This in fact can be expected since Cro has to curve the molecule, and the energy required is linearly proportional to the stiffness [eqn 2].

In non-operator sequences, on the other hand, ΔG is lower for stiffer sequences (R=-0.99), i.e. the stiffer the sequence, the stronger the binding. For the explanation of this phenomenon we consider a simplistic model (Figure 4, inset) in which Cro first binds to the oligonucleotide in a non-specific manner and reduces the free movement (thermal fluctuations) of DNA, which results in an entropy loss. Since the elastic entropy can be calculated from the $<\theta>^{1/2}$ root mean square fluctuations of the model *(21, 33, 41)*, the entropy change can be calculated as

$$\Delta S = nR\ln\left[\frac{<\theta^2_{bound}>^{1/2}}{<\theta^2_{free}>^{1/2}}\right] = nR\ln\left[\frac{E_{free}}{E_{bound}}\right] \quad [7]$$

where E_{free} is the average Young's modulus from eqn[3], n is the number of degrees of freedom and E_{bound} is the Young's modulus of the bound (quasi immobilized) DNA. Since E_{free} is smaller than E_{bound}, this equation an adverse relationship similar to that shown in Figure 4, and which is in fact very near to linear in the range of the experimental data shown in Figure 4. In other terms, the relationship shown in Figure 4 is qualitatively explained by the entropy loss, in accordance with the intuitive expectation that the "immobilization" of a stiff DNA substrate will take less energy on binding.

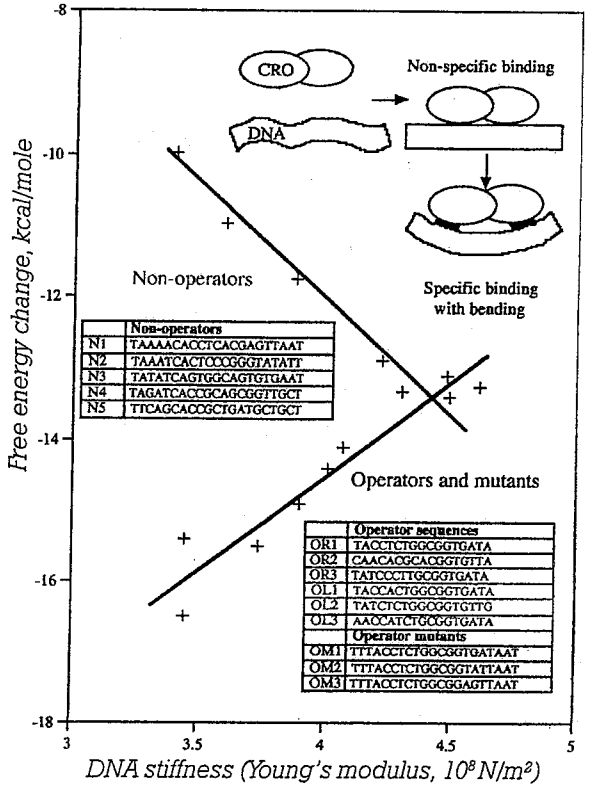


FIGURE 4. *Relationship between the average stiffness (Young's modulus) of DNA and the free energy change (DG) of operator and non-operator DNA sequences. Equation [7] was fitted to the non-cognate data (R=0.99) and a straight line (eqn.[2]) was fitted to the cognate data (R=0.95). The sequences and the model are shown in the insets* (22).

Prediction Tools

By plotting bendability parameters along a DNA sequence provides a qualitative picture on bendable or rigid regions. In protein/DNA complexes with known 3D structures, bendability shows a good correlation with the roll angles found in the crystals. The relative bendability differences are however small, so bendability plots of long DNA sequences do not show striking features. However, helical circle diagrams - a technique widely used for protein α-helices - shows characteristic differences between curved, straight and rigid segments (Figure 5). In these diagrams, the bendability is plotted as a vector pointing towards the major groove of each basepair. In randomly chosen DNA segments these plots are close to symmetrical with a vector sum are close to zero.. In curved DNA, the plots are asymmetrical and there is a substantial resulting vector. Based on this observation, we calculate and index for the helical asymmetry of the distribution, H, which we use as a measure of *curvature propensity*:

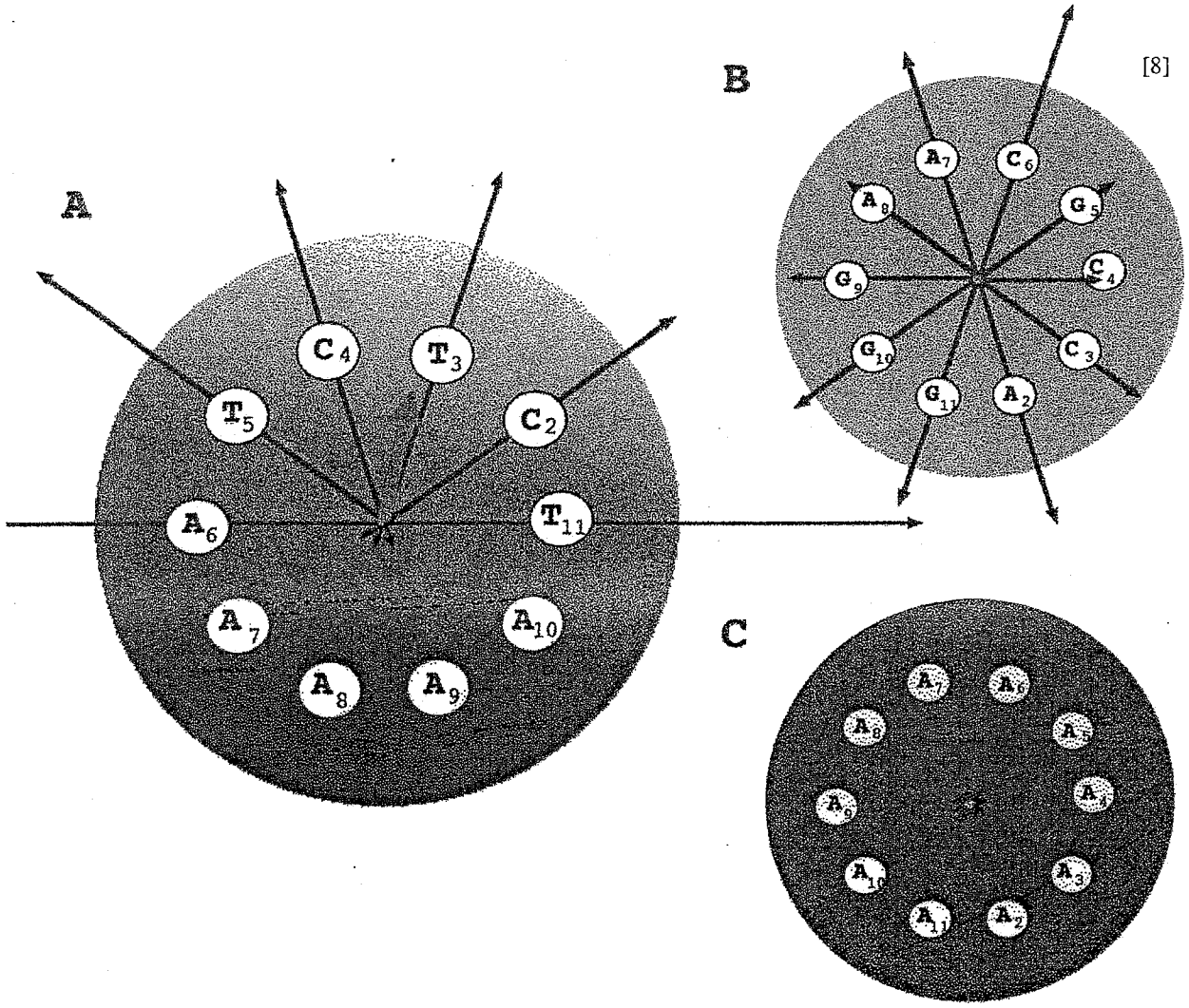


FIGURE 5. *Vectorial representation of bendability (helical circle diagrams). The length of black arrows is proportional with that of the bendability parameter at the given sequence position. The white arrow is the vectorial average of the bendability vectors (given in eqn [1]. The length of the average vector is negligible in Fig. 1C and 1D, so it is denoted by a dot only). The radius of the shaded circle indicates the average bendability of genomic sequences (about 5.3). A: A curved sequence motif (A)CTCTAAAAAT(A) designed by Ulanovsky et al. (47) B: A straight but bendable sequence from the lambda phage OR_3 operator region (30, 31, 47). (C)ACCGCAAGGG(A) C: poly-A sequence.*

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

where f_i is the bendability parameter (taken from Table 1) for position i, w is the twist angle (36° for ideal B-DNA) and n is the number of vectors in the segment (usually a segment length of 32 residues i.e. approximately 3 helical turns are used for the calculation). H will be positive for curved sequences and close to zero for straight ones (Table 2). Originally, equation [8] was used with the DNAseI -derived bendability data. Table 2 also includes figures calculated with the nucleosome-based parameters as well as the so-called consensus bendability scale *(16)*. In general there is a good agreement between the various methods. Therefore this calculation - like all predictions - has to be considered approximate, so absolute threshold values may not be defined. Nevertheless, the correlation with experimental curvature values in these and in previously reported examples *(16)* is satisfactory. Even though we tried to select sequence motifs that have been found curved by various research groups, it has to be mentioned that curvature values determined by gel mobility analysis strongly depend on the experimental conditions *(42)* and on the length of the tested molecule *(13)*, so a quantitative correlation between prediction and measurement can not be expected. The "curved" and "straight" DNA

TABLE 2

Analysis of curved and straight sequence motifs with various methods

No.	Origin (reference)	Sequence		Bendability [a.u.]		Curvature propensity [a.u.]		Curvature [degree/helical turn]			
			G+C conten	DNAse I *(10, 11)*	Consensus [this work, *(16)*]	DNAse I *(10, 11)*	Consensus [this work, *(16)*]	Bolshoy et al l, elfo *(6)*	Olson et al, X-rays*(20)*	Uljanov& James, NMR *(48)*	Nucleo-some *(19, 39)*
		Curved DNA									
1	Synthetic *(23)*	$(aaaattttgc)_n$	0.200	2.78	2.4	21.1	9.83	26.221	6.892	18.333	13.738
2	Synthetic *(23)*	$(aaaattttcg)_n$	0.200	2.22	2.3	8.18	9.35	21.034	3.801	13.193	17.697
3	Synthetic *(14)*	$(tctcaaaaaacgcgaaaaaaccggaaaaaagc)_n$	0.375	3.14	3.25	7.94	9.24	27.078	8.196	19.430	17.126
4	Synthetic *(26)*	$(ccgaaaaagg)_n$	0.500	3.86	4.31	14.0	17.69	14.730	6.844	20.999	23.335
5	Synthetic *(47)*	$(tctctaaaaaatatataaaaa)_n$	0.095	4.79	3.19	52.3	10.48	27.767	3.010	4.880	10.927
6	Synthetic *(26)*	(ggcaaaaaac)n	0.400	3.20	3.26	14.6	15.46	26.788	12.016	20.093	20.404
7	*L.tarentolae* kinetoplast *(9)*	ccaaaaatgtcaaaaaataggcaaaaaatgcc	0.313	3.76	3.53	20.0	17.25	26.005	6.438	16.027	19.606
8	Synthetic *(9)*	aaaaactctctaaaaactctccctagaggggccctagagggc	0.500	5.19	4.68	3.03	3.88	19.405	7.815	5.893	13.330
9	Synthetic *(9)*	aaaaactctctaaaaactctaagaggggccctagagggccc	0.488	4.90	4.53	3.82	2.12	18.298	6.650	6.579	16.312
10	*C.risortia* bent sat. DNA *(25)*	agaattgggacaaaaattggaaatttttaaggg	0.303	2.86	2.90	4.36	2.87	18.208	6.794	13.579	11.915
		Straight DNA									
11	Synthetic *(4)*	$(atctaatctaacacaacaca)_n$	0.300	5.14	4.44	0.00	0.00	0.769	0.456	2.091	1.275
12	OR3 operator region *(30)*	actacgttaaatctatcaccgcaagggataaa	0.375	5.01	4.43	1.72	0.39	10.394	5.536	5.193	5.859
13	OR3 region, mutated *(31)*	actacgttaaatctatcaccacaagggataaa	0.344	4.96	4.39	1.78	0.45	10.960	5.502	4.302	6.165
14	poly-A[2]*(53)*	$(a)_n$	0.000	0.10	0.063	0.00	0.00	0.008	0.000	0.008	0.000
15	Synthetic *(23)*	$(ttttaaaacg)_n$	0.200	2.86	2.55	0.0025	0.03	1.562	7.080	14.593	10.693
16	Synthetic *(23)*	$(ttttaaaagc)_n$	0.200	3.60	3.28	0.0072	0.87	1.709	0.820	16.002	16.325

[1]The angular deflection [degree per helical turn] was determined with the BEND algorithm *(19)* using a window size w=31 (~3 helical turns) and the values were corrected to a helical repeat length of 10.5 nucleotides.

[2]By definition, homopolymers should give zero curvature. The non-zero value indicates the numeric precision of the calculation.

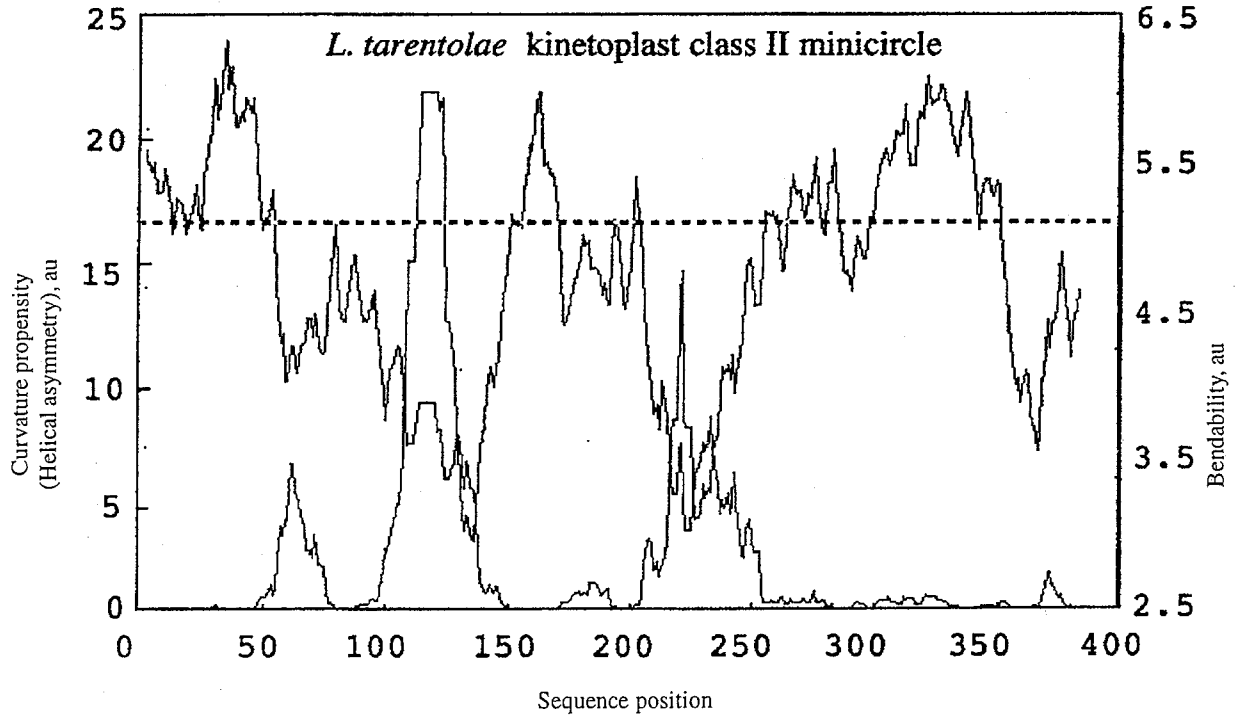


FIGURE 6. *Curvature propensity (Helical asymmetry) (equation [8]) and bendability versus sequence plot (Sequence: Genbank LEIKPMNC2). The dotted line indicates the average bendability value of DNA. The average helical asymmetry value (which is close to zero) is not indicated. (based on a 1D plot from http://www.icgeb.trieste.it/dna/bend_it.html).*

motifs used in this comparison were classified primarily on the basis of gel-electrophoresis data - in fact, the dinucleotide model of Bolshoy et al. *(6)*, which is itself based on gel electrophoresis data correctly predicts all of them while overpredicts two of the two repressor cognates that are known to be straight in the absence of the repressor. The model of Ulyanov and James *(48)* is based on NMR measurements (with no fitting to electrophoresis data) and gives slightly different predictions. The qualitative bendability based predictions are grossly similar and both allow one to distinguish between curved and straight motifs, as do the nucleosome model and the dinucleotide models of Bolshoy et al. *(6)* Naturally it is possible that the apparent differences are not mispredictions but reflect the true differences between the experimental techniques (e.g. NMR or X-rays vs. electrophoresis).

Both curvature propensity and average bendability can be plotted against the sequence, and curved motifs will appear as peaks in the curvature propensity profile (Figure 6). 2D plots are especially useful for longer sequences: In these, curvature propensity is plotted against bendability and each 32 residue long segment of the sequence will appear as one dot in the plot. Curved, rigid and flexible segments will occupy different regions (Figure 7A). While most DNA segments will fall in the region of average bendability and low curvature propensity, sequences containing curved segments will have a characteristic, asymmetric distribution (Figure 7B) which disappears if the sequences are random shuffled. This shape of distribution is characteristic of kinetoplast sequences containing curved segments. Finally, for very long sequences, such as genomes, a histogram-like 3D plot can be used in which the third, vertical dimension is the frequency of occurrence of a segment corresponding to a given bendability-curvature range. Such histograms can reveal interesting differences between DNA regions. For example, an analysis of the human T-cell receptor locus shows that most of the rigid, flexible and curved regions are in the non-coding regions while the protein coding regions show a comparatively tight distribution (Figure 8).

The plotting programs described here are available via WWW at *http://www.icgeb.trieste.it/dna/dnatools.html/*. In addition to the curvature propensity calculated according to the sequence-dependent DNA bendability (*SDAB*) model, the server includes (1D and 2D) plots for more than thirty DNA sequence parameters as well as various static curvature models based on the BEND algorithm of Goodsell and Dickerson *(19)*.

Distribution of Bendability and Curvature within Genomes

The bendability and static curvature values were determined for a number of complete genomes currently available (Figures 9—10).

The average curvature in genomic sequences (Table 3) is between 5.4 and 6.8 degrees per helical turn, even though less than 20% of DNA is below the limit of

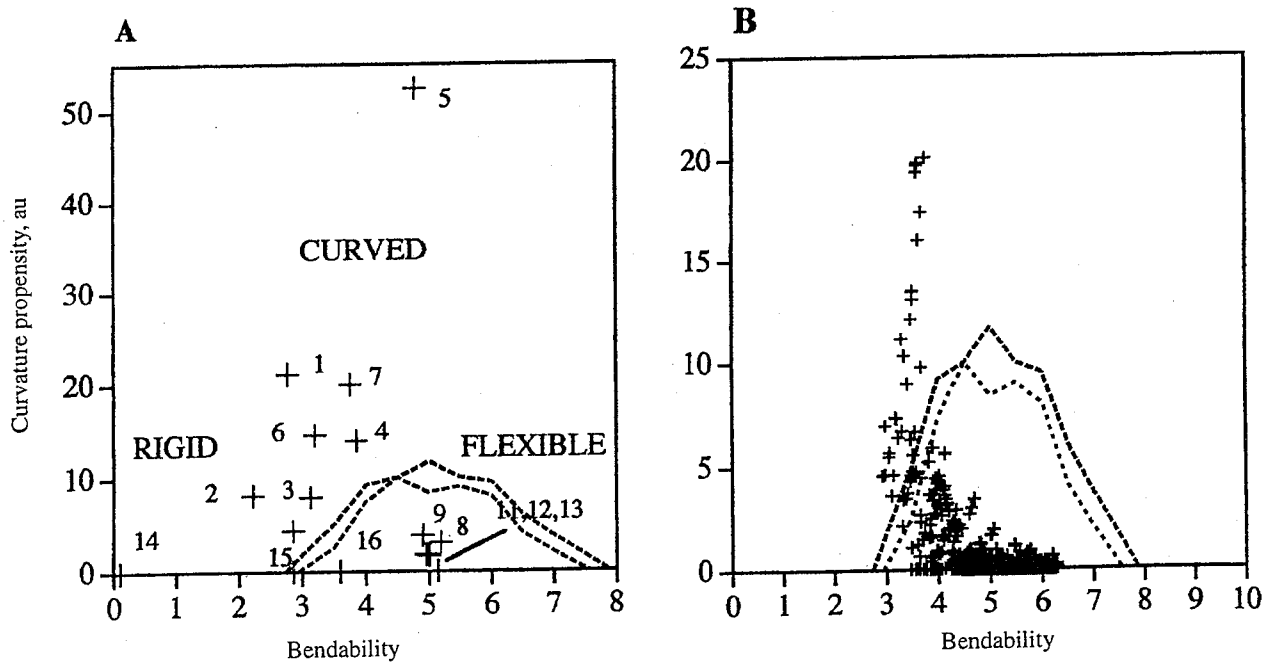


FIGURE 7. *Curvature vs. bendability plots. A: Curved (1-10) and straight (11-16) sequences from Table 2. B:* Leishmania tarentolae *class II minicircle (Sequence: Genbank LEIKPMNC2). The values are calculated for 30 bp sequence segments. (based on 2D plots from http://www.icgeb.trieste.it/dna/curve_it.html). The dashed lines indicate the border of random sequences obtained by random-shuffling of the sequences of* H. influenzae *genome and yeast chromosome III., respectively.*

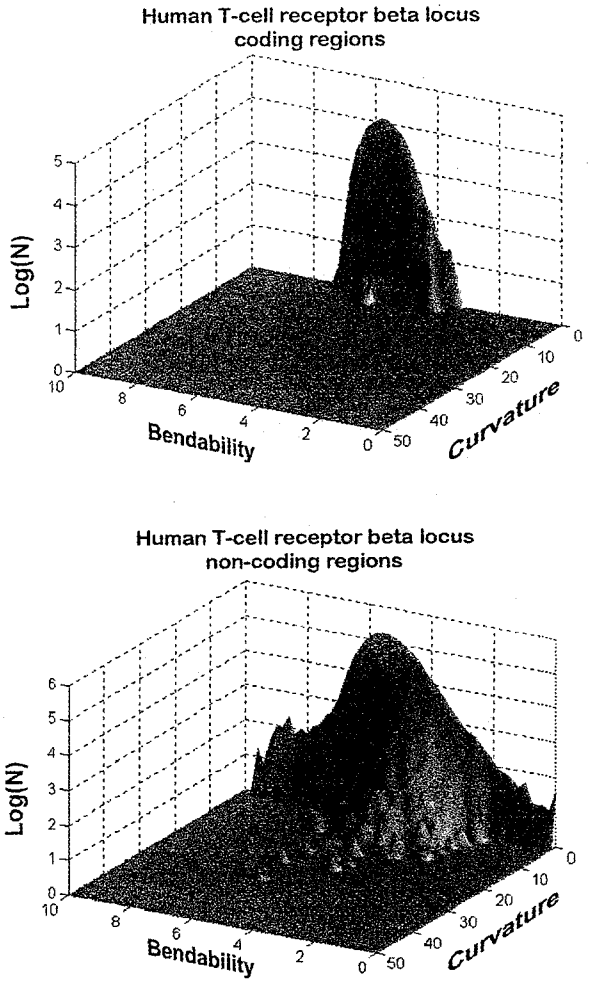


FIGURE 8. *3D Curvature propensity (eqn.[8]) vs. bendability histogram of the Human T-cell receptor locus (Genbank: humtcrb) The protein coding regions (a) have a tight distribution around average bendability and low curvature. The non-protein-coding regions (b) have a higher number of stiff (low bendability), flexible and curved segments.*

3°/helical turn. This in accordance with the intuitive expectation that average DNA is reasonably straight if not exposed to external factors. Short sequence segments may have very different average values, as shown by the example of the *L. tarentolae* kinetoplast minicircle (Table 3). The maximum values found in the genomes (Table 3, column 6) are quite similar to those found with artificially designed sequence motifs (Table 2). However, it is conspicuous, that the longest stretches of continuous curvature (Table 3, column 9) do not reach the length of 100-200 bp, i.e. the oligonucleotide length used for quantitating curvature by gel mobility analysis *(26)*.

The curvature distributions were calculated with two models, that of Bolshoy et al. *(6)* and that of Ulyanov and James *(48)*, using the BEND algorithm as incorporated into the server software described above. In both cases, distribution of curvature appears quite similar in all the genomes tested (Figure 9). It seems to follow a typical, even though not symmetrical random distribution. The distribution is reminiscent of a gamma function which is often found with randomly distributed variables whose value can not be negative - curvature is actually such a case. The distribution is smooth, apparently there are no preferred values of curvature. There seems to be no clear-cut separation between the eukaryotes and the prokaryotes but a very weak separation into two groups is apparent. *E. coli*, *Synechocystis*, *B. subtilis* and *H. Influenzae* have an apparently higher average curvature than the other genomes, while *C. elegans* and *S. cerevisiae*, as well as the two *Mycoplasma* genomes are closer to the human genomic sequences. *M. jannaschii* is closer to the human sequences than to the other group that includes *E. coli*, *Synechocystis*, *B. subtilis* and *H . influenzae*. The average of bacterial genomes (not shown) contains more curved segments than the averages of higher organisms *(18)*. If we assume that most of the curvature in DNA occurs around promoter regions, DNA in prokaryotes should then be more curved on the average since transcription units are smaller. Another difference is that the DNA is not packed in the same way in pro- and eukaryotes. In *E.coli*, there are not enough copies of non- specific histone-like proteins like HU and HNS to condense all the bacterial as tightly as the histones do it in eukaryotes. So a

TABLE 3

Genomic averages of DNA curvature [degree/helical turn] in genomic DNA.

Genomic DNA	Size	Average bendability (s.d.) [au]	Average curvature (s.d.) [°/hel.turn]	Average G+C content (s.d.)	Max. curvature [°/.hel.turn] (G+C content)	% below 3°/hel. turn	% above 15°/hel. turn	Longest segment above 15°/hel. turn [bp]
1	2	3	4	5	6	7	8	9
B. subtilis	1983kbp	5.249±0.644	7.324±3.744	0.468	25.439 (0.533)	12.07	3.26	61
C. elegans	10080kbp	4.829±2.295	6.799±3.437	0.465	27.136 (0.533)	14.14	2.05	122
E. coli	4639kbp	5.470±0.694	7.609±3.848	0.509	27.410 (0.467)	11.11	4.23	90
H. influenzae	1830kbp	5.010±0.107	7.561±3.828	0.374	26.188 (0.433)	11.30	3.93	56
M. genitalium	580kbp	4.840±0.198	6.652±3.399	0.338	22.703 (0.333)	14.34	1.53	49
M. jannaschii	1665kbp	5.030±0.244	6.913±3.537	0.337	24.285 (0.433)	13.51	2.10	59
M. pneumoniae	816kbp	4.944±0.174	7.023±3.584	0.427	24.403 (0.600)	12.97	2.36	78
S. cerevisiae	12063kbp	4.731±2.668	6.673±3.299	0.374	27.351 (0.633)	14.51	1.64	69
Synechocystis	3573kbp	5.065±0.470	7.836±3.964	0.475	29.760 (0.633)	10.52	5.01	90
H. sapiens c fragments	278kbp	5.238±0.172	6.678±3.434	0.368	24.161 (0.467)	14.45	1.50	39

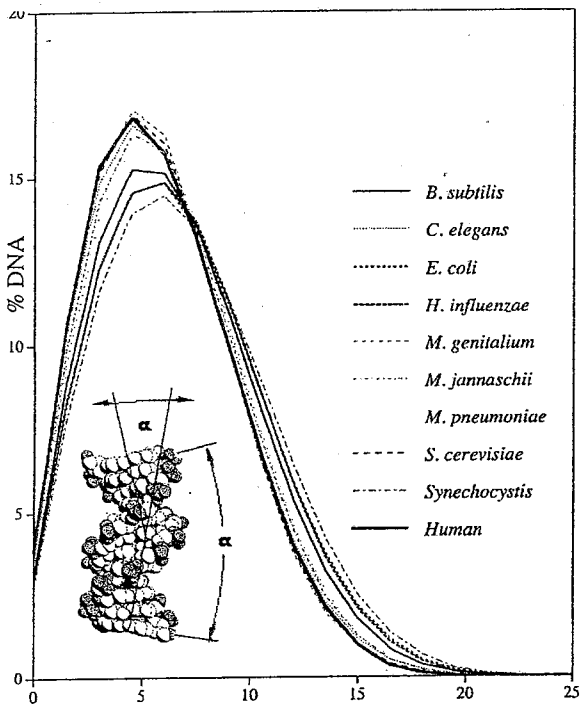


FIGURE 9. *Distribution of curvature in genomic DNA. The curvature was calculated with the BEND algorithm* (19) *using the dinucleotide parameters of Ulyanov and James* (48). *The genomes are listed in Table 3.*

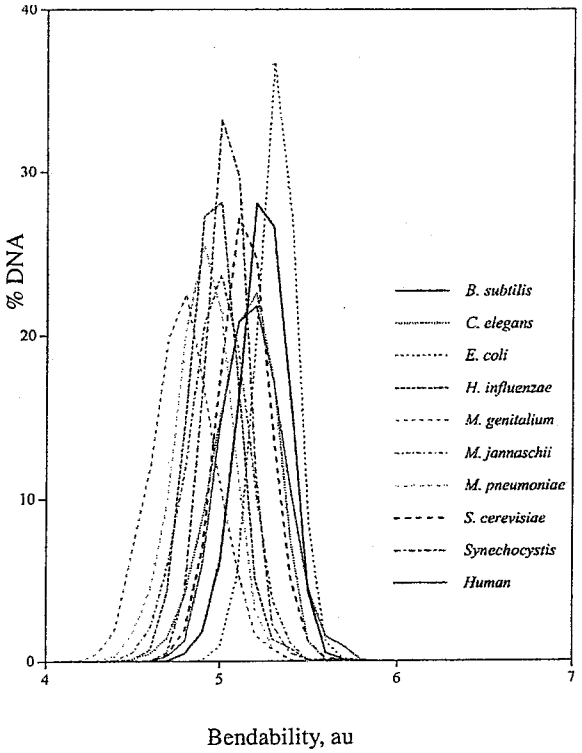


FIGURE 10. *Distribution of bendability in genomic DNA. The average bendability* (11) *was calculated for windows of 1000 bp. The genomes are listed in Table 3.*

higher 'intrinsinc' curvature in prokaryotes would certainly help compacting so much DNA in such a small space. We emphasize that, at least for the moment, these explanations are purely speculative. What is apparent, however, is that the maximum curvature values found in genomes are in the range found with synthetic oligonucleotides (Tables 2, 3).

The distribution of bendability follows a smooth, symmetrical distribution more reminiscent of a bell-shape. The average bendability of the genomes falls in the range of 4.5-5.5 arbitrary units, i.e. all the genomes are of "average bendability". We find differences between genomes even with windows as long as 1000 nucleotides (Figure 9). There is no clear-cut separation between prokaryotes and eukaryotes, even though the human sequences, *C. elegans* and *S. cerevisiae* fall near each other.

Discussion and Conclusions

The sequence-dependent DNA bendability model was developed with the aim of incorporating DNA bendability data into a simple physical model that *i*) can be tested with mechanical modelling (finite element) methods, and, *ii*) can be used for scanning large amounts of sequence data for bendable, rigid and potentially curved segments. It is an elastic rod model with sequence-dependence, anisotropy, and the helical symmetry of B-DNA. The rationale of using bendability data is that these can be more readily obtained from experiment than geometry parameters. The larger datasets might make it easier, on the one hand, to obtain more detailed and accurate model parameters, and, on the other, to include context dependence beyond the limits of static dinucleotide models. The idea of anisotropic DNA bendability is not new; early work by Schellman *(40)* and Zhurkin *(54)* already pointed out that anisotropy is important for explaining bending phenomena *(36)*. On the other hand, individual anisotropic models can differ in the way they are parametrized. The work presented here is mostly based on bendability data derived from DNAse I experiments *(10)*, in some cases we included data based on a more recent consensus bendability scale that had proven efficient in detecting both AA-type and GC-type curvature. Also, the philosophy of *SDAB* is slightly different from that of static geometry models since here DNA is considered as an originally straight elastic rod, and all curvature phenomena arise as a consequence of sequence-dependent and anisotropic bendability. This is an intended oversimplification which allows one to test whether bendability - i.e. the only property represented in a sequence-dependent fashion - can explain different phenomena. There are various ways to represent flexibility, e.g. Sarai et al. *(38)* as well as Goodsell and Dickerson *(19)* use average torsional angle values; however we preferred to use the Young's modulus which is more widely used in mechanics. Scal-

ing of bendability to Young's modulus is based on the approximation that the average of the parameters should correspond to the experimentally known average value. In fact, if the Young's modulus is helically phased along the sequence (i.e. has a periodicity close to that of B-DNA) then the macroscopic bendability of the rod model will be anisotropic and the net result will be an apparent curvature due to thermal fluctuations. In other terms, *SDAB* can account for "static" curvature, and essentially with one single parameter per basepair unit. (We note that dinucleotide models assign values to dinucleotide steps while the trinucleotide models assign the parameters to the central basepair. The dimensions of the elements used for mechanical modelling are identical, however).

One other difference with respect to static geometry models is the fact that the bendability parameters were not derived from fitting the model to electrophoresis data but come from an independent (even though naturally not unbiased) measurement. In spite of this, a good agreement was found first with protein/DNA complex data, and more recently with static curvature (17). Generally speaking, there is a good qualitative agreement between the models tested here since the differences between curved and straight motifs can be seen sufficiently by all models (Table 2).

The shape of bendability and curvature distributions do not show striking differences between genomes. Based on the average values of the distributions, one sees apparent groups in both cases, however the groups do not correspond to the ones expected based on phylogenetic (or compositional) similarities.

References

1. BANSAL M, BHATTACHARYYA D, RAVI B 1995 NUPARM and NUCGEN: Software for analysis and generation of sequence dependent nucleic acid structures. *Comput Appl Biosci 11:* 281-7
2. BATHE K J 1992 Finite element procedures in engineering analysis. Prentice-Hall Inc., New Jersy, USA.
3. BAUER W R, LUND R A, WHITE J H 1993 Twist and writhe of a DNA loop containing intrinsic bends. *Proc Natl Acad Sci USA 90:* 833-7
4. BEDNAR J, FURRER P, KATRITCH V, STASIAK A Z, DUBOCHET J, STASIAK A 1995 Determination of DNA persistence length by cryo-electron microscopy. Separation of the static and dynamic contributions to the apparent persistence length of DNA. *J Mol Biol 254:* 579-94
5. BOLSHOY A 1995 CC dinucleotides contribute to the bending of DNA in chromatin [letter]. *Nat Struct Biol 2:* 446-8
6. BOLSHOY A, MCNAMARA P, HARRINGTON R E, TRIFONOV E N 1991 Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc Natl Acad Sci USA 88:* 2312-6
7. BRESLAUER K J, FRANK R, BLOCKER H, MARKY L A 1986 Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci USA 83:* 3746-50
8. BRUKNER I, BELMAAZA A, CHARTRAND P 1997 Differential behavior of curved DNA upon untwisting. *Proc Natl Acad Sci U S A 94:* 403-406
9. BRUKNER I, DLAKIĆ M, SAVIĆ A, SUŠIĆ S, PONGOR S, SUCK D 1993 Evidence for opposite groove-directed curvature of GGGCCC and AAAAA sequence elements [published erratum appears in Nucleic Acids Res 1993 Mar 11;21(5):1332]. *Nucleic Acids Res 21:* 1025-9
10. BRUKNER I, SANCHEZ R, SUCK D, PONGOR S 1995 Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *Embo J 14:* 1812-8
11. BRUKNER I, SANCHEZ R, SUCK D, PONGOR S 1995 Trinucleotide models for DNA bending propensity: comparison of models based on DNaseI digestion and nucleosome packaging data. *J Biomol Struct Dyn 13:* 309-17
12. BRUKNER I, SUŠIĆ S, DLAKIĆ M, SAVIĆ A, PONGOR S 1994 Physiological concentration of magnesium ions induces a strong macroscopic c urvature in GGGCCC-containing DNA. *J Mol Biol 236:* 26-32
13. CALLADINE C R, COLLIS C M, DREW H R, MOTT M R 1991 A study of electrophoretic mobility of DNA in agarose and polyacrylamide gels. *J Mol Biol 221:* 981-1005
14. CALLADINE C R, DREW H R, MCCALL M J 1988 The intrinsic curvature of DNA in solution. *J Mol Biol 201:* 127-37
15. DE SANTIS P, PALLESCHI A, SAVINO M, SCIPIONI A 1990 Validity of the nearest-neighbor approximation in the evaluation of the electrophoretic manifestations of DNA curvature. *Biochemistry 29:* 9269-73
16. GABRIELIAN A, PONGOR S 1996 Correlation of intrinsic DNA curvature with DNA property periodicity. *Febs Lett 393:* 65-8
17. GABRIELIAN A, SIMONCSITS A, PONGOR S 1996 Distribution of bending propensity in DNA sequences. *Febs Lett 393:* 124-30
18. GABRIELIAN A, VLAHOVIČEK K, PONGOR S 1997 Distribution of Sequence-dependent curvature in genomic DNA sequences. *FEBS Letters 406:* 69-74
19. GOODSELL D S, DICKERSON R E 1994 Bending and curvature calculations in B-DNA. *Nucleic Acids Res 22:* 5497-503
20. GORIN A A, ZHURKIN V B, OLSON W K 1995 B-DNA twisting correlates with base-pair morphology. *J Mol Biol 247:* 34-48
21. GROMIHA M M, MUNTEANU M G, GABRIELIAN A, PONGOR S 1996 Anisotropic elastic bending models of DNA. *J Biol Phys 22:* 227-243
22. GROMIHA M M, MUNTEANU M G, SIMON I, PONGOR S 1997 The role of DNA bending in Cro protein-DNA interactions. *Biophysical Chemistry:* in press
23. HAGERMAN P J 1986 Sequence-directed curvature of DNA. *Nature 321:* 449-50
24. IVANOV V I, KRILOV D Yu, SHEHYOLKINA A K, CHERNOV B K, MIRCHENKOVA L E 1995 Decimal code contolling the B to A transition of DNA. *J Biomol Str Dyn 12:* A102
25. KITCHIN P A, KLEIN V A, RYAN K A, GANN K L, RAUCH C A, KANG D S, WELLS R D, ENGLUND P T 1986 A highly bent fragment of Crithidia fasciculata kinetoplast DNA. *J Biol Chem 261:* 11302-9
26. KOO H S, WU H M, CROTHERS D M 1986 DNA bending at adenine thymine tracts. *Nature 320:* 501-6
27. LAHM A, SUCK D 1991 DNase I-induced DNA conformation. 2 A structure of a DNase I-octamer complex *J Mol Biol 222:* 645-67
28. LANGOWSKI J, OLSON W K, PEDERSEN S C, TOBIAS I, WESTCOTT T P, YANG Y 1996 DNA supercoiling, localized bending and thermal fluctuations [letter]. *Trends Biochem Sci 21:* 50

29. LEWIS J, SANKEY O 1995 Geometry and energetics of DNA basepairs and triplets from first principles quantum molecular relaxations. *Biophy J 69:* 1068-1076

30. LYUBCHENKO Y, SHLYAKHTENKO L, CHERNOV B, HARRINGTON R E 1991 DNA bending induced by Cro protein binding as demonstrated by gel electrophoresis. *Proc Natl Acad Sci USA 88:* 5331-4

31. LYUBCHENKO Y L, SHLYAKHTENKO L S, APPELLA E, HARRINGTON R E 1993 CA runs increase DNA flexibility in the complex of lambda Cro protein with the OR3 site. *Biochemistry 32:* 4121-7

32. MANNING G S 1985 Packaged DNA. An elastic model [published erratum appears in Cell Biophys 1986 Feb;8(1):86]. *Cell Biophys 7:* 57-89

33. MARKO J F, SIGGIA E D 1994 Fluctuations and supercoiling of DNA. *Science 265:* 506-8

34. OLSON W K 1996 Simulating DNA at low resolution. *Curr Opin Struct Biol 6:* 242-56

35. OLSON W K, BABCOCK M S, GORIN A, LIU G, MARKY N L, MARTINO J A, PEDERSEN S C, SRINIVASAN A R, TOBIAS I, WESTCOTT T P, et al. 1995 Flexing and folding double helical DNA. *Biophys Chem 55:* 7-29

36. OLSON W K, ZHURKIN V B 1996 Twenty years of DNA bending. *In:* Sarma R H, Sarma M H *(ed)* Biological Structure and Dynamics. Adenine Press, Schenectady, p 341-370

37. SANTALUCIA J, ALLAWI H, SENEVIRATNE P A 1996 Improved nearest-neighbour parameters for predicting DNA duplex stability. *Biochemistry 35:* 3555-3562

38. SARAI A, MAZUR J, NUSSINOV R, JERNIGAN R L 1989 Sequence dependence of DNA conformational flexibility. *Biochemistry 28:* 7842-9

39. SATCHWELL S C, DREW H R, TRAVERS A A 1986 Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol 191:* 659-75

40. SCHELLMAN J A 1974 Flexibility of DNA. *Biopolymers 13:* 217-26

41. SCHLICK T 1995 Modeling superhelical DNA: recent analytical and dynamic approaches. *Curr Opin Struct Biol 5:* 245-62

42. SHLIAKHTENKO L S, LIUBCHENKO I, CHERNOV B K, ZHURKIN V B 1990 [The effect of temperature and ionic strength on the electrophoretic motility of synthetic DNA fragments]; Vliianie temperatury i ionnoi sily na elektroforeticheskuiu podvizhnost' sinteticheskikh fragmentov DNK. *Molekularnaya Biologija 24:* 79-95

43. SMITH S B, CUI Y, BUSTAMANTE C 1996 Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. *Science 271:* 795-9

44. SUGIMOTO N, NAKANO S, YONEYAMA M, HONDA K 1996 Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucl Acid Res 24:* 4501-4505

45. TAKEDA Y, SARAI A, RIVERA V M 1989 Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. *Proc Natl Acad Sci USA 86:* 439-43

46. TRAVERS A A, KLUG A 1990 *In:* N.R.a.W. Cozzarelli, J.C. *(ed)* DNA topology and its biological effects. Cold Spring Harbor laboratory, Cold Spring Harbor, p 57-106

47. ULANOVSKY L, BODNER M, TRIFONOV E N, CHODER M 1986 CURVED DNA: design, synthesis, and circularization. *Proc Natl Acad Sci U S A 83:* 862-6

48. ULYANOV N B, JAMES T L 1995 Statistical analysis of DNA duplex structural features. *Methods Enzymol 261:* 90-120

49. WESTON S A, LAHM A, SUCK D 1992 X-ray structure of the DNase I-d(GGTATACC)2 complex at 2.3 A resolution. *J Mol Biol 226:* 1237-56

50. WOOTTON 1996 Analysis of compositionally biased regions in sequence databases. *Methods Enzymo 266:* 554-571

51. YAKUSHEVICH L V 1994 Nonlinear DNA dynamics. *Physica D 79:* 77-86

52. YANG Y, TOBIAS I, OLSON W K 1993 Finite element analysis of DNA supercoiling. *J Chem Phys 98:* 1673-1686

53. YOUNG M A, RAVISHANKER G, BEVERIDGE D L, BERMAN H M 1995 Analysis of local helix bending in crystal structures of DNA o ligonucleotides and DNA-protein complexes. *Biophys J 68:* 2454-68

54. ZHURKIN V B, LYSOV Y P, IVANOV V I 1979 Anisotropic flexibility of DNA and the nucleosomal structure. *Nucleic Acids Res 6:* 1081-96

55. ZIENKIEWICZ O C a T, R.L. 1991 The finite element methods. McGraw-Hill.