

## SUPPLEMENTARY MATERIALS

to the article: A.F. Muterko “Quaternion modeling of the helical path for analysis of the shape of the DNA molecule”

### Biological processes and phenomena that depend on DNA bending and curving

Process or phenomenon	References
Recombination	Schultes, Szostak, 1991; Milot et al., 1992
Replication	Zahn, Blattner, 1985; Hertz et al., 1987; Caddle et al., 1990; Gimenes et al., 2008
The excision of damaged nucleotides from DNA	Shi et al., 1992
Transcription regulation	Rees et al., 1993; Iyer, Struhl, 1995; Nickerson, Achberger, 1995; Pérez-Martín, de Lorenzo, 1997; Koch, Thiele, 1999; Bolshoy, Nevo, 2000; Hizver et al., 2001; Bi et al., 2004; Kozobay-Avraham et al., 2006; Gimenes et al., 2008; Raveh-Sadka et al., 2012; Struhl, Segal, 2013 and references therein
DNA-protein recognition and interactions	Dickerson, Chiu, 1997; Rohs et al., 2009
Nucleosome organization and positioning	Segal et al., 2006; Rohs et al., 2009; Segal, Widom, 2009; Struhl, Segal, 2013
Chromosome folding	Tolstorukov et al., 2005
DNA packaging in the nucleoid	Tolstorukov et al., 2005; Mrazek, 2010
Protection of the chromosome against integration of prophages	Abel, Mrazek, 2012; Tong, Mrazek, 2014
Supercoiling	Herzel et al., 1998; Schieg, Herzel, 2004

Furthermore, DNA curvature related sequence elements provide a general genetic mechanism for regulation of promoter activity by different transcription factors and fine-tuning expression in a predictable manner, with resolution that can be even finer than that attained through altering transcription factor sites (Raveh-Sadka et al., 2012). The function of such a promoter element mainly depends on its intrinsic structure, not its interaction with sequence-specific DNA-binding proteins (Iyer, Struhl, 1995). For this reason it is important to predict both intrinsically curved and easy to bend, flexible DNA regions.

There is an assumption that the molecular shape of DNA is under selection and that it reflects evolutionary history (Parker et al., 2009).

## Methods of obtaining parameters of the base pair steps

Method	References
Circularization	Ulanovsky, Trifonov, 1987; Bolshoy et al., 1991
Gel-mobility	Calladine et al., 1988; Liu, Beveridge, 2001
DNase I digestion	Brukner et al., 1995
Nucleosome positioning	Calladine, Drew, 1986; Satchwell et al., 1986
X-ray crystallography	Calladine et al., 1988; Bansal et al., 1995; Gorin et al., 1995; Olson et al., 1998
NMR spectroscopy	Ulyanov, James, 1995
Conformational energy minimization	De Santis et al., 1990
Computer simulation	Dixit et al., 2005; Fujii et al., 2007; Lavery et al., 2010; Beveridge et al., 2012

**Base pair step parameters used in the present study for reconstruction of the DNA helical path**

The calculation of reconstructed DNA trajectories was performed based on average values of the local helical parameters deduced from the published B-DNA solution structures, PDB ID: 1RVI, 1RVH, 1FZX, 1G14, 1AXP, 2L8Q, 2M2C, 2MCI, 1DK9, 2K0V, 1NEV, 1LAI according to recommendations described in the present article.

The structural properties of dinucleotides in the DNA molecule depend upon the flanking base pairs (Beveridge et al., 2004; Dixit et al., 2005; Yonetani, Kono, 2009). The effect of the highly flexible steps on the structure of positions located farther than the immediate neighbor is consistently large and would be present outside the di- and even tetra-nucleotide (Dixit et al., 2005). For this reason, parameters of base pair steps were applied in tetra-nucleotide context, where possible.

In di-nucleotide context:

	Shift (Å)	Slide (Å)	Rise (Å)	Tilt (°)	Roll (°)	Twist (°)
AA	0.01	-0.72	3.22	-0.81	0.19	35.19
TT	-0.01	-0.72	3.22	0.81	0.19	35.19
AG	0.03	-0.80	3.24	0.31	-0.60	35.62
CT	-0.03	-0.80	3.24	-0.31	-0.60	35.62
AT	0.01	-0.67	3.54	0.54	-2.60	33.84
CG	0.03	-0.60	3.14	0.53	1.89	37.17
GA	-0.02	-0.64	3.08	0.70	0.06	36.75
TC	0.02	-0.64	3.08	-0.70	0.06	36.75
GT	0.01	-0.44	3.14	0.64	1.93	35.82
AC	-0.01	-0.44	3.14	-0.64	1.93	35.82
TG	0.03	-0.32	3.11	-0.05	3.50	34.80
CA	-0.03	-0.32	3.11	0.05	3.50	34.80
GC	0.16	-0.31	3.28	1.13	1.08	37.50
GG	0.18	-0.27	3.29	1.65	2.86	38
CC	-0.18	-0.27	3.29	-1.65	2.86	38
TA	-0.30	0.06	3.13	-1.03	2.50	35.06

In tetra-nucleotide context:

	Shift (Å)	Slide (Å)	Rise (Å)	Tilt (°)	Roll (°)	Twist (°)
AAAT	-0.23	-0.4	3.15	-4.84	0.39	37.47
ATTT	0.23	-0.4	3.15	4.84	0.39	37.47
AAGA	0.08	-1.15	3.15	0.34	0.39	31.61
TCTT	-0.08	-1.15	3.15	-0.34	0.39	31.61
AAGC	0	-1.26	3.19	-0.55	-1.2	35.47
GCTT	0	-1.26	3.19	0.55	-1.2	35.47
AATT	0.02	-0.98	3.89	0.02	-13.47	32.03
ACGC	0.01	-0.37	3.2	1.05	8.27	35.65
GCGT	-0.01	-0.37	3.2	-1.05	8.27	35.65
ACGG	-0.17	-0.55	3	3.28	3.43	34.56
CCGT	0.17	-0.55	3	-3.28	3.43	34.56
AGAA	0.26	-0.74	3.17	1.67	1.83	36.56
TTCT	-0.26	-0.74	3.17	-1.67	1.83	36.56
AGAG	-0.21	-0.27	3.04	-0.9	1.84	38.19
CTCT	0.21	-0.27	3.04	0.9	1.84	38.19
AGTA	-0.09	-1.05	3.31	4.1	-5.9	40.9
TACT	0.09	-1.05	3.31	-4.1	-5.9	40.9

ATGA	-0.32	-0.76	2.8	3	1.2	39.1
TCAT	0.32	-0.76	2.8	-3	1.2	39.1
CAAA	-0.01	-0.29	3.13	-1.62	3	35.48
TTTG	0.01	-0.29	3.13	1.62	3	35.48
CAAG	0.28	-0.67	2.96	-1.02	2.32	34.75
CTTG	-0.28	-0.67	2.96	1.02	2.32	34.75
CATG	0.03	-0.68	3.15	0.32	-3.84	32.28
CCTC	-0.21	0.78	3.52	0.3	14.8	32.4
GAGG	0.21	0.78	3.52	-0.3	14.8	32.4
CGCA	-0.14	-0.55	3.36	-1.13	3.6	36.61
TGCG	0.14	-0.55	3.36	1.13	3.6	36.61
CGCG	0.02	0.02	3.1	0.02	-2.54	36.77
CGGC	0.41	0.11	3.27	0.74	1.05	40.44
GCCG	-0.41	0.11	3.27	-0.74	1.05	40.44
CGGT	0.39	-0.32	3.28	1.8	4.2	38.8
ACCG	-0.39	-0.32	3.28	-1.8	4.2	38.8
CGTT	0.16	-0.8	3.39	4.27	-0.89	34.89
AACG	-0.16	-0.8	3.39	-4.27	-0.89	34.89
CTAC	-0.23	-0.07	3.21	-0.19	7.1	37.87
GTAG	0.23	-0.07	3.21	0.19	7.1	37.87
CTCC	-0.06	-0.43	3.24	1.2	-6.2	30.8
GGAG	0.06	-0.43	3.24	-1.2	-6.2	30.8
CTGG	-0.72	0.67	2.95	-1.7	6.6	36.5
CCAG	0.72	0.67	2.95	1.7	6.6	36.5
GAAA	-0.1	-0.7	3.27	-1.87	5.47	36.53
TTTC	0.1	-0.7	3.27	1.87	5.47	36.53
GAAG	0.04	-0.99	2.99	-0.77	-1.29	35.33
CTTC	-0.04	-0.99	2.99	0.77	-1.29	35.33
GAGA	0.09	-1.05	3.13	1.11	-4.27	32.43
TCTC	-0.09	-1.05	3.13	-1.11	-4.27	32.43
GAGT	-0.25	-1.05	3.08	-2.7	3.7	35.7
ACTC	0.25	-1.05	3.08	2.7	3.7	35.7
GCAA	0.06	-0.66	3.07	1.11	7.98	31.8
TTGC	-0.06	-0.66	3.07	-1.11	7.98	31.8
GCAT	-0.42	-0.14	3.17	0.45	5.16	37.29
ATGC	0.42	-0.14	3.17	-0.45	5.16	37.29
GCGC	0.74	-0.14	3.27	0.46	2.53	33.95
GCGG	-0.03	0.1	2.84	-0.35	2.1	40.6
CCGC	0.03	0.1	2.84	0.35	2.1	40.6
GCTA	-0.16	-0.55	3.12	2.47	-3.12	32.79
TAGC	0.16	-0.55	3.12	-2.47	-3.12	32.79
GCTG	-0.63	-0.17	3.56	1.13	8.83	34.79
CAGC	0.63	-0.17	3.56	-1.13	8.83	34.79
GGCA	0	-0.65	3.14	1.94	-1.54	35.97
TGCC	0	-0.65	3.14	-1.94	-1.54	35.97
GGCT	0.22	-0.02	3.52	1.79	0.61	39.04
AGCC	-0.22	-0.02	3.52	-1.79	0.61	39.04
GGTC	0.43	-1	3.19	1.4	-0.4	29.5

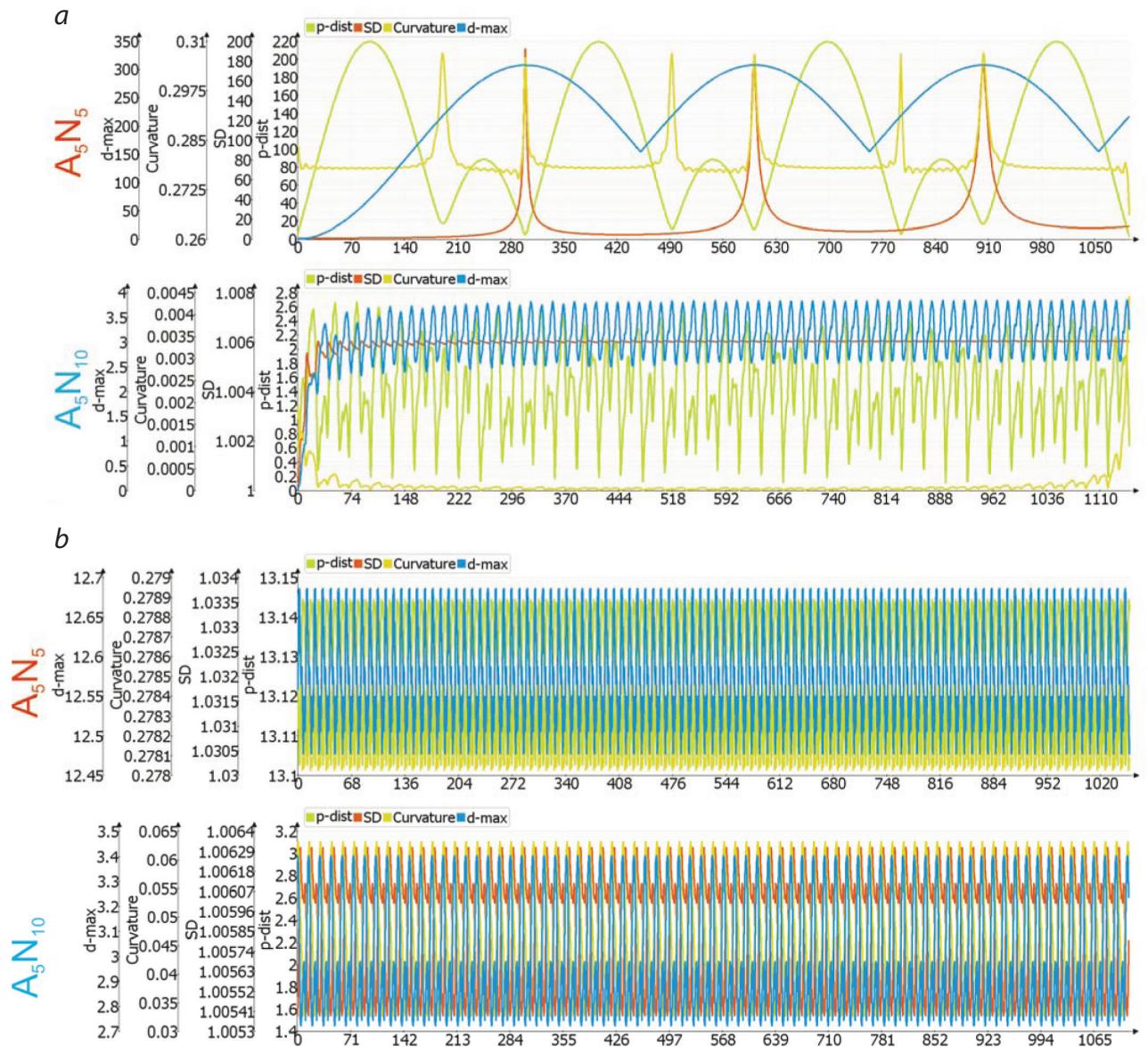
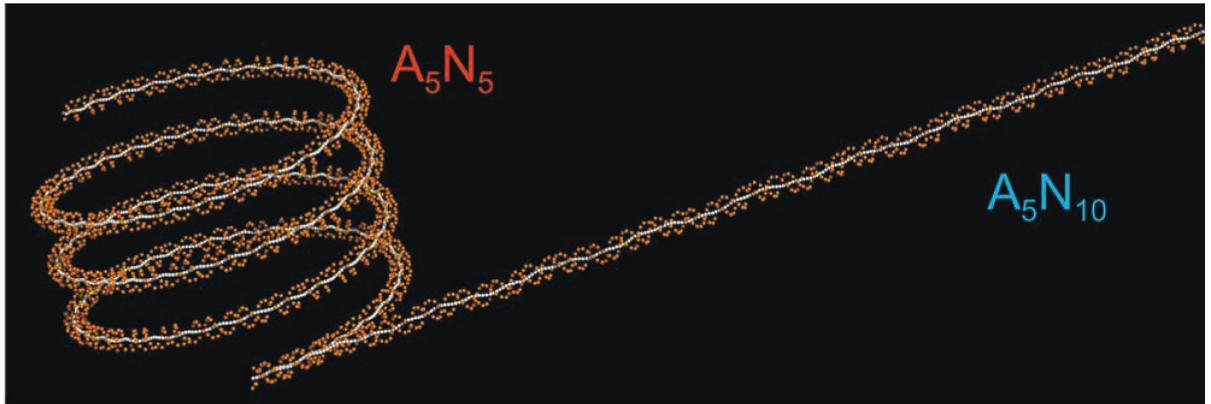
GACC	-0.43	-1	3.19	-1.4	-0.4	29.5
GGTG	-0.44	-0.54	3.26	-0.8	-4.2	34.8
CACC	0.44	-0.54	3.26	0.8	-4.2	34.8
GTAC	0.41	-0.53	2.98	4.3	8	39.7
GTCC	-0.26	0.18	3.47	-2.9	5.3	42.1
GGAC	0.26	0.18	3.47	2.9	5.3	42.1
GTCG	-0.16	-0.05	3.26	-1.25	4.99	37.11
CGAC	0.16	-0.05	3.26	1.25	4.99	37.11
GTCT	0.03	-0.61	3.03	0.7	5.7	36.5
AGAC	-0.03	-0.61	3.03	-0.7	5.7	36.5
GTGT	0.43	-0.19	3.22	3.7	3.7	40.8
ACAC	-0.43	-0.19	3.22	-3.7	3.7	40.8
GTTT	0.5	-0.5	3.32	1.12	0.87	37.53
AAAC	-0.5	-0.5	3.32	-1.12	0.87	37.53
TACG	0.14	-0.53	3.28	-1.53	0.95	33.24
CGTA	-0.14	-0.53	3.28	1.53	0.95	33.24
TCCG	-0.58	0.42	3.01	-2.4	-2.2	36.4
CGGA	0.58	0.42	3.01	2.4	-2.2	36.4
TCGG	-0.12	0.42	4.03	0.6	0.6	32.1
CCGA	0.12	0.42	4.03	-0.6	0.6	32.1
TCTG	-0.36	-0.26	3.47	1.4	-0.4	37.5
CAGA	0.36	-0.26	3.47	-1.4	-0.4	37.5
TGAG	0.15	-0.82	3.11	3	3.4	39.2
CTCA	-0.15	-0.82	3.11	-3	3.4	39.2
TGCT	-0.04	-0.72	3.49	-1.62	0.07	38.81
AGCA	0.04	-0.72	3.49	1.62	0.07	38.81
TGGT	0.85	0.1	3.46	0.8	8.3	36.3
ACCA	-0.85	0.1	3.46	-0.8	8.3	36.3
TGTC	0.28	-0.1	3.27	0.1	-1.8	35.6
GACA	-0.28	-0.1	3.27	-0.1	-1.8	35.6
TTAA	0.44	-0.71	3.78	-0.22	9.44	37.19
TTTA	0.34	-1.24	3.12	3.89	-2.98	27.94
TAAA	-0.34	-1.24	3.12	-3.89	-2.98	27.94
TTTT	0.24	-0.82	3.22	0.59	-1.71	36.91
AAAA	-0.24	-0.82	3.22	-0.59	-1.71	36.91

Three-dimensional DNA structures analyzed in the present study were modeled based on algebra of unit quaternions using equation (1) and visualized by the PyMOL Molecular Graphics System, Version 1.7.2 Schrödinger, LLC.

### Techniques of the DNA curvature distribution analysis

The distribution of  $C$ ,  $SD$  ( $Qc$ ) and  $p$ -*dist* can be analyzed in “local” and “global” scales using sliding window and fixed frame techniques, respectively. Fixed frame can be translated through DNA molecule like to a sliding window of very large size, but measurements within this system are performed differently. The sliding window technique assumes a single measurement at a given position of boundaries of the window. In contrast, within a fixed frame, the series of measurements are carried out along the DNA fragment at a fixed position of one or both boundaries of the frame. Thus, the measurements within sliding window characterize the DNA segments in scale of window, whereas measurements within fixed frame characterize the analyzed features in the current position, relative to the boundaries of frame. For example, the distribution of  $p$ -*dist* within the fixed frame represents a set of perpendicular distances from the current position on a trajectory (within frame) to the straight line connecting the ends of frame, over all positions. Distribution of  $SD$  ( $Qc$ ) and  $d$ -*max* is estimated for the DNA segments from a fixed start position (left boundary) to current position over the entire frame.

In order to demonstrate the differences between these techniques,  $(A_5N_5)_n$  (curved DNA) and  $(A_5N_{10})_n$  (straight DNA) were modeled and the distribution of such features of the DNA molecule as *curvature*,  $p$ -*dist*,  $d$ -*max* and  $SD$  were analyzed in sliding window and fixed frame (**Figure S1**). It was shown that using the sliding window technique the distributions of estimating features for both, curved and straight DNA fragments are very similar, and fluctuations of their values are insignificant (they are much smaller than base pair distances), indicating the uniformity of these distributions along the DNA molecules. Conversely, analysis of the same features in fixed frame clearly distinguishes the space organization of the  $(A_5N_5)_n$  and  $(A_5N_{10})_n$  DNA fragments (see **Figure S1**).



**Figure S1.** Distribution of *curvature*, *SD*, *p-dist* and *d-max* for the  $((A_5N_5)_n)$  and  $((A_5N_{10})_n)$  DNA fragments, calculated using fixed frame (a) and sliding window (b) techniques.

### Radius of gyration

A widely used measure of the chain dimension (size and shape) is the radius of gyration ( $R_G$ ) of the molecule about its center of gravity. Its second moment around the center of the mass of the chain ( $R_G^2$ ) is defined as the mean square distance of the collection of segments from their common center of mass ( $r_c$ ):

$$R_G^2 = \frac{1}{N} \sum_{i=1}^N (r_i - r_c)^2, \quad r_c = \frac{1}{N} \sum_{i=1}^N r_i,$$

where  $r_i$  represents the vector position of the  $i$ -th base pair center and  $N$  is the number of base pair. It is assumed that base pair have the same mass and are connected by massless bonds.

As the  $R_G^2$  is half of the square distance between two monomers on the chain, the  $R_G^2$  can be obtained avoiding the first calculation of  $r_c$ :

$$R_G^2 = \frac{1}{N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^N (r_i - r_j)^2,$$

where  $r_i - r_j$  is the length of the vector connecting segments  $i$  and  $j$ .

The gyration tensor is defined as:

$$T = \begin{bmatrix} T_x^2 & T_{xy} & T_{xz} \\ T_{xy} & T_y^2 & T_{yz} \\ T_{xz} & T_{yz} & T_z^2 \end{bmatrix} \xrightarrow{\text{diagonalization}} \begin{bmatrix} \lambda_x^2 & 0 & 0 \\ 0 & \lambda_y^2 & 0 \\ 0 & 0 & \lambda_z^2 \end{bmatrix},$$

where

$$T_{mn} = \frac{1}{N} \sum_{i=1}^N (r_m^i - r_{c_m}^i)(r_n^i - r_{c_n}^i)$$

or avoiding calculation of  $r_c$ :

$$T_{mn} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=i+1}^{N-1} (r_m^i - r_m^j)(r_n^i - r_n^j).$$

The eigenvalues of the matrix  $T$  denoted by  $\lambda_x^2$ ,  $\lambda_y^2$  and  $\lambda_z^2$  are called the principal moments of the gyration tensor and are squares of the three principal radii of gyration (orthogonal components in the system of principal axes of gyration, the variances of the coordinates along the principal axes directions), which are chosen such that the eigenvalues of  $T$  are sorted in the order  $\lambda_x^2 \leq \lambda_y^2 \leq \lambda_z^2$ .

### Persistence length

From the tangent correlation function (equation (9)) it follows that if  $L=P$  the  $\langle \cos(\theta) \rangle = e^{-1}$  and  $-\ln\langle \cos(\theta) \rangle = 1$ . In this sense the persistence length is the length of the polymer over which the average deflection of the polymer axis is one radian ( $\text{acos}(e^{-1})$ ). For this reason, it is likely that in some cases when  $L \gg P$ , there are segments with  $-\ln\langle \cos(\theta) \rangle > 1$ , the persistence length can be taken as the length of a segment, which corresponds to the point  $e^{-1}$  or 1 in distribution of  $\langle \cos(\theta) \rangle$  or  $-\ln\langle \cos(\theta) \rangle$ , respectively, or derived from the intersection of distribution of  $P$ , for different segment lengths, with diagonal.

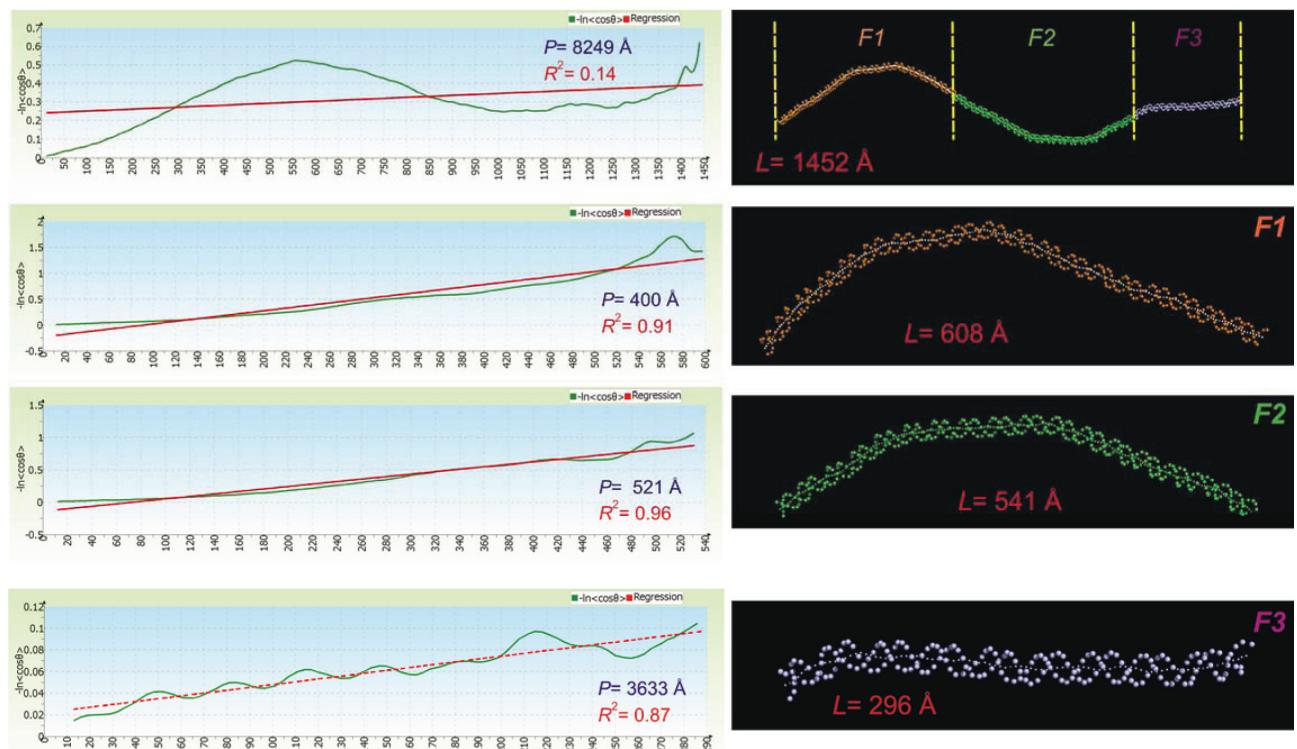
If  $l \ll P$  and  $\theta \ll \pi/2$  then:

$$\begin{cases} \cos(\theta) \approx 1 - \frac{\theta^2}{2}, & e^{-l/P} \approx 1 - \frac{l}{P} \\ \langle \theta^2 \rangle \approx \frac{2l}{P} \end{cases},$$

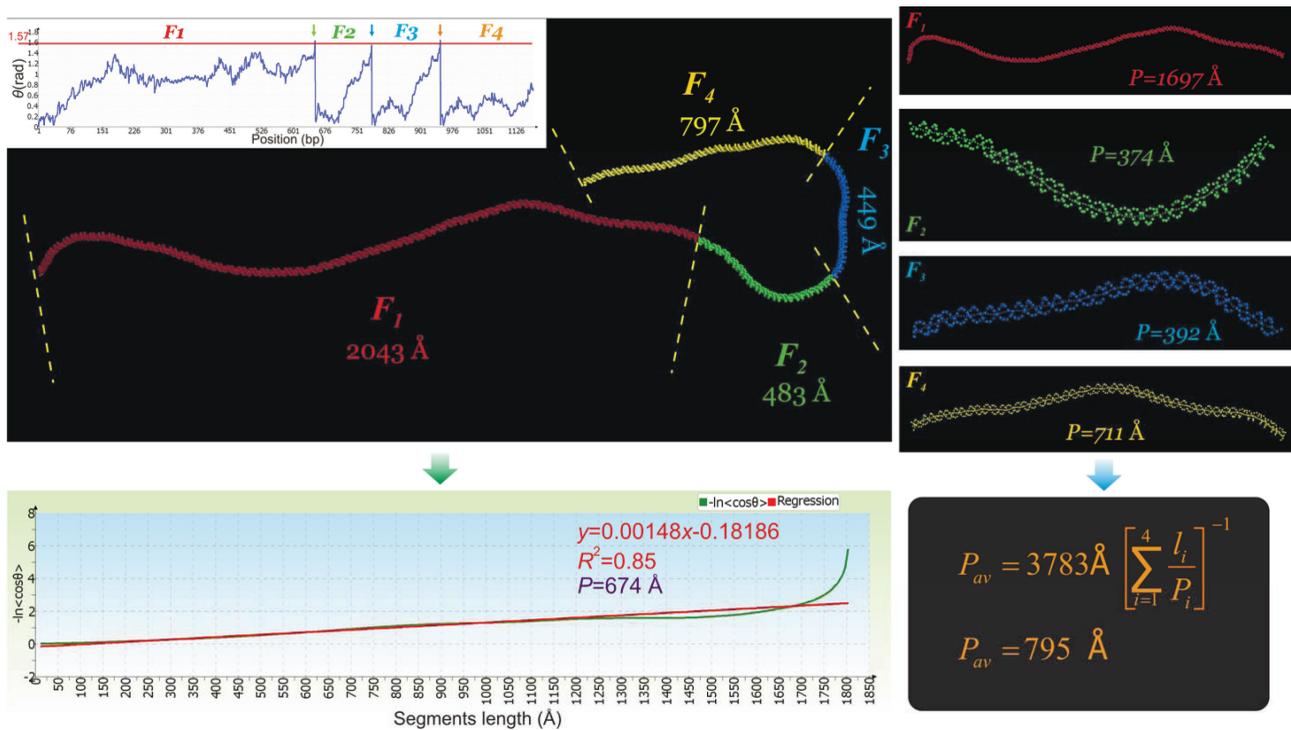
where  $\theta$  is estimated in radians.

Thus, persistence length  $P$  can be obtained from the distribution of  $\langle \theta^2 \rangle$  versus  $2l$  (over all  $l$  and along the entire chain) as the inverse value to the tangent of the slope of the linear regression.

It should be remembered that any global (over all chain) averaging of  $P$  for the heterogeneous chains will increase the errors of local measurements, performed for separate segments with the different intrinsic curvature and flexibility. For this reason, the distributions averaged over all lengths of segments and along all chain should be used with caution.



**Figure S2.** Estimation of the static persistence length of the PPD-A1 promoter region PCR-fragment (GenBank: KF834265) based on reconstructed DNA helical path. The deflection angle  $\theta$  between tangents to the start position of the fragment and point on DNA trajectory, along the entire fragment is less than  $90^\circ$ . The molecule was divided into three fragments (F1–F3) and  $P$  was estimated for each fragment using the orientational correlation function, from the inverse of the regression slope of the  $-\ln\langle \cos(\theta) \rangle$  distribution, over all segment lengths and along the fragment. The coefficient of determination ( $R^2$ ) for linear regression of  $-\ln\langle \cos(\theta) \rangle$  distribution for each fragment is indicated. Deflection angle was estimated between vectors tangent to trajectory of central axis and spaced from 1 to  $L-2$  bp, with step of 1 bp. The average persistence length ( $P_{av}$ ) evaluated from equation (10) is 550 Å, while this estimated from equation (8) is 1103 Å.

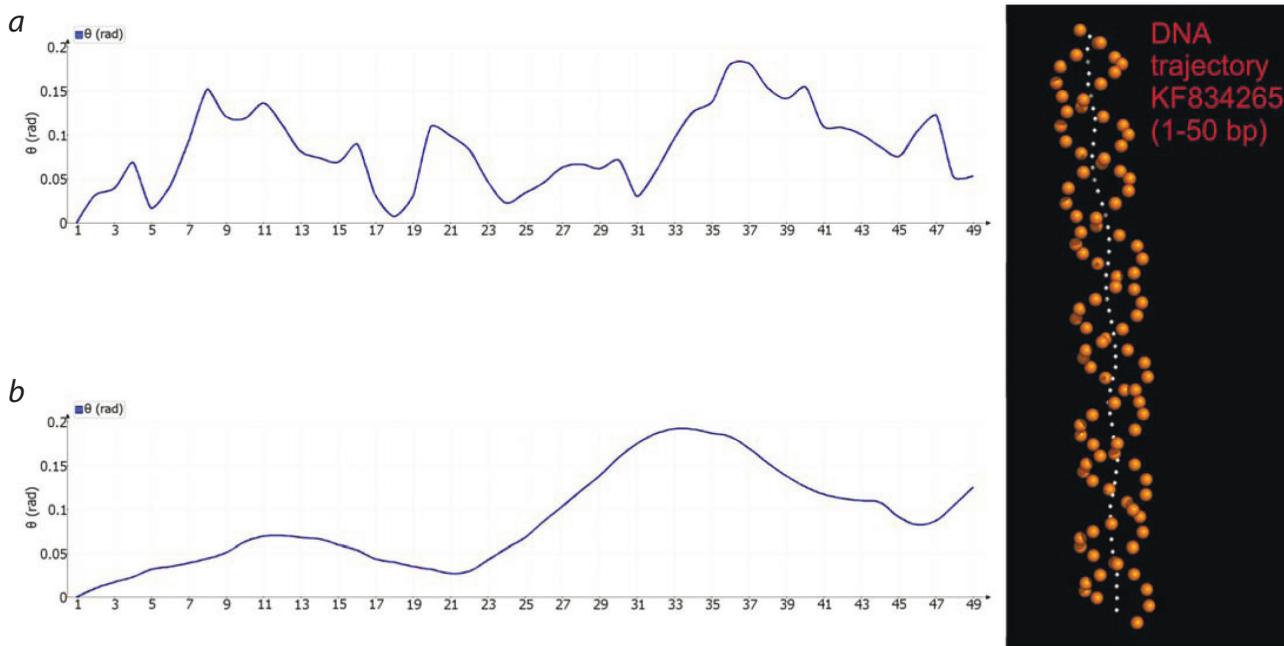


**Figure S3.** Estimation of the static persistence length of DNA molecule of the kinetoplast minicircle of *Trypanosoma vivax* (GenBank: KM386506) based on reconstructed DNA helical path. The molecule was divided into four fragments ( $F_1$ – $F_4$ ) according to the distribution of the deflection angle  $\theta$  between tangents to the start position of the fragment and point on the DNA trajectory, while  $\theta < 90^\circ$ .  $P$  was estimated for each fragment based on the magnitude of projection vector (equation (8)). Computation of the  $-\ln\langle\cos(\theta)\rangle$  distribution for the whole molecule is completed when the length of segment is about 1800 Å. This indicates that there are segments with length  $>1800$  Å with the deflection angle between the directions of ends  $\geq 90^\circ$ . Thus, the persistence length calculated for the whole molecule ( $P = 674$  Å), as for a heterogeneous chain, will be limited by the length of the most curved segments. Localization of these segments does not affect the result, because  $-\ln\langle\cos(\theta)\rangle$  at the given length of the segment calculated from the average cosine  $\theta$  over all positions of the chain. The value of  $P$  estimated for whole molecule using equation (13) is 748 Å.

Such non-Gaussian deformations as kinking make it impossible to use the WLC equations that are valid only for Gaussian chains. In present review the simple solution of this problem has been proposed. According to it's the DNA molecule can be divided into a set of fragments and estimate persistence length for each of these fragments separately. These fragments are statistically independent, so they can be distinguished by the change of distribution function of  $\langle\cos(\theta)\rangle$  ( $-\ln\langle\cos(\theta)\rangle$ ), or magnitude of projection vector with increasing length of segments.

For example, the length of the fragments can be expanded, while distribution of  $\langle\cos(\theta)\rangle$  decay exponentially ( $-\ln\langle\cos(\theta)\rangle$  has a good linear fit), then the next fragment begins. Further, the average persistence length can be estimated from equation (10) based on values of  $P$  each of these fragments (**Figure S2**). However, it should be remembered that use of the orientational correlation function for estimation of persistence length should be justified by a good linear fit of the  $-\ln\langle\cos(\theta)\rangle$  distribution with coefficient of determination  $R^2 \geq 0.9$ .

Since the computations of both the tangent-tangent correlation function and the magnitude of the projection vector is completed if  $\theta \geq 90^\circ$  (there is no  $\ln\langle\cos(\theta)\rangle$  and magnitude of projection  $\leq 0$ ) in some cases the different fragments can be distinguished based on this feature ( $\theta \geq 90^\circ$ ). For example, the **Figure S3** represents estimation of persistence length of a DNA molecule separated into four fragments with angular deflection of the double helix between ends of fragments close to  $90^\circ$ . The orientational correlation function can not be applied to estimation of  $P$  of the  $F_1$ ,  $F_3$  and  $F_4$  fragments, due to low values of  $R^2$  (0.57, 0.69 and 0.55 respectively).



**Figure S4.** Angular deflection between the tangents to the start position and each point on DNA trajectory of 50 bp length (fragment of KF834265), with step of 1 bp. (a) The tangents are unit vectors tangent to the trajectory of central axis (indicate the base pair directions). (b) The tangents are unit vectors tangent to the surface of molecule (normalized helical turn centroids).

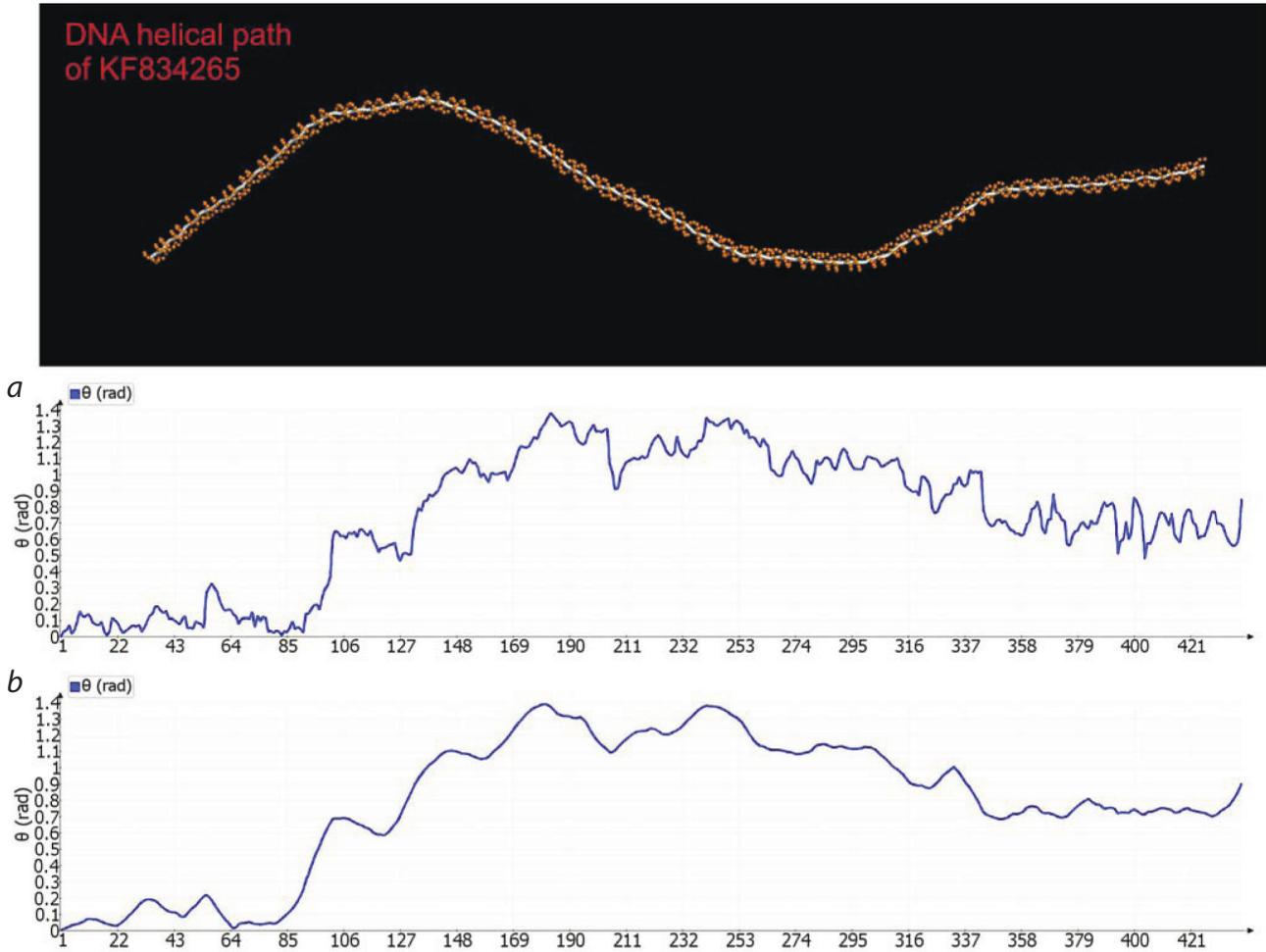
For this reason, the persistence length is estimated based on the magnitude of projection vector for all four fragments (equation (8)) and the average  $P$  of the DNA molecule calculated according to equation (10).

Other, more difficult methods for estimation of the persistence length of bent DNA can be found in (Schellman, Harvey, 1995; Rivetti et al., 1998; Anselmi et al., 2005).

As a rule, the contribution of the intrinsic curvature to  $P_a$  is much less than the contribution of thermal fluctuations (Vologodskaja, Vologodskii, 2002). However, this conclusion is valid under the condition that DNA fragments do not contain intrinsic curvature related sequences such as A-tracts (Geggier, Vologodskii, 2010). While flexibility of base pair steps in overall correlates well with local bends, determined by *roll* and *tilt* (Olson et al., 1998; Packer et al., 2000), those involved in A-tracts are relatively rigid (Sherer et al., 1999; Strahs, Schlick, 2000; Macdonald et al., 2001; McConnell, Beveridge, 2001; Nikolova et al., 2012). This leads to the fact that contribution of intrinsic DNA bends in measuring the value of  $P$  increases considerably, when the sequence contains motifs that are associated with substantial intrinsic curvature, such as A-tracts. Unfortunately, in the course of DNA shape analysis we cannot clearly distinguish the contribution of A-tracts and general sequence to changes in the deflection of DNA axes. So care must be taken when interpreting the results of *in silico* persistence length analysis of the A-tract enriched DNA molecules.

The values of persistence length strongly depend on the scale of measurements. The figures below demonstrate distribution of the axial deflection along the molecule, measured as the angles between the base pair directions (unit vectors tangent to the trajectory of central axis) and as the angles between the helical turn centroids (unit vectors tangent to the surface of molecule), respectively. It is shown that although the molecule is almost straight with average deflection between the adjacent centroids 0.005–0.013 rad the trajectory of the central axis is spiral, with an average deflection angle of 0.05 rad (**Figure S4**).

It is likely that use of tangent to the surface of molecule will be useful in division of DNA on the statistically independent fragments with different persistence length as well as in positioning of large static deformations along a molecule. Furthermore using of centroids is more preferable in calculations related to the spatial orga-



**Figure S5.** Distribution of the deflection angle  $\theta$  between tangents to the start position of the reconstructed DNA molecule of KF834265 and point on DNA trajectory. (a) The tangents are unit vectors tangent to the trajectory of central axis. (b) The tangents are unit vectors tangent to the surface of molecule (normalized helical turn centroids).

nization of the DNA molecules in large scale, because significant reduce noise, is determined by fluctuations of the individual base pair steps conformation parameters (**Figure S5**).

It is assumed that the average bending rigidity (apparent persistence length) of a fragment can be calculated from a set of specific values of persistence lengths assigning to each of 10 distinct base pair steps (Geggier, Vologodskii, 2010):

$$\frac{1}{P} = \sum_{XY} v_{XY} \frac{1}{P^{XY}},$$

where  $v_{XY}$  is the fraction of the dinucleotide step  $XY$  in the fragment. This allows us to estimate of apparent persistence length of molecule directly from DNA sequence. However, it is important to note that this approach was not tested on DNA fragments containing intrinsic bends intentionally (Geggier, Vologodskii, 2010), so that we cannot guarantee its validity for molecules carrying A-tracts or other intrinsic curvature related sequences. The differential DNA bendability along the sequence can be predicted based on flexibility, rigidity and bendability parameters as well as double helix thermodynamic stability in di-, three- and tetra-nucleotide context (Olson et al., 1998; Packer et al., 2000; Gabrielian, Pongor, 1996; Anselmi et al., 2000)

## DFT analysis

DFT analysis was previously applied in the phasing of DNA bendability and prediction of macroscopic curvature (Gabrielian, Pongor, 1996), in analysis of A-tract periodicity (Tolstorukov et al., 2005) and analysis of hydroxyl radical cleavage data related to local bends of DNA (Price, Tullius, 1993). Sliding-window DFT analysis of A-tracts and other nucleotide sequences are associated with DNA curvature was carried out for more than 1000 prokaryotic and eukaryotic genomes (Mrazek, 2009, 2010; Mrazek et al., 2011; Abel, Mrazek, 2012).

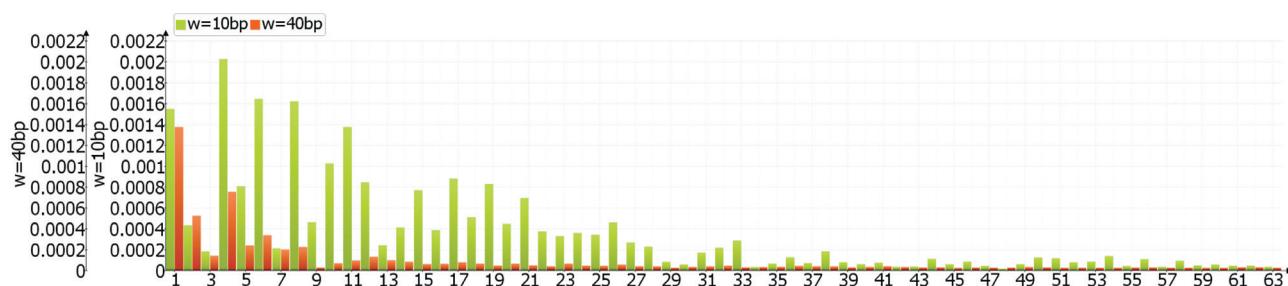
The computing of the discrete Fourier transform is high complexity and takes a lot of time. For this reason the set of data point should be minimized. In overall, the size of dataset up to 128 will be enough to clearly detect the change of curvature below  $1/42 \text{ bp}^{-1}$  ( $f=3$ ) at  $\Delta f=1$ , while for identification of the more rare systematic bends, preserving the dataset size, the resolution of DFT spectrum should be decreased through increase of the step of sliding window (decrease of  $fs$ ).

With increasing resolution of the curvature measurements (decreasing size of sliding window) the magnitudes of DFT power spectrum also increases, so that useful frequency domains can be clearly identified. However this leads to increase of high frequency noise in signal and DFT power spectral. In order to preserve the high resolution and in the same time reduce the high frequency noise the low-pass filter (LPF) can be applied.

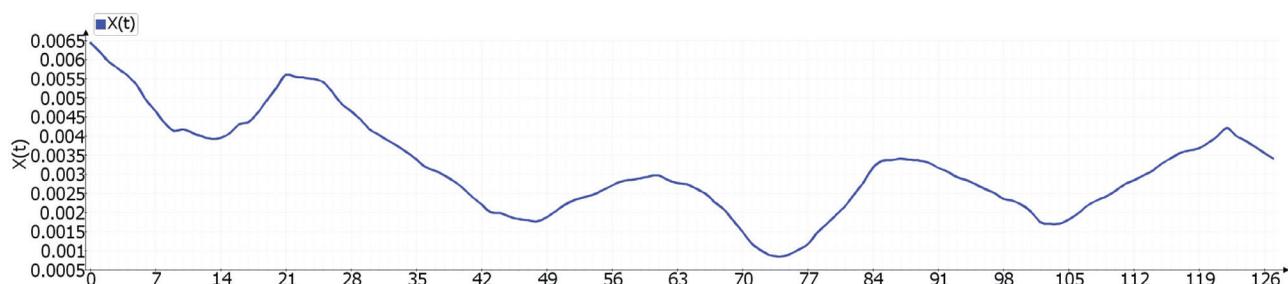
As a rule, local bends during a helical turn, as well as the beds between neighboring turns of DNA helix with general nucleotide sequence are mutually compensated. For the most cases the use of LPF that cut off frequency bins more than  $N/10$  (one helical turn) is recommended. The calculation of DFT can be completed upon reaching this frequency. Furthermore, unnecessary frequencies can be filtered by set to 0 (real and imaginary components) of the corresponding frequency bins.

The figures below demonstrate the use of low-pass filter in DFT analysis of curvature distribution along 128 bp of the kinetoplast minicircle of *Trypanosoma vivax* (GenBank: KM386506).

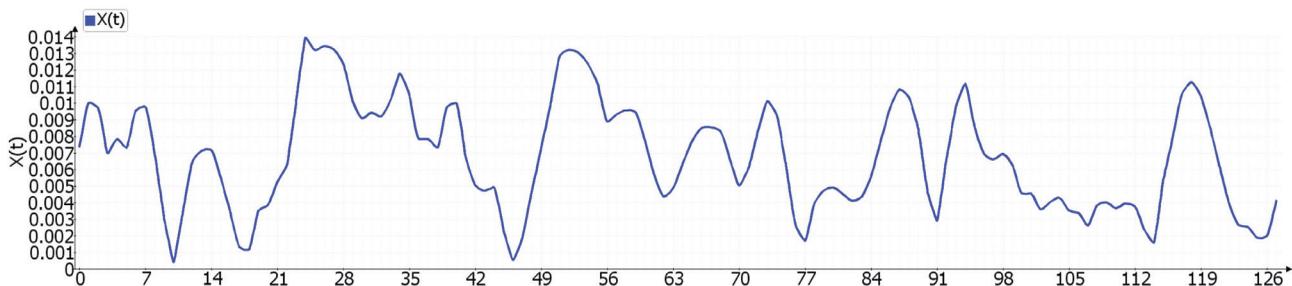
### DFT power spectral of curvature measured in sliding window of 10 and 40 bp sizes:



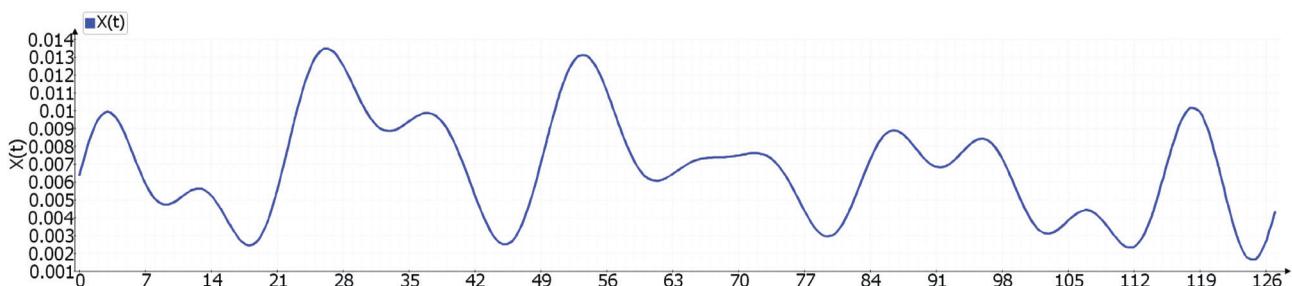
### Curvature signal for $w = 40 \text{ bp}$ :



### Curvature signal for $w = 10$ bp:



### Curvature signal for $w = 10$ bp and LPF for frequencies $> 1/10$ bp $^{-1}$ :



## References

- Abel J., Mrazek J. (2012) Differences in DNA curvature-related sequence periodicity between prokaryotic chromosomes and phages, and relationship to chromosomal prophage content. *BMC Genomics*. 13:188.
- Anselmi C., Bocchinfuso G., De Santis P., Savino M., Scipioni A. (2000) A theoretical model for the prediction of sequence-dependent nucleosome thermodynamic stability. *Biophys. J.* 79(2):601-613.
- Anselmi C., DeSantis P., Scipioni A. (2005) Nanoscale mechanical and dynamical properties of DNA single molecules. *Biophys. Chem.* 113(3):209-221.
- 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(3):281-287.
- Beveridge D.L., Barreiro G., Byun K.S., Case D.A., Cheatham T.E. 3rd, Dixit S.B., Giudice E., Lankas F., Lavery R., Maddocks J.H., Osman R., Seibert E., Sklenar H., Stoll G., Thayer K.M., Varnai P., Young M.A. (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps. *Biophys. J.* 87(6): 3799-3813.
- Beveridge D.L., Cheatham T.E., Mezei M. (2012) The ABCs of molecular dynamics simulations on B-DNA, circa 2012. *J. Biosci.* 37(3):379-397.
- Bi X., Yu Q., Sandmeier J.J., Zou Y. (2004) Formation of boundaries of transcriptionally silent chromatin by nucleosome-excluding structures. *Mol. Cell. Biol.* 24(5):2118-2131.
- 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(6):2312-2316.
- Bolshoy A., Nevo E. (2000) Ecologic genomics of DNA: upstream bending in prokaryotic promoters. *Genome Res.* 10:1185-1193.
- Brukner I., Sánchez R., Suck D., Pongor S. (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J.* 14(8):1812-1828.
- Caddle M.S., Dailey L., Heintz N.H. (1990) RIP60, a mammalian origin-binding protein, enhances DNA bending near the dihydrofolate reductase origin of replication. *Mol. Cell. Biol.* 10(12):6236-6243.
- Calladine C.R., Drew H.R. (1986) Principles of sequence-dependent flexure of DNA. *J. Mol. Biol.* 192(4):907-918.
- Calladine C.R., Drew H.R., McCall M.J. (1988) The intrinsic curvature of DNA in solution. *J. Mol. Biol.* 201(1):127-137.
- 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(39):9269-9273.
- Dickerson R.E., Chiu T.K. (1997) Helix bending as a factor in protein/DNA recognition. *Biopolymers.* 44(4):361-403.
- Dixit S.B., Beveridge D.L., Case D.A., Cheatham T.E. 3rd, Giudice E., Lankas F., Lavery R., Maddocks J.H., Osman R., Sklenar H., Thayer K.M., Varnai P. (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: Sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.* 89(6):3721-3740.

- Fujii S., Kono H., Takenaka S., Go N., Sarai A. (2007) Sequence-dependent DNA deformability studied using molecular dynamics simulations. *Nucleic Acids Res.* 35(18):6063-6074.
- Gabrielian A., Pongor S. (1996) Correlation of intrinsic DNA curvature with DNA property periodicity. *FEBS Lett.* 393(1):65-68.
- Geggier S., Vologodskii A. (2010) Sequence dependence of DNA bending rigidity. *Proc. Natl. Acad. Sci. USA.* 107(35):15421-15426.
- Gimenes F., Takeda K.I., Fiorini A., Gouveia F.S., Fernandez M.A. (2008) Intrinsically bent DNA in replication origins and gene promoters. *Genet. Mol. Res.* 7(2):549-558.
- Gorin A.A., Zhurkin V.B., Olson W.K. (1995) B-DNA twisting correlates with base-pair morphology. *J. Mol. Biol.* 247(1):34-48.
- Hertz G.Z., Young M.R., Mertz J.E. (1987) The A+T-rich sequence of the simian virus 40 origin is essential for replication and is involved in bending of the viral DNA. *J. Virol.* 61(7):2322-2325.
- Herzel H., Weiss O., Trifonov E.N. (1998) Sequence periodicity in complete genomes of archaea suggests positive supercoiling. *J. Biomol. Struct. Dyn.* 16:341-345.
- Hizver J., Rozenberg H., Frolow F., Rabinovich D., Shakked Z. (2001) DNA bending by an adenine-thymine tract and its role in gene regulation. *Proc. Natl. Acad. Sci. USA.* 98(15):8490-8495.
- Iyer V., Struhl K. (1995) Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* 14(11):2570-2579.
- Koch K.A., Thiele D.J. (1999) Functional analysis of a homopolymeric (dA-dT) element that provides nucleosomal access to yeast and mammalian transcription factors. *J. Biol. Chem.* 274(34):23752-23760.
- Kozobay-Avraham L., Hosid S., Bolshoy A. (2006) Involvement of DNA curvature in intergenic regions of prokaryotes. *Nucleic Acids Res.* 34:2316-2327.
- Lavery R., Zakrzewska K., Beveridge D., Bishop T.C., Case D.A., Cheatham T. 3rd, Dixit S., Jayaram B., Lankas F., Loughton C., Maddocks J.H., Michon A., Osman R., Orozco M., Perez A., Singh T., Spackova N., Sponer J. (2010) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.* 38(1):299-313.
- Liu Y., Beveridge D.L. (2001) A refined prediction method for gel retardation of DNA oligonucleotides from dinucleotide step parameters: reconciliation of DNA bending models with crystal structure data. *J. Biomol. Struct. Dyn.* 18(4):505-526.
- Macdonald D., Herbert K., Zhang X., Polgruto T., Lu P. (2001) Solution structure of an A-tract DNA bend. *J. Mol. Biol.* 306(5):1081-1098.
- McConnell K.J., Beveridge D.L. (2001) Molecular dynamics simulations of B'-DNA: sequence effects on A-tract-induced bending and flexibility. *J. Mol. Biol.* 314:23-40.
- Milot E., Belmaaza A., Wallenburg J.C., Gusew N., Bradley W.E., Chartrand P. (1992) Chromosomal illegitimate recombination in mammalian cells is associated with intrinsically bent DNA elements. *EMBO J.* 11(13):5063-5070.
- Mrazek J. (2009) Phylogenetic signals in DNA composition: limitations and prospects. *Mol. Biol. Evol.* 26(5):1163-1169.
- Mrazek J. (2010) Comparative analysis of sequence periodicity among prokaryotic genomes points to differences in nucleoid structure and a relationship to gene expression. *J. Bacteriol.* 192(14):3763-3772.
- Mrazek J., Chaudhari T., Basu A. (2011) PerPlot & PerScan: tools for analysis of DNA curvature-related periodicity in genomic nucleotide sequences. *Microb. Inform. Exp.* 1(1):13.
- Nickerson C.A., Achberger E.C. (1995) Role of curved DNA in binding of Escherichia coli RNA polymerase to promoters. *J. Bacteriol.* 177(20):5756-5761.
- Nikolova E.N., Bascom G.D., Andricioaei I., Al-Hashimi H.M. (2012) Probing sequence-specific DNA flexibility in A-tracts and pyrimidine-purine steps by nuclear magnetic resonance <sup>13</sup>C relaxation and molecular dynamics simulations. *Biochem.* 51(43):8654-8664.
- Olson W.K., Gorin A.A., Lu X.J., Hock L.M., Zhurkin V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA.* 95(19):11163-11168.
- Packer M.J., Dauncey M.P., Hunter C.A. (2000) Sequence-dependent DNA structure: tetranucleotide conformational maps. *J. Mol. Biol.* 295(1):85-103.
- Parker S.C., Hansen L., Aabaan H.O., Tullius T.D., Margulies E.H. (2009) Local DNA topography correlates with functional noncoding regions of the human genome. *Science.* 324(5925):389-392.
- Pérez-Martín J., de Lorenzo V. (1997) Clues and consequences of DNA bending in transcription. *Annu. Rev. Microbiol.* 51:593-628.
- Price M.A., Tullius T.D. (1993) How the structure of an adenine tract depends on sequence context: a new model for the structure of TnAn DNA sequences. *Biochemistry.* 32(1):127-136.
- Raveh-Sadka T., Levo M., Shabi U., Shany B., Keren L., Lotan-Pompan M., Zeevi D., Sharon D., Weinberger A., Segal E. (2012) Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Gen.* 44(7):743-750.
- Rees W.A., Keller R.W., Vesenska J.P., Yang G., Bustamante C. (1993) Evidence of DNA bending in transcription complexes imaged by scanning force microscopy. *Science.* 260(5114):1646-1649.
- Rivetti C., Walker C., Bustamante C. (1998) Polymer chain statistics and conformational analysis of DNA molecules with bends or sections of different flexibility. *J. Mol. Biol.* 280(1):41-59.
- Rohs R., West S.M., Sosinsky A., Liu P., Mann R.S., Honig B. (2009) The role of DNA shape in protein-DNA recognition. *Nature.* 461:1248-1253.
- Satchwell S.C., Drew H.R., Travers A.A. (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* 191(4):659-675.
- Schellman J.A., Harvey S.C. (1995) Static contributions to the persistence length of DNA and dynamic contributions to DNA curvature. *Biophys. Chem.* 55(1-2):95-114.
- Schieg P., Herzel H. (2004) Periodicities of 10–11 bp as indicators of the supercoiled state of genomic DNA. *J. Mol. Biol.* 343:891-901.

- Schultes N.P., Szostak J.W. (1991) A poly(dA. dT) tract is a component of the recombination initiation site at the ARG4 locus in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 11(1):322-328.
- Segal E., Fondufe-Mittendorf Y., Chen L., Thastrom A., Field Y., Moore I.K., Wang J.P., Widom J. (2006) A genomic code for nucleosome positioning. *Nature.* 442:772-778.
- Segal E., Widom J. (2009) Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.* 19(1):65-71.
- Sherer E.C., Harris S.A., Soliva R., Orozco M., Laughton C.A. (1999) Molecular dynamics studies of DNA A-tract structure and flexibility. *J. Am. Chem. Soc.* 121:5981-5991.
- Shi Q., Thresher R., Sancar A., Griffith J. (1992) Electron microscopic study of (A)BC excinuclease. DNA is sharply bent in the UvrB-DNA complex. *J. Mol. Biol.* 226(2):425-432.
- Straus D., Schlick T. (2000) A-tract bending: insights into experimental structures by computational models. *J. Mol. Biol.* 301(3):643-663.
- Struhl K., Segal E. (2013) Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* 20(3):267-273.
- Tolstorukov M.Y., Virnik K.M., Adhya S., Zhurkin V.B. (2005) A-tract clusters may facilitate DNA packaging in bacterial nucleoid. *Nucleic Acids Res.* 33(12):3907-3918.
- Tong H., Mrázek J. (2014) Investigating the interplay between nucleoid-associated proteins, DNA curvature, and CRISPR elements using comparative genomics. *PLoS One.* 9(3):e90940.
- Ulanovsky L.E., Trifonov E.N. (1987) Estimation of wedge components in curved DNA. *Nature.* 326:720-722.
- Ulyanov N.B., James T.L. (1995) Statistical analysis of DNA duplex structural features. *Methods Enzymol.* 261:90-120.
- Vologodskaja M., Vologodskii A. (2002) Contribution of the intrinsic curvature to measured DNA persistence length. *J. Mol. Biol.* 317(2):205-213.
- Yonetani Y., Kono H. (2009) Sequence dependencies of DNA deformability and hydration in the minor groove. *Biophys. J.* 97(4):1138-1147.
- Zahn K., Blattner F.R. (1985) Sequence-induced DNA curvature at the bacteriophage lambda origin of replication. *Nature.* 317(6036):451-453.