DNA crookedness regulates DNA mechanical properties at short length scales

A. Marín-González, J. G. Vilhena, Fernando Moreno-Herrero,* and Rubén Pérez*

E-mail: fernando.moreno@cnb.csic.es; ruben.perez@uam.es

Supplementary Materials

Materials and Methods

1 Derivation of the model

We will denote the extension of the molecule by $x(0)$ and the distance between the base pair centers of the i^{th} step by $l_i(0)$. Then, from the definition of the crookedness, β, provided in the main text:

$$\cos\beta(0) \equiv \frac{x(0)}{\sum_{i=1}^{N-1} l_i(0)} \Longrightarrow x(0) = \cos\beta(0) \times \sum_{i=1}^{N-1} l_i(0), \tag{S1}$$

where the summatory is over the $N-1$ base pair steps. When a stretching force is exerted on the molecule, it will induce a change in the parameters $\cos\beta$ and l_i and, according to Eq.S1 a change in the total extension. This is then equal to

$$x(F) = \cos\beta(F) \times \sum_{i=1}^{N-1} l_i(F). \tag{S2}$$

We approximated $x(F)$ as a Taylor expansion of the function of N variables $x(l_i(F), \cos\beta(F))$ around the point $(l_i = l_i(0), \cos\beta = \cos\beta(0))$. This yields

$$\begin{aligned}
\Delta x(F) &= \sum_{i=1}^{N-1} \left(\frac{\partial x(l_j, \cos\beta)}{\partial l_i} \right)_{\substack{l_j = l_j(0) \\ \cos\beta = \cos\beta(0)}} \Delta l_i(F) + \left(\frac{\partial x(l_j, \cos\beta)}{\partial(\cos\beta)} \right)_{\substack{l_j = l_j(0) \\ \cos\beta = \cos\beta_0}} \Delta\cos\beta(F) \\
&= \sum_{i=1}^{N-1} \cos\beta(0)\ \Delta l_i(F) + \left(\sum_{i=1}^{N-1} l_i(0) \right)\ \Delta\cos\beta(F) \\
&\equiv \sum_{i=1}^{N-1} \Delta x_{l,i}(F) + \Delta x_\beta(F).
\end{aligned} \tag{S3}$$

We checked that this approximation holds for our simulated molecules in the range of forces studied (1 - 20 pN) see Fig.S6.

Notably, Eq.S3 shows that we are decomposing the molecule elongation, $\Delta x(F)$, as a sum of $N-1$ *local* contributions coming from elongating individual base pair steps, $\Delta x_{l,i}(F)$; and

the *global* contribution of aligning the base pairs with the helical axis, $\Delta x_\beta(F)$. Assuming that these deformations are linear with the force (see below for justification) we can define their associated elastic constants as:

$$k_{l,i} \equiv \frac{x(0)F}{\Delta x_{l,i}(F)}, \tag{S4}$$

$$k_\beta \equiv \frac{x(0)F}{\Delta x_\beta(F)}, \tag{S5}$$

in analogy with the definition of the stretch modulus

$$S \equiv \frac{x(0)F}{\Delta x(F)}. \tag{S6}$$

Using these three definitions and Eq. S3 one can easily derive the following expression:

$$S^{-1} = \sum_{i=1}^{N-1} k_{l,i}^{-1} + k_\beta^{-1}. \tag{S7}$$

Therefore, we arrive at an expression where the stretching stiffness of a DNA molecule is determined by N parameters, $k_{l,i}$ ($i = 1, \ldots, N-1$) and k_β. In what follows we will show that these paramters can be univocally determined from DNA sequence and structure. In other words, Eq.S7 allows us to determine the stretch modulus of any given DNA sequence by solely *looking* at its equilibrium conformation. Figure 2c, main text is a schematic representation of Eq.S7, showing that the stretching response of a DNA molecule is being modelled as a set of N springs in series with elastic constants $k_{l,i}$ and k_β.

In the following we will elaborate on the relation between these elastic constants and DNA sequence and structure. Starting from the definition of $k_{l,i}$, Eq. S4, one arrives at

$$k_{l,i} \equiv \frac{x(0)F}{\Delta x_{l,i}(F)} = \frac{x(0)F}{\cos\beta(0)\ \Delta l_i(F)} = \left(\sum_{j=1}^{N-1} l_j(0)\right) \times \frac{F}{\Delta l_i(F)} \equiv \left(\sum_{j=1}^{N-1} l_j(0)\right) \times \tilde{k}_{l,i}, \tag{S8}$$

where we have used the definitions of $\Delta x_{l,i}(F)$ and $\cos\beta(0)$ given in Eq. S1 and Eq. S3

respectively. This equation illustrates that $k_{l,i}$ is nothing but the stiffness of separating two consecutive base pairs, $\tilde{k}_{l,i}$, multiplied by a prefactor. We may argue that this $\tilde{k}_{l,i}$ is sequence dependent, since it is closely related with the base pair stacking interactions. Moreover, for computing $\tilde{k}_{l,i}$ we will resort to the nearest neighbour approximation, that is, we will assume that this parameter is solely dependent on the base pairs composing the step. Accordingly, there will be ten different values of $\tilde{k}_{l,i}$, corresponding to the ten different dinucleotides or *step kinds*.

We computed $\tilde{k}_{l,i}$ for the ten *step kinds* from our constant forces simulations. To that end we analyzed the six 18-mer molecules with sequences CGCG$(\mathrm{NN})_5$CGCG, where NN = AA, AC, AG, AT, CG and GG. Note that each of the ten *step kinds* is present at least four times in this set of sequences. We first computed the mean base pair step separation for each step kind at different forces, $l_i(F)$. This separation was obtained using the 3DNA software as $l = \sqrt{Slide^2 + Shift^2 + Rise^2}$.[1] Then we plotted $\Delta l_i(F)$ as a function of the force taking the F = 1 pN simulation as reference, see Fig S7. These data sets showed a linear dependence with the force, supporting the assumption made above that this deformation is elastic in this range of forces . Following Eq. S8, the values of $\tilde{k}_{l,i}$ are obtained as the inverse of the slopes of the linear fits to these datasets. Knowing the $\tilde{k}_{l,i}$ for all step kinds, one only needs to measure the sum of base pair distances at zero force, $\sum_j^{N-1} l_j(0)$, to obtain $k_{l,i}$.

In Table S1 we show the computed values of $\tilde{k}_{l,i}$. As anticipated, this parameter is highly dependent on the step kind. In particular, a closer inspection at Table S1 reveals that the AT step is the stiffest, a result which is in line with previous works which coincide in that this step is the least flexible.[2–4] In contrast, the GA, CG and CA steps showed the smallest $\tilde{k}_{l,i}$ values. This result is in agreement with a study on DNA crystal structures,[3] where these three steps showed the highest standard deviation of *Rise*. Moreover, we reproduced the

[0]Only the GG step showed a larger dispersion from the linear response. This is probably due to convergence issues, since this step has an unusually high value of the slide parameter, which is highly variable and strongly affects the value of l_i. We extended our simulations to 2 μs and ran additional simulations at 12 and 18 pN and this deviation from linearity persisted. It is unlikely that we will achieve convergence for this step in the μs timescale.

tendency of *pyrimidine-purine* steps being generally the most flexible, followed by *purine-purine* and *purine-pyrimidine*.[4,5] Interestingly, this same trend was recently reported for the dinucleotides stacking energies, where *purine-pyrimidine* interactions are mostly among the strongest and *pyrimidine-purine* among the weakest.[6]

From the MD simulations one can also determine the k_β of the six simulated sequences. In order to do so, we shall write

$$k_\beta \equiv \frac{x(0)F}{\Delta x_\beta(F)} = \frac{x(0)F}{\left(\sum_{i=1}^{N-1} l_i(0)\right)\ \Delta\cos\beta(F)} = \frac{F}{\Delta\cos\beta(F)/\cos\beta(0)}. \tag{S9}$$

The values of $\cos\beta(F)$ were computed for each molecule using the definition of the crookedness, Eq. S1, where the extension of the molecule was computed as the sum of the helical rises using the software 3DNA.[1] We then plotted $\Delta\cos\beta(F)/\cos\beta(0)$ as a function of the force taking the F = 1 pN simulation as reference. This ratio showed a linear dependence on the force, validating the assumption that $\Delta x_\beta(F)$ is elastic. We fitted these data sets to linear functions and obtained the k_β for each molecule as the inverse of the slopes, according to Eq. S9. k_β is represented as a function of β in Figure 3, main text.

2 Simulation Details

DNA duplexes were built using the software NAB.[7] The sequence of interest, *SEQ*, was sandwiched between two CGCG handles, such that the entire molecule reads d(5'-CGCG(*SEQ*)CGCG-3'). The sequences were divided into two kinds: benchmark and testing sequences.

Molecules were neutralized with sodium counterions and no additional salt was added. The systems were then placed in a box of dimensions $\sim$ 85Å×85Å×85Å (depending on the number of base pairs of the molecule) filled with explicit water molecules. Systems were energy minimized in 5000 steps with restrains on the DNA followed by 5000 steps of unrestrained minimization. Then, the systems were heated up to 300 K and equilibrated

for 20 ns in the isobaric-isothermal (NPT) ensemble (P=1 atm, T=300 K). After NPT equilibration, and starting from the last configuration of the equilibration, five simulations were run for each sequence at constant forces of 1, 5, 10, 15 and 20 pN in the NVT ensemble following the protocol described in.[8] Constant forces simulations were extended up to times $t \gtrsim 1\mu$s. The CGCG handles were excluded for the data analysis.

We used the AMBER14 software suite[7] with NVIDIA GPU acceleration.[9–11] Parmbsc0[12] with the χ0L3 modification[13] of the Cornell ff99 force field[14] was used to describe DNA. Water was described using the TIP3P model,[15] while Joung/Cheatham parameters were used to describe the sodium counterions.[16,17] For the description of the deoxy-5-methylcitosine we used the Amber parameters derived in[18]. Periodic boundary conditions and Particle Mesh Ewald (with standard defaults and a real-space cutoff of 9Å) were used to account for long-range electrostatics interactions. Van der Waals contacts were truncated at the real space cutoff. SHAKE algorithm was used to constrain bonds containing hydrogen, thus allowing us to use an integration step of 2 fs. Coordinates were saved every 1000 steps.

Supplementary Figures

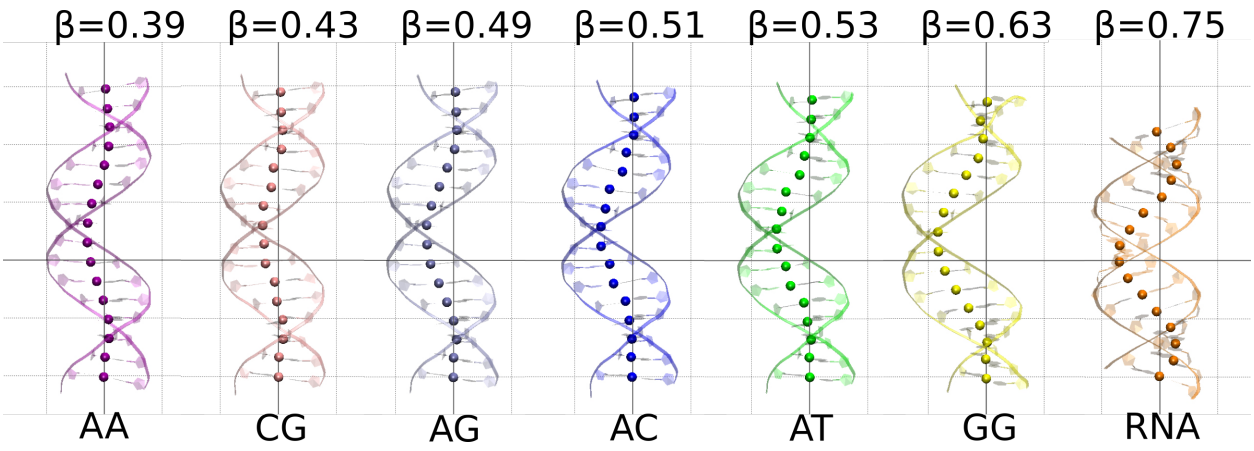

Figure S1: **Average structures of molecules with variable crookedness.** Average structures were computed over the 1 μs MD simulation at 1 pN force of the benchmark sequences CGCG$(NN)_5$CGCG with NN=AA, CG, AG, AC, AT and GG. The beads represent the centers of the base pairs. The terminal base pairs have been omitted in the representation. For comparison, a 16 bp's double-stranded RNA molecule is shown.

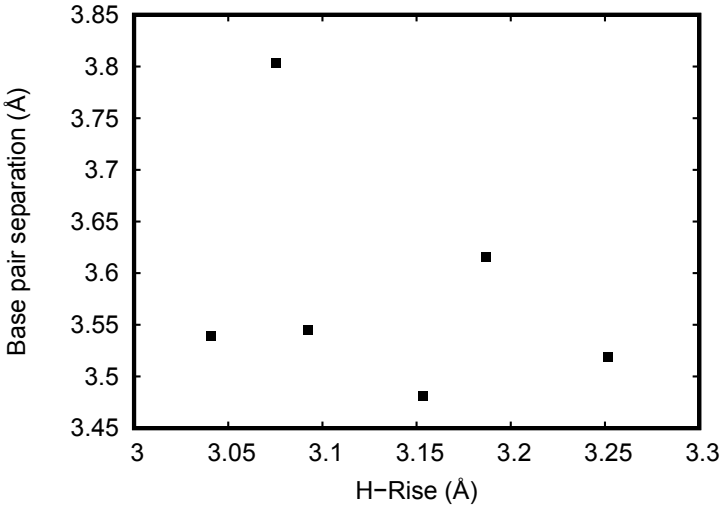

Figure S2: **Molecule extension and base pair seperation.** The base pair separation and helical rise per base pair step were computed using the 3DNA software[1] for our benchmark sequences at 1 pN force excluding the CGCG handles. Both quantities were averaged for the 1μs simulation time. Base pair step separations were computed using the formula[1] $l_i = \sqrt{Slide^2 + Shift^2 + Rise^2}$

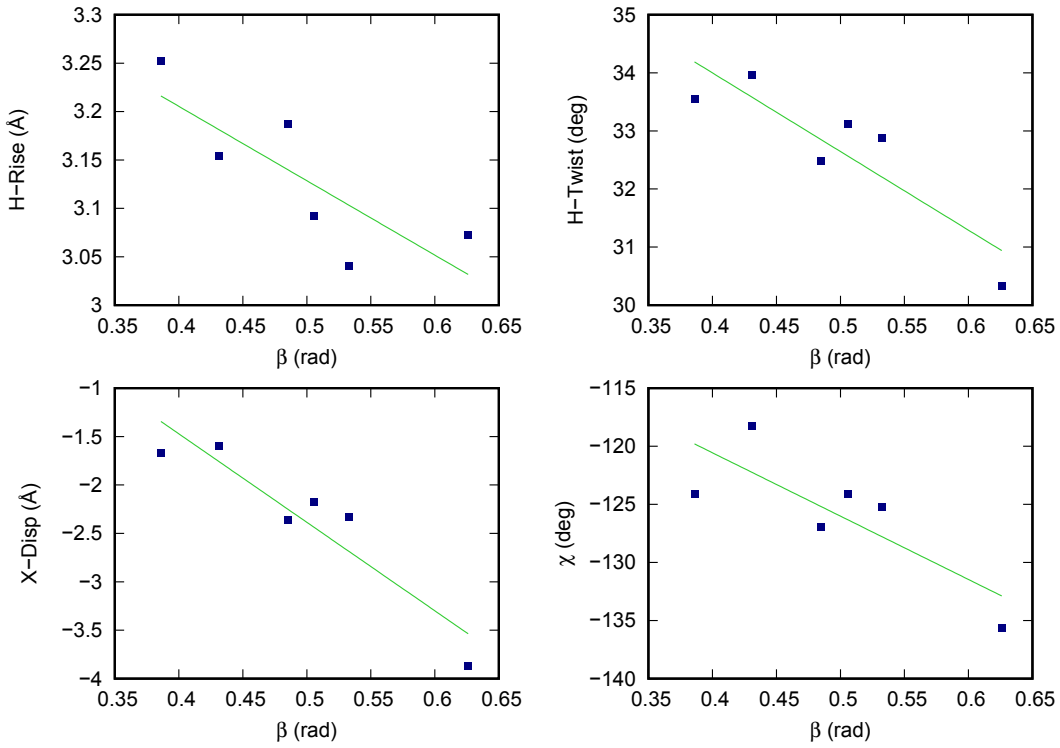

Figure S3: **Comparison of β with A/B DNA structural parameters.** A-form DNA is known to have a lower helical rise and helical twist than B-DNA in addition to a larger x-displacement in absolute value[19] and a lower glycosidic torsion angle, χ[20]. These four parameters were computed for our benchmark molecules using the software *3DNA*[1] and *cpptraj*[7] and averaged over the 1μs simulation time. These are represented as a function of the crookedness, β, in blue squares. As a guide to the eye we plotted the linear fit of these data sets, showing that β is anti-correlated with these parameters. Therefore, sequences with high β can be reasonably associated to DNA conformations close to the $A-$form and sequences with lower β can be associated to $B-$form conformations.

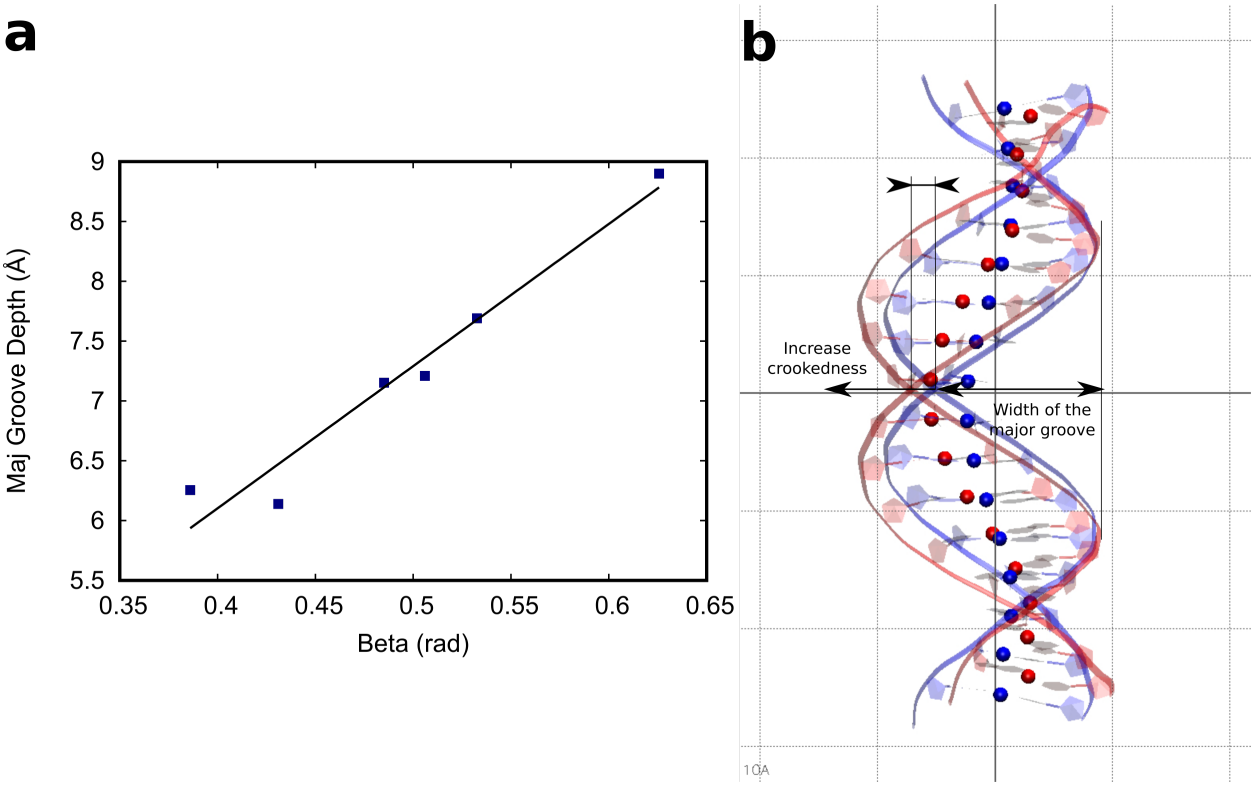

Figure S4: **Mechanism of increasing the crookedness by deepening the major groove. a,** The depth of the major groove was computed using the software Curves+ [21] and is represented as a function of the crookedness for the benchmark sequences. The data yields a correlation of r = 0.975. **b,** average structures of the poly-G (red) and the poly-A (blue) molecules to illustrate the structural relation between the crookedness curvature and the major groove depth. As usual, the color beads represent the crookedness deformation. As the crookedness increases, *i.e.* the beads deviate more from a straight line, the depth of the major groove increases. This highlights the close relation between the crookedness and a short-scale curvature of the DNA modulated by the major groove.

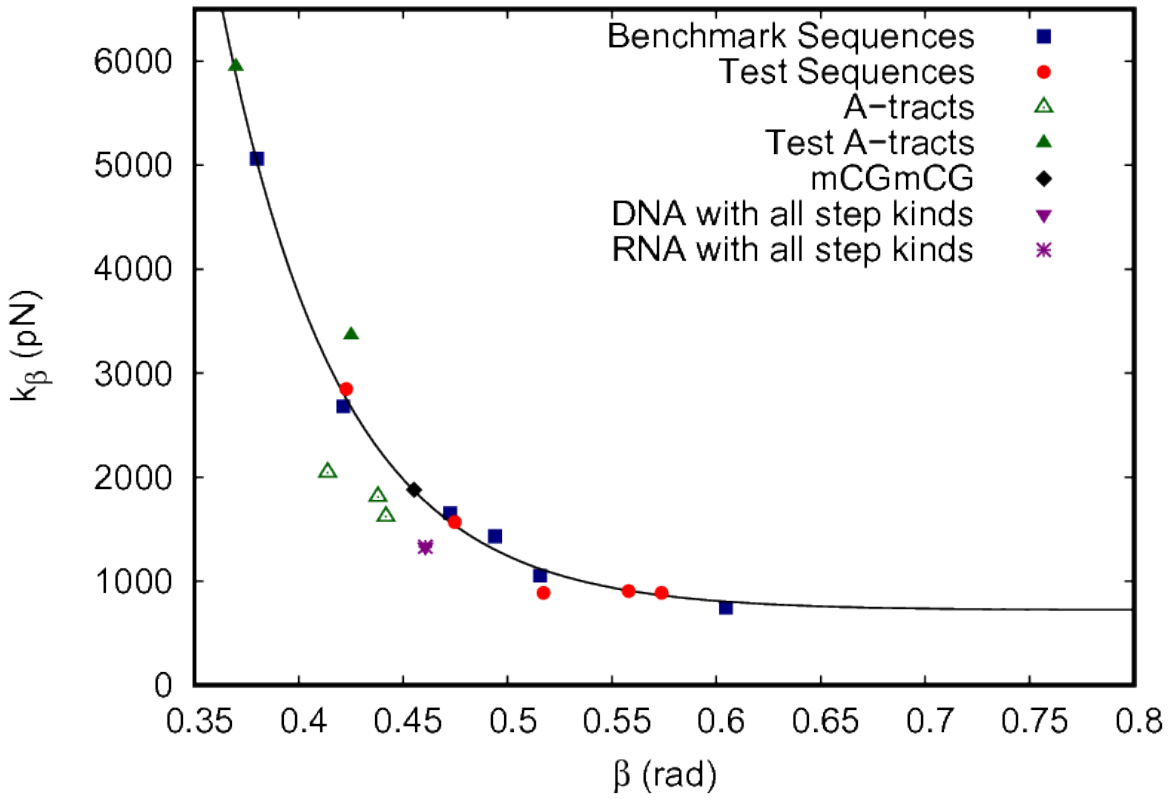

Figure S5: **Relation between β and k_β at 10pN force.** The values of k_β as computed in the main text were represented as a function of β values computed at 10 pN force. The black line represents a fit to the same function as the one used in the text. The one-to-one correspondence between these two parameters presented in the text is conserved when defining β at 10pN. Notice that the plot is slightly shifted towards higher β values as induced by the 10 pN force.

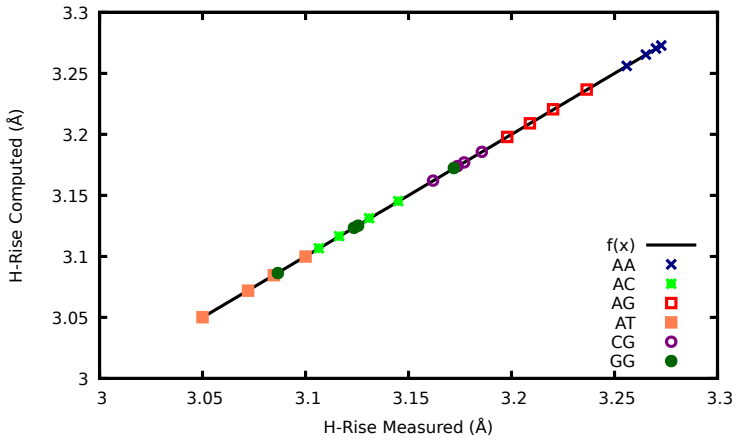

Figure S6: **Check of the validity of the first order Taylor expansion.** The values of the molecule extension at forces F = 5, 10, 15 and 20 pN were computed for our benchmark sequences using the formula derived from the Taylor expansion shown above, Eq. (S3). This was calculated from the values of $x(F), \cos\beta(F)$ and $l(F)$ obtained at different forces, taking the F = 1 pN value as reference.

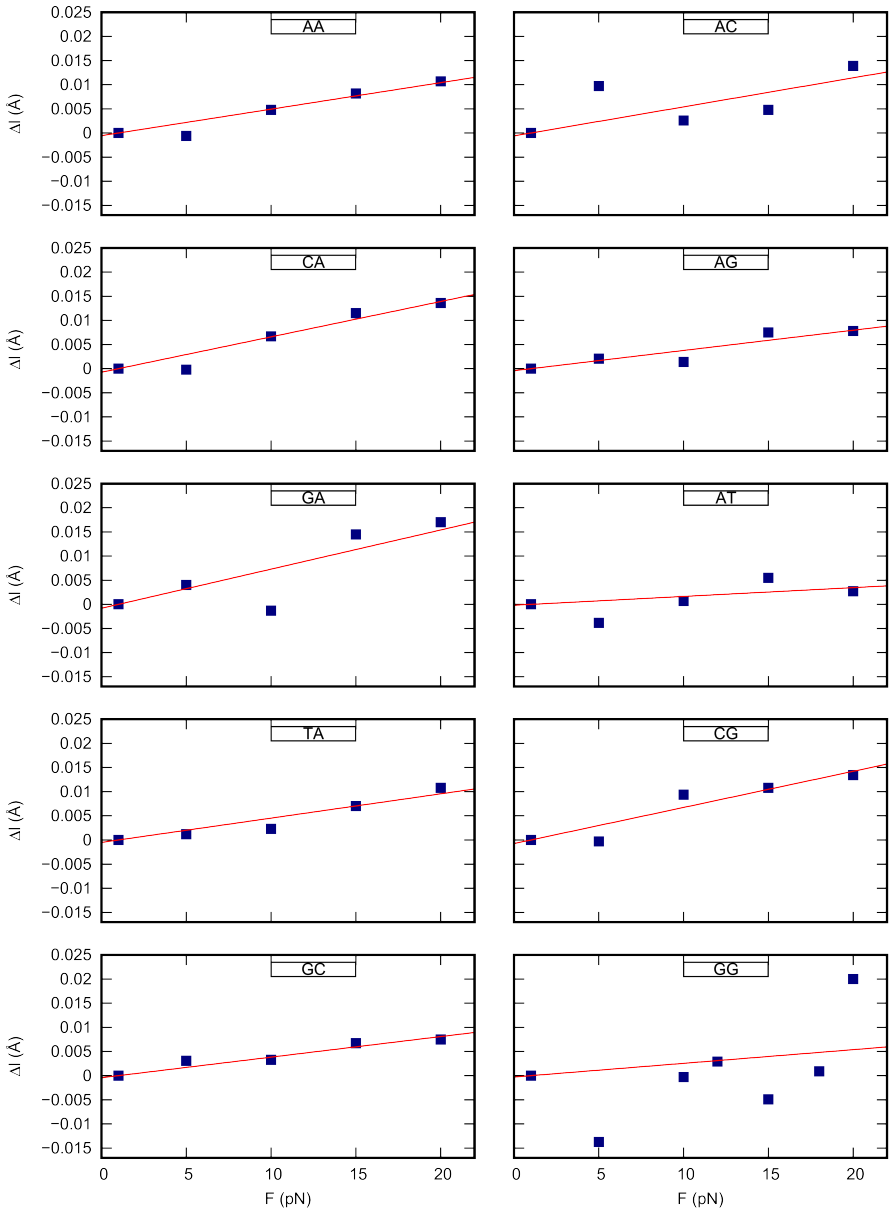

Figure S7: **Force-induced base pair step elongation of the ten dinucleotide step kinds.** The base pair step separation, $l_i(F)$ was computed for each step and averaged over the steps of the same kind and over the 1 μs simulation time for our benchmark sequences at each constant force simulation. We represented the elongation $\Delta l_i(F)$ with respect to the $l_i(0)$ value, taken at 1 pN force, as a function of the applied force. The data sets were fitted to a linear function constrained to go through the (1,0) point. The inverse of the slopes are the $\tilde{k}_{l,i}$ of each step kind given in Table S1.

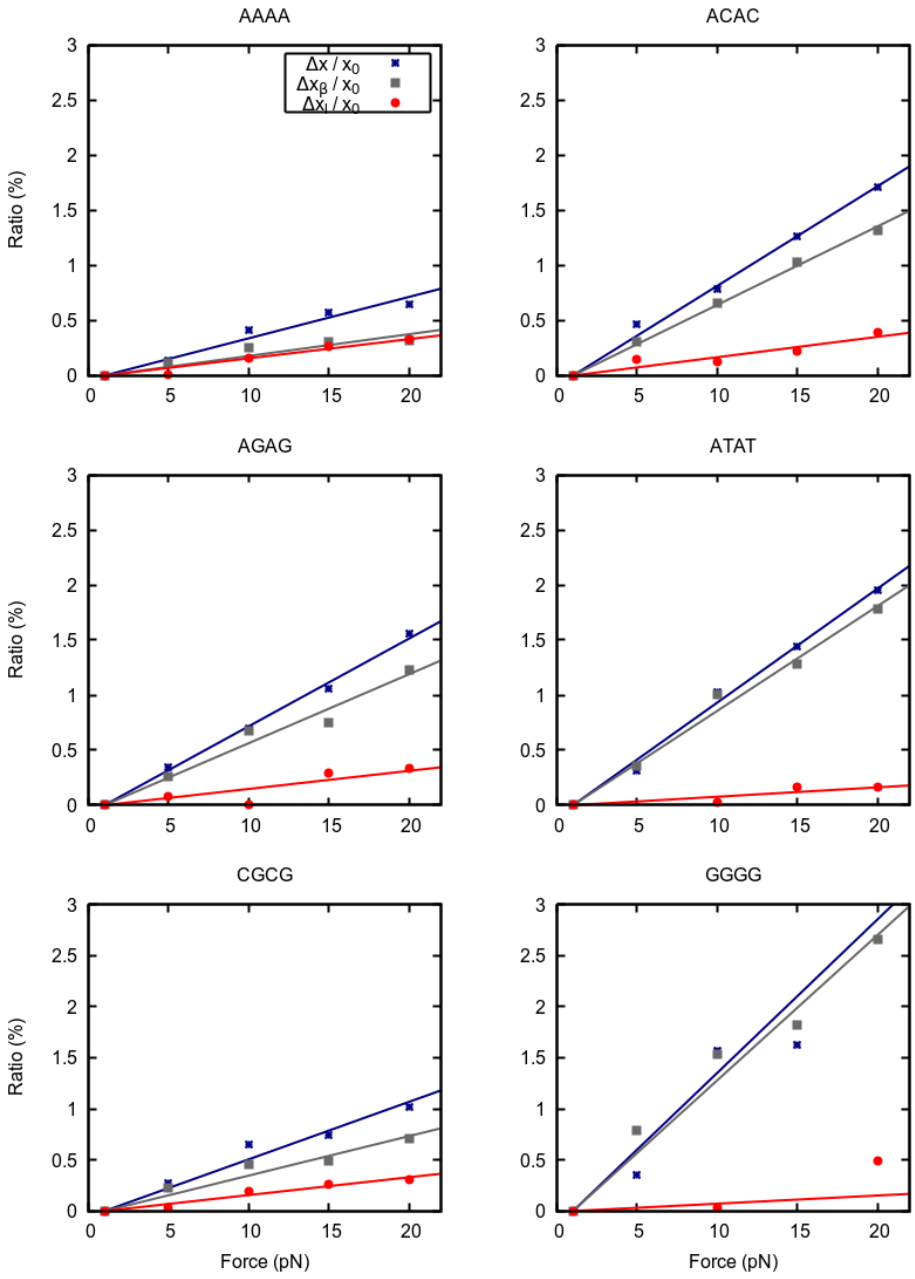

Figure S8: **Contributions to the force-induced elongation of the benchmark sequences.** The total elongation, Δx; the contribution to the elongation coming from aligning the base pairs with the helical axis, Δx_β; and the contribution coming from elongating base pair steps, $\Delta x_l \equiv (\sum \Delta x_{l,i})/N$ were computed from the MD simulations of our benchmark sequences. The ratio of these quantities and the molecule extension at 1 pN force, x_0 are represented as a function of the applied force. All the quantities were computed using the *3DNA software* in the same way as described in[8]. The inverse of the linear fits of $\Delta x(F)/x_0$ and $\Delta x_\beta(F)/x_0$ yield respectively the stretch modulus, S, and the crookedness flexibility, k_β, that are shown in Figure 3, main text.

Supplementary Tables

Table S1: The values of $\tilde{k}_{l,i}$ were computed for the ten different dinucleotide steps as the inverse of the slopes of the linear fits of Fig S7.

Step kind	AA	AC	CA	AG	GA	AT	TA	CG	GC	GG
$\tilde{k}_{l_i}$ (pN/Å)	1820	1660	1370	2390	1230	5500	1990	1340	2350	3500

Table S2: Analyzed sequences to build and test the model. All newly simulated sequences (that is all the molecules except DNA and RNA with all steps[8]) were sandwiched betweeen CGCG handles. mC stands for deoxy-5-methyl cytosine.

Benchmark Sequences			
Label	Name	Sequence	Ref
AA	poly-A	AAAAAAAAAA	
CG	poly-CG or CG Island	CGCGCGCGCG	
AG		AGAGAGAGAG	
AC		ACACACACAC	
AT		ATATATATAT	
GG	poly-G	GGGGGGGGGG	
Testing Sequences			
DDD	Drew-Dickerson Dodecamer	CGCGAATTCGCG	[22]
TATA	TATA-element	TATAAAAG	[23]
TFBS	Transcription Factor Binding Site	GGATGGGAG	[24]
G_4CG_4		GGGGCGGGG	
G_4AAG_4		GGGGAAGGGG	
A-tracts			
DUE	DNA Unwinding Element	GATCTATTTATTT	[25]
A_4TA_4		AAAATAAAA	
A_4GGA_4		AAAAGGAAAA	
Test A-tracts			
A_8T		AAAAAAAAT	
A_8GG		AAAAAAAAGG	
mCGmCG			
mCGmCG	Hypermethylated CG Island	mCGmCGmCGmCGmCG	
DNA with all step kinds			
DNA with all steps	DNA containing all step kinds	GCGCAATGGAGTACG	[8,26]
RNA with all step kinds			
RNA with all steps	RNA containing all step kinds	GCGCAAUGGAGUACG	[8,26]

Table S3: Comparison of the unusually high value of the stretch modulus of the unmethylated and hypermethylated poly-CG obtained from our simulations and measured in optical tweezers experiments[27]. Our value of S was computed from the force extension curves as described in[8].

Molecule	S (pN) from[27]	S (pN) this work
CG-Island	1828.5 (52.5)	1809 (70)
Hypermethylated CGI	1514.5 (66.3)	1345 (54)

References

(1) Lu, X.; Olson, W. K. *Nucleic Acids Research* **2003**, *31*, 5108.

(2) Lankaš, F.; Šponer, J.; Langowski, J.; Cheatham, T. E. *Biophysical Journal* **2003**, *85*, 2872–2883.

(3) Olson, W. K.; Gorin, A. A.; Lu, X.-J.; Hock, L. M.; Zhurkin, V. B. *Proceedings of the National Academy of Sciences* **1998**, *95*, 11163–11168.

(4) Fujii, S.; Kono, H.; Takenaka, S.; Go, N.; Sarai, A. *Nucleic Acids Research* **2007**, *35*, 6063–6074.

(5) K. Olson, W.; V. Colasanti, A.; Li, Y.; Ge, W.; Zheng, G.; Zhurkin, V. B. In *Computational studies of RNA and DNA*; Sponer, J., Lankas, F., Eds.

(6) Kilchherr, F.; Wachauf, C.; Pelz, B.; Rief, M.; Zacharias, M.; Dietz, H. *Science* **2016**, *353*.

(7) Case, D. et al. AMBER 14. 2014.

(8) Marín-González, A.; Vilhena, J.; Pérez, R.; Moreno-Herrero, F. *Proceedings of the National Academy of Sciences of the United States of America* **2017**, *114*, 7049–7054.

(9) Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. *J. Chem. Theory Comput.* **2013**, *9*, 3878–3888.

(10) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. *J. Chem. Theory Comput.* **2012**, *8*, 1542–1555.

(11) Grand, S. L.; Götz, A. W.; Walker, R. C. *Computer Physics Communications* **2013**, *184*, 374 – 380.

(12) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; III, T. E. C.; Laughton, C. A.; Orozco, M. *Biophysical Journal* **2007**, *92*, 3817 – 3829.

(13) Zgarbová, M.; Otyepka, M.; Sponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E.; Jurecka, P. *Journal of Chemical Theory and Computation* **2011**, *7*, 2886–2902, PMID: 21921995.

(14) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *Journal of the American Chemical Society* **1995**, *117*, 5179–5197.

(15) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(16) Joung, I. S.; Cheatham, T. E. *The Journal of Physical Chemistry B* **2009**, *113*, 13279–13290.

(17) Li, P.; Roberts, B. P.; Chakravorty, D. K.; Merz, K. M. *Journal of Chemical Theory and Computation* **2013**, *9*, 2733–2748.

(18) Lankaš, F.; Cheatham, T. E.; Špačáková, N.; Hobza, P.; Langowski, J.; Šponer, J. *Biophysical Journal* **2002**, *82*, 2592 – 2609.

(19) Bloomfield, V. A.; Crothers, D. M.; Tinoco Jr, I. *Nucleic Acids. Structures, Properties and Functions*; University Science Books, 2000.

(20) Waters, J. T.; Lu, X.-J.; Galindo-Murillo, R.; Gumbart, J. C.; Kim, H. D.; Cheatham, T. E.; Harvey, S. C. *The Journal of Physical Chemistry B* **2016**, *120*, 8449–8456.

(21) Lavery, R.; Moakher, M.; Maddocks, J. H.; Petkeviciute, D.; Zakrzewska, K. *Nucleic Acids Res* **2009**, *37*, 5917–5929.

(22) Wing, R.; Drew, H.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E. *Nature* **1980**, *287*, 755 EP –.

(23) Kim, J. L.; Nikolov, D. B.; Burley, S. K. *Nature* **1993**, *365*, 520–527.

(24) McCall, M.; Brown, T.; Hunter, W. N.; Kennard, O. *Nature* **1986**, *322*, 661–664.

(25) Bramhill, D.; Kornberg, A. *Cell* **1988**, *52*, 743 – 755.

(26) Liebl, K.; Drsata, T.; Lankas, F.; Lipfert, J.; Zacharias, M. *Nucleic Acids Research* **2015**, *43*, 10143–10156.

(27) Pongor, C. I.; Bianco, P.; Ferenczy, G.; Kellermayer, R.; Kellermayer, M. *Biophysical Journal* **2017**, *112*, 512 – 522.