UNIVERSITAT DE BARCELONA

Information Transfer and Dynamics of Nucleic Acids studied by Theoretical Approaches

Alexandra Balaceanu

ANNEX | SUPPLEMENTARY INFORMATION

2 Parmbsc1: a refined force field for DNA simulations

SUPPLEMENTARY DISCUSSION

QM data fitting.

As shown in the **Supplementary Table 12** refined parmbsc1 parameters fit very well high-level QM data. The *syn-anti* equilibrium, which was non-optimal in parmbsc0, is now well reproduced (**Supplementary Fig. 26**). The fitting to sugar puckering profile was improved by increasing the East barrier, and by displacing the North and South minima to more realistic regions (**Supplementary Table 12** and **Supplementary Fig. 27**). Additionally, parmbsc1 provides ε and ζ conformational map almost indistinguishable from the CCSD(T)/CBS results in solution (**Supplementary Fig. 28**), with errors in the estimates of relative BI/BII stability and transition barrier equal to 0.2 and 0.0 kcal mol^{-1} respectively.

Force-field benchmark simulations.

It is not our purpose here to perform a comprehensive comparison of parmbsc1 with previous force-fields. This would require the analysis of >100 structures with up to six other force-fields, clearly out of the scope of this work. We performed, however, a first critical evaluation of the most used force-fields using the well-known Drew Dickerson dodecamer as reference. We tested parmbsc0[1–3], parmbsc0-OL1[4] (ε and ζ corrections from Šponer's group), parmbsc0-OL4[5] (χ corrections), parmbsc0-OL1+OL4[4,5], CHARMM36[6], and a modified parmbsc0 developed by mixing corrected χ values and scaled-down van der Waals interactions (parmbsc0-CG, Cheng-Garcia)[7]. In all cases simulations were extended for at least 1 µs under identical simulation conditions. The value of this benchmark must not be overestimated, since different behavior may be found for other DNA sequences or conformations, but it can be useful to obtain an approximate idea of the range of error expected in parmbsc1 with respect to other modern force-fields. Results are summarized in **Supplementary Table 2** and **Supplementary Figs. 29–31**. All the force-fields are able to maintain the general B-like conformation in the central part of the duplex. However, significant distortions are found in the terminal pairs for parmbsc0, parmbsc0-OL1 (ε and ζ corrections), and CHARMM36, which show large openings (**Supplementary Fig. 29**) and very frequent fraying, with the formation of non-canonical interactions. The distortion induced by the opening of the terminal C-G pairs is especially dramatic in CHARMM36 simulations (**Supplementary Fig. 29**), but it is not negligible for parmbsc0[8] and parmbsc0-OL1, where aberrant *trans* Watson-Crick contacts involving a cytosine in *syn*, are dominant (**Supplementary Fig. 30**). It is clear that duplexes are flexible and reversible opening and closing of terminal base pair should exist, as found for example in parmbsc1 simulations (**Supplementary Fig. 30**). However, detailed analysis of new NMR spectra (**Supplementary Fig. 31**) shows that there are just minor differences between terminal and interior base pairs, which mean that open states should be short-lived, and not prevalent as in CHARMM36 simulations. Furthermore, no NMR evidence exists (**Supplementary Fig. 31**) supporting the existence of stable unusual contacts involving terminal pairs, or the prevalence of non-*anti* conformations, which are observed in parmbsc0, parmbsc0-OL1 or CHARMM36 simulations.

The introduction of χ corrections removes the excessive fraying of terminal pairs, preserving better the integrity of the entire helix in parmbsc1, parmbsc0-OL4[8], parmbsc0-CG (Cheng-Garcia, and parmbsc0-OL1+OL4 (ε, ζ, and χ corrections together) trajectories (**Supplementary Figs. 29** and **30**). The duplex sampled from parmbsc0-CG calculations is however far from the experimental structures: RMSd around 4 Å (compared to values clearly below 2.0 Å for parmbsc1 simulations), strong under-twisting, poor groove geometry and incorrect description of the BI/BII equilibrium

(**Supplementary Table 2**). The sequence dependence of the helical properties, which is clear for the rest of bsc0-based force-fields, is also lost here (**Supplementary Fig. 29**).

Parmbsc0-OL4 and parmbsc0-OL1+OL4 provide reasonable representations of the DDD geometry. However, the use of parmbsc1 leads to clear improvements in all structural descriptors. Thus, parmbsc1 balances better the sugar puckering (see **Supplementary Fig. 29**), leads to a better balance of BI/BII states (**Supplementary Table 2**), improves very significantly the average roll which is now very close to the NMR estimates, avoiding the excess of roll found in other calculations (**Supplementary Table 2** and **Supplementary Fig. 29**). Parmbsc1 improves very clearly the average twist and its sequence-dependence (RMSd difference between NMR and parmbsc1 twist profiles is 1.9 º, compared with 3.7 º for parmbsc1-OL1+OL4, or 5.6 º for CHARMM36. Not surprisingly, the improvement in twist, roll and puckering is reflected in much more realistic groove dimensions. For example the average difference in groove widths is only 0.3 Å between parmbsc1 and NMR values, while for the parmbsc0-OL1+OL4 force-field error is above 1 Å. In summary, at least for DDD, parmbsc1 provide results of better quality than those obtained with the most recent force-fields for DNA available.

The effect of ionic strength and the nature of counterion.

To evaluate potential differences in simulations arising from the ionic strength we performed additionally 2 µs simulations of DDD with extra salt: Na^+Cl^- 150 mM, and 500 mM. These additional calculations were performed using the same conditions outlined previously, showing results that are quite independent on the exact choice (in the 0–500 mM range) of the added extra salt (**Supplementary Fig. 25**).

SUPPLEMENTARY REFERENCES

1. Pérez, A. *et al. Biophys. J.***92,** 3817–3829 (2007).

2. Cornell, W.D. *et al. J. Am. Chem. Soc.***117,** 5179–5197 (1995).

3. Cheatham III, T.E., Cieplak, P. & Kollman, P.A. *J. Biomol. Struct. Dyn*.**16,** 845–862 (1999).

4. Zgarbová, M. *et al. J. Chem. Theory Comput.***9,** 2339–2354 (2013).

5. Krepl, M. *et al. J. Chem. Theory Comput.***8,** 2506–2520 (2012).

6. Hess, B., Kutzner, C., Van Der Spoel, D. & Lindahl, E. *J. Chem. Theory Comput.***4,** 435–447 (2008).

7. Cheng, A.A., Garcia, A.E. *Proc. Natl. Acad. Sci. USA***110**, 16820–25 (2013).

8. Zgarbová, M., Otyepka, M., Šponer, J., Lankaš, F. & Jurečka, P. *J. Chem. Theory Comput.***10,** 3177–3189 (2014).

SUPPLEMENTARY TABLES

Supplementary Table 1. DNA sequences used for validation of the parmbsc1 force-field. The nature of the structure, the origin of the starting conformation and the length of the production trajectories are also reported. The validation set is divided in several blocks separated in the table by double lines (from top to bottom): i) Normal B-DNA structures (including mismatches, epigenetic modifications and polymeric sequences); ii) very large oligomers; iii) Complexes of DNA with proteins or drugs; iv) Unusual DNA structures; v) dynamic transitions.; parmbsc1 validation; and vi) parmbsc1 benchmarking.

Sequence	Family	Origine / PDB id	Length (ns)
d(CGCGAATTCGCG)$_2$	B-DNA	1BNA, 1NAJ	1x 800 2x 1000 1x 12001x 10000
d(CCATACaATACGG)$_2$	B-DNA mismatch AA	Fiber	500
d(CCATACgATACGG)$_2$	B-DNA mismatch GG	Fiber	500
d(CGCGA*5mC*GTCGCG)$_2$	B-DNA 5methylC	Fiber	250
d(CGCGA*5hmC*GTCGCG)$_2$	B-DNA 5hydroxy-methylC	Fiber	250
d(CGCGT*5mC*GACGCG)$_2$	B-DNA 5methylC	Fiber	500
d(CGCGACGTCGCG)$_2$	B-DNA	Fiber	500
d(CGCGTCGACGCG)$_2$	B-DNA,	Fiber	500
d(GCCTATAAACGCCTATAA)$_2$	B-DNA	Fiber	1000
d(CTAGGTGGATGACTCATT)$_2$	B-DNA	Fiber	1000
d(CACGGAACCGGTTCCGTG)$_2$	B-DNA	Fiber	1000
d(GGCGCGCACCACGCGCGG)$_2$	B-DNA	Fiber	1000
d(GCCGAGCGAGCGAGCGGC)$_2$	B-DNA	Fiber	1000
d(GCCTAGCTAGCTAGCTGC)$_2$	B-DNA	Fiber	1000
d(GCTGCGTGCGTGCGTGGC)$_2$	B-DNA	Fiber	1000
d(GCGATCGATCGATCGAGC)$_2$	B-DNA	Fiber	1000
d(GCGAGGGAGGGAGGGAGC)$_2$	B-DNA	Fiber	1000
d(GCGCGGGCGGGCGGGCGC)$_2$	B-DNA	Fiber	1000
d(GCGGGGGGGGGGGGGGGC)$_2$	B-DNA	Fiber	1000
d(GCGTGGGTGGGTGGGTGC)$_2$	B-DNA	Fiber	1000
d(CTCGGCGCCATC)$_2$	B-DNA	2HKB	590
d(CCTCTGGTCTCC)$_2$	B-DNA	2K0V	590

d(CGCATGCTACGC)$_2$	B-DNA	2L8Q	590
d(GGATATATCC)$_2$	B-DNA	2LWG	590
d(GCGCATGCTACGCG)$_2$	B-DNA	2M2C	590
d(CCTCAGGCCTCC)$_2$	B-DNA	2NQ1	590
d(CGCGAAAAAACG)$_2$	B-DNA (A-track)	1D89	200
d(GGCAAAAAACGG)$_2$	B-DNA (A-track)	1FZX	200
d(GCAAAATTTTGC)$_2$	B-DNA (A-track)	1RVH	200
d(CTTTTAAAAG)$_2$	B-DNA (A-track)	1SK5	200
d(AGGGGCCCCT)$_2$	B-DNA (A-track)	440D	200
d(GGCAAGAAACGG)$_2$	B-DNA (A-track)	1G14	1000
d(CGATCGATCG)$_2$	B-DNA crystal	1D23	32x 2000
d(ATGGATCCATAGACCAGAACATGATGTTCTCA)$_2$	B-DNA 32mer	Fiber	1000
d(CGCGATTGCCTAACGAGTACTCGTTAGGCAATCGCG)$_2$	B-DNA 36mer	Fiber	2x 300
d(CGCGATTGCCTAACGGACAGGCATAGACGTCTATGCCTGTC CGTTAGGCAATCGCG)$_2$	B-DNA 56mer	Fiber	1x 290 1x 500
d(CGTGGCGGCAGTAGCGCGGTGGTCCCACCTGACCCCATGCC GAACTCAGAAGTGCG)$_2$	B-DNA 56mer	Fiber	300
d(CGCCGGCAGTAGCCGAAAAAATAGGCGCGCTCAAAAAAA TGCCCCATGCCGCGC)$_2$	B-DNA 56mer	Fiber	1x 360 1x 440 1x 500
d(ATCTTTGCGGCAGTTAATCGAACAAGACCCGTGCAATGCTA TCGACATCAAGGCCTATCGCTATTACGGGGTTGGGAGTCAATG GGTTCAGGATGCAGGTGAGGAT)$_2$	106-mer circle 10 turns (reg A)	Fiber	100
d(ATCTTTGCGGCAGTTAATCGAACAAGACCCGTGCAATGCTA TCGACATCAAGGCCTATCGCTATTACGGGGTTGGGAGTCAATG GGTTCAGGATGCAGGTGAGGAT)$_2$	106-mer circle 10 turns (reg B)	Fiber	100
d(ATCTTTGCGGCAGTTAATCGAACAAGACCCGTGCAATGCTA TCGACATCAAGGCCTATCGCTATTACGGGGTTGGGAGTCAATG GGTTCAGGATGCAGGTGAGGAT)$_2$	106-mer circle 10 turns (reg C)	Fiber	100
d(ATCTTTGCGGCAGTTAATCGAACAAGACCCGTGCAATGCTA TCGACATCAAGGCCTATCGCTATTACGGGGTTGGGAGTCAATG GGTTCAGGATGCAGGTGAGGAT)$_2$	106-mer circle 9 turns	Fiber	50
d(ATCTTGGCAGTTAATCGAACAAGACCCGTGCAATGCTATCG ACATCAAGGCCTATCGTTACGGGGTTGGGAGTCAATGGGTTCA GGATGCAGGTGAGGAT)$_2$	100-mer circle 9 turns	Fiber	100
147mer nucleosome	DNA-histones	1KX5	500
DNA:HU complex	DNA-HU protein	1P71	1000
DNA:HU complex	DNA-HU protein	1P71 (without mismatches and flipped bases)	1000

DNA:TRP repressor	DNA-repressor	1TRO	1000
DNA:leucine zipper	DNA-transc factor	2DGC	1000
DNA:P22 c2	DNA-represor	3JXC	1000
d(CGCAAATTTGCG)$_2$-distamycin	DNA-mG binder	2DND	700
d(CTTTTCGAAAAG) $_2$-Hoescht	Drug cooperativity	1QSX	10x 10
d(CGTACG)$_2$-daunomycin	DNA-intercalator	1D11	600
d(GGGG)$_4$	PS quadruplex	352D (without Thymine loops)	440
d(GGGG)$_4$	APS quadruplex	156D (without Thymine loops)	440
d(T•A•T)$_{10}$	PS triplex	Fiber	440
d(G•G•C)$_{10}$	PS triplex	Fiber	440
d(G•G•C)$_{10}$	APS triplex	Fiber	440
d(ATATATATATAT)$_2$	H-duplex	1GQU	720
d(CGATATATATAT)$_2$	H-duplex	2AF1	400
d(AAGGGTGGGTGTAAGTGTGGGTGGGT)	G_quadruplex	2LPW	5000
d(AGGGTTAGGGTTAGGGTTAGGG)	G-loop quadruplex(HTQ)	1KF1	1000
d(GGGGTTTTGGGG)$_2$	G quadruplex (OxyQ)	1JRN	1000
d(CCGGTACCGG)$_4$	Holliday Junction	1DCW	1000
d(CGCGCGCGCG)$_2$	Z-DNA, duplex	1I0T	2x 385
d(GCGAAGC)	Hairpinfold (REXMD)	1PQT	1000
d(CGCGAATTCGCG)$_2$	A-form in ethanol	1BNA	200
d(CGCGAATTCGCG)$_2$	A to B transition (H_2O)	1BNA	5x40
d(GGCGCC)$_2$	DNA unfolding (Pyridine)	1P25	400
d(CGCGAATTCGCG)$_2$	DDD, 0.15M NaCl	1BNA	2000
d(CGCGAATTCGCG)$_2$	DDD, 0.5M NaCl	1BNA	3000
d(CGCGAATTCGCG)$_2$	parmBSC0	1BNA	1500
d(CGCGAATTCGCG)$_2$	parmBSC0-OL1	1BNA	1500
d(CGCGAATTCGCG)$_2$	parmBSC0-OL4	1BNA	1500
d(CGCGAATTCGCG)$_2$	parmBSC0-OL1-OL4	1BNA	1500
d(CGCGAATTCGCG)$_2$	parmBSC0-Cheng-Garcia	1BNA	1500
d(CGCGAATTCGCG)$_2$	CHARMM36	1BNA	1500
d(CGCGAATTCGCG)$_2$	DDD, Amber GPU	1BNA	100
d(CGCGAATTCGCG)$_2$	DDD, Amber CPU	1BNA	100
d(CGCGAATTCGCG)$_2$	DDD, Gromacs GPU	1BNA	100
d(CGCGAATTCGCG)$_2$	DDD, Gromacs CPU	1BNA	100

Supplementary Table 2. MD-averaged helical parameters (on 1.2 μs simulation time) of Drew-Dickerson dodecamer in parmbsc1 simulations (and, as a control, other modern force-fields) compared with the NMR and X-ray estimates. [a]

	Twist	Roll	Slide	Rise	Shift	Tilt	BI(%)	Major groove width	Minor groove width
Parmbsc1	**34.3±5.4**	**1.5±5.4**	**-0.3±0.5**	**3.3±0.3**	**0.0±0.8**	**0.0±4.5**	**77**	**11.9±1.7**	**5.4±1.2**
Parmbsc0	32.8±5.8	2.7±5.8	-0.4±0.6	3.3±0.3	0.0±0.7	0.0±4.3	84	12.9±1.8	3.9±1.2
OL1	33.3±5.7	2.7±5.9	-0.2±0.6	3.3±0.3	0.0±0.7	0.0±4.4	83	12.2±1.4	6.1±1.3
OL4	33.3±6.4	2.6±5.9	-0.1±0.6	3.3±0.3	0.0±0.7	0.0±4.5	85	12.1±1.4	6.5±1.3
OL1+OL4	33.0±6.1	2.8±5.7	-0.3±0.6	3.3±0.3	0.0±0.7	0.0±4.3	86	12.4±1.5	6.0±1.2
C36 [d]	34.5±11	5.1±8.8	0.8±1.0	3.6±0.8	-0.1±1.1	0.9±8.0	66	10.5±1.5	8.3±1.7
Cheng-Garcia(CG)	32.5±3.4	1.5±5.2	-1.7±0.5	3.4±0.3	0.0±0.4	0.0±4.3	100	15.3±1.6	5.5±0.9
X-ray [b]	**35.2±0.6**	**-0.7±1.1**	**0.1±0.1**	**3.3±0.1**	**-0.1±0.1**	**-0.4±0.9**		**11.2±0.1**	**4.6±0.3**
NMR [c]	**35.6±0.8**	**1.6±1.0**	**-0.3±0.1**	**3.2±0.1**	**0.0±0.1**	**0.0±0.7**	**73[e]**	**11.9±0.3**	**4.7±0.3**

[a] Translational parameters and groove widths are in Å, while rotational parameters are in degrees. Note that for MD trajectories the standard deviations are computed from sequence-averages and time-averages. [b] X-ray mean values and standard deviations were obtained averaging the following structures (PDB id): 1BNA[1], 2BNA[2], 7BNA[3] and 9BNA[4]. [c] NMR mean values and standard deviations were obtained by averaging over the ensemble of structures contained in the PDB id 1NAJ[5].[d] These average values are contaminated by the opening of terminal base pairs (note large standard deviations in roll and twist). [e] Average value of BI population taken by averaging direct NMR estimates[6,7]. See also Supplementary Discussion and **Supplementary Figs. 29-31** for a discussion on the relative performance of parmbsc1 with respect to other force-fields.

1. Drew, H.R. *et al*. *Proc. Natl. Acad. Sci. U. S. A.***78,** 2179–2183 (1981).
2. Drew, H.R., Samson, S. & Dickerson, R.E. *Proc. Natl. Acad. Sci. U. S. A.***79,** 4040–4044 (1982).
3. Holbrook, S.R. *et al. Acta Crystallogr.,Sect.B***41,** 255–262 (1985).
4. Westhof, E. *J. Biomol. Struct. Dyn.***5,** 581–600 (1987).
5. Wu, Z., Delaglio, F., Tjandra, N., Zhurkin, V.B. & Bax, A.*J. Biomol. NMR***26,** 297–315 (2003).
6. Tian, Y., Kayatta, M., Shultis, K., Gonzalez, A., Mueller, L.J. & Hatcher, M.E., *J. Phys. Chem. B***113**, 2596–2603 (2008).
7. C. D. Schwieters, C.D. & Clore, G.M. *Biochemistry***46**, 1152–1166 (2007).

Supplementary Table 3. Ability of MD-ensembles obtained from parmbsc0 and parmbsc1 force fields to reproduce NMR observables for Drew-Dickerson dodecamer. The first block correspond to residual dipolar couplings Q-factor, $q = \sqrt{\sum\left(RDC_{calc} - RDC_{exp}\right)^2} \Big/ \sqrt{\sum RDC_{exp}^2}$, where $\mathrm{RDC_{exp}}$ has been determined using PALES[1], and the second block to NOEs (146 restraints).

	NMR	**X-ray**	**Fiber model B-DNA**	**Fiber model A-DNA**	**BSC1**	**BSC0**
Bicelles, 1NAJ [a], 129 RDCs	0.17	0.49	0.51	0.87	0.32	0.36
Bicelles, 1DUF [b], 204 RDCs	0.23	0.53	0.66	0.92	0.34	0.38
Sum of violations (A)	0.01	10.0	7.6	42.01	0.4	2.6
Largest violation (A)	0.01	1.0	0.4	1.3	0.2	1.3
Num. of violated restraints	1	35	36	84	2	5

[a] Data taken from ref. 2. [b] Data taken from ref. 3.

1. Zweckstetter, M. *Nat. Protoc.*, ***3***, 679-690 (2008).
2. Wu, Z., Delaglio, F., Tjandra, N., Zhurkin, V.B. & Bax, A.*J. Biomol. NMR***26,** 297–315 (2003).
3. Tjandra, N., Tate, S. I., Ono, A., Kainosho, M. & Bax, A. *J. Am. Chem. Soc.***122,** 6190–6200 (2000).

Supplementary Table 4. Different metrics showing the quality of parmbsc1 simulations for B-DNA duplexes.[a]

DNA seq or PDB id	Ref	RMSd	RMSd/bp	% H-bond	Avg. twist	Avg. roll
1BNA (12mer)	C	2.1 / 1.7	0.18 / 0.17	96 / 98	35.6 / 34.3	2.8 / 1.5
1NAJ (12mer)	N	1.7 / 1.4	0.15 / 0.15	96 / 98	35.6 / 34.3	2.8 / 1.5
CCATACgATACGG[b]	N	2.9 / 2.3	0.22 / 0.21	91 / 91	33.5 / 34.2	8.8 / 1.6
CCATACaATACGG[c]	N	3.3 / 3.1	0.26 / 0.28	93 / 94	33.7 / 34.1	2.7 / 2.5
CGCGACGTCGCG	F	2.0 / 1.5	0.17 / 0.15	98 / 99	34.8 / 34.6	3.1 / 2.0
CGCGTCGACGCG	F	2.6 / 1.5	0.22 / 0.16	97 / 99	34.1 / 34.5	3.4 / 2.3
GCGAGGGAGGGAGGGAGC	F	2.7 / 2.3	0.15 / 0.15	97 / 99	33.5 / 33.3	2.5 / 2.9
GCGCGGGCGGGCGGGCGC	F	2.3 / 2.0	0.13 / 0.13	97 / 99	33.7 / 33.7	2.8 / 3.3
GCGGGGGGGGGGGGGGGC	F	3.0 / 2.7	0.17 / 0.17	98 / 99	32.8 / 32.6	3.0 / 3.5
GCGTGGGTGGGTGGGTGC	F	2.2 / 1.9	0.12 / 0.12	97 / 99	33.1 / 33.0	2.7 / 3.2
GCCGAGCGAGCGAGCGGC	F	2.9 / 2.4	0.17 / 0.15	98 / 99	34.7 / 34.5	2.1 / 2.6
GCCTAGCTAGCTAGCTGC	F	2.2 / 1.9	0.13 / 0.12	97 / 98	34.3 / 34.2	1.6 / 2.1
GCTGCGTGCGTGCGTGGC	F	2.2 / 2.0	0.13 / 0.13	97 / 98	32.6 / 34.5	2.3 / 2.8
GCGATCGATCGATCGAGC	F	2.0 / 1.8	0.11 / 0.12	97 / 98	34.8 / 34.7	1.9 / 2.3
GCCTATAAACGCCTATAA	F	2.9 / 2.8	0.17 / 0.18	94 / 97	34.7 / 34.4	1.6 / 2.0
CTAGGTGGATGACTCATT	F	3.3 / 2.9	0.18 / 0.18	94 / 97	30.9 / 31.8	1.2 / 4.6
CACGGAACCGGTTCCGTG	F	3.0 / 2.9	0.17 / 0.18	95 / 97	34.6 / 33.8	2.7 / 2.0
GGCGCGCACCACGCGCGG	F	3.4 / 2.7	0.19 / 0.17	96 / 98	33.2 / 34.4	3.5 / 2.4
1D89 (12mer)	C	2.3 / 1.9	0.19 / 0.19	93 / 98	35.6 / 33.9	3.0 / 1.7
1FZX (12mer)	N	1.8 / 1.7	0.16 / 0.18	95 / 96	33.9 / 33.8	2.4 / 2.3
1RVH (12mer)	N	1.9 / 1.7	0.16 / 0.17	98 / 98	33.9 / 34.0	2.2 / 2.6
1SK5 (10mer)	C	2.1 / 1.8	0.21 / 0.23	93 / 97	34.2 / 34.3	1.7 / 1.7
CGATATATATATCG	F	1.9 / 1.6	0.16 / 0.17	96 / 97	34.4 / 34.4	2.9 / 1.7
2HKB (12mer)	N	1.8 / 1.7	0.15 / 0.17	96 / 97	34.1 / 33.8	2.3 / 2.6
2K0V (12mer)	N	2.4 / 2.1	0.20 / 0.22	95 / 96	33.9 / 33.5	2.2 / 1.9
2L8Q (12mer)	N	1.9 / 1.5	0.16 / 0.16	95 / 97	34.4 / 34.1	2.7 / 2.5
2LWG (10mer)	N	1.8 / 1.5	0.18 / 0.19	98 / 99	34.5 / 34.6	2.4 / 1.5
2M2C (14mer)	N	2.5 / 2.3	0.18 / 0.20	96 / 97	34.4 / 34.0	2.7 / 2.5

[a] The reference structures used for comparison were taken from X-ray crystallography (C), NMR (N) or fiber (F) data, as available. Except otherwise mentioned, all the duplexes were self-complementary and only one strand is noted. For structures available in the Protein Data Bank we display only the PDB code. RMSd are in Å and average rotational parameters are in degrees. Note that the first value in each cell corresponds to a sequence average considering the complete oligomer, while the second value in each cell was computed excluding the terminal residues. [b] Structure containing a G:G mismatch. The NMR structure used as reference was solved after parmbsc1 was derived[1]. [c] Same than [b] but containing an A:A mismatch.

1. Rossetti, G., Dans, P.D. *et al. Nucleic Acids Res.***43**, 4309-4321 (2015).

Supplementary Table 5. Long oligomers RMSd, helical parameters, and bending (reported herein as % of shortening) values, for all the residues or excluding the terminal ones, with respect to the ideal helix built using average dinucleotide X-ray helical parameters.

	Seq1[c]	**Seq2a**	**Seq2b**	**Seq3**	**Seq4a**	**Seq4b**
RMSd	4.4±1.3	4.2±1.5	4.3±1.3	6.7±2.8	7.2±2.7	7.4±2.7
RMSd (no ends)	4.2±1.2	4.0±1.4	4.1±1.2	6.4±2.6	6.9±2.6	7.0±2.5
RMSd / bp[a]	0.14	0.12	0.12	0.12	0.14	0.13
RMSd / bp (no ends)	0.14	0.12	0.12	0.12	0.13	0.13
Avg. twist (º)	34.9±7.3	35.0±5.3	34.5±5.4	34.2±5.6	34.8±5.3	34.3±5.8
Avg. roll (º)	2.1±8.4	1.5±5.8	1.7±5.8	2.2±5.7	1.7±5.8	2.0±6.0
Avg. slide (Å)	-0.4±0.7	-0.2±0.5	-0.3±0.6	-0.4±0.6	-0.2±0.5	-0.3±0.5
Shortening[b]	4±2 (16)	5±2 (20)	5±2 (17)	6±3 (18)	6±3 (23)	6±3(21)

[a] Values per base pair are indicated to avoid size-inconsistency. [b] Note that for helix shortening the maximum shortening percentages are reported in bracket.
[c] Seq1: ATGGATCCATAGACCAGAACATGATGTTCTCA in TIP3P water;
Seq2a: CGCGATTGCCTAACGAGTACTCGTTAGGCAATCGCG in SPCE water;
Seq2b: idem Seq2a in TIP3P water;
Seq3: CGCCGGCAGTAGCCGAAAAAATAGGCGCGCGCTCAAAAAAATGCCCCATGCCGCGC in TIP3P water;
Seq4a: CGCGATTGCCTAACGGACAGGCATAGACGTCTATGCCTGTCCGTTAGGCAATCGCG in SPCE water;
Seq4b: idem Seq4a in TIP3P water.

Supplementary Table 6. Statistic of NOE restraints violations for different nucleic acids (include: normal duplexes, hairpins, quadruplexes, and A-tracks).[a]

Structure (PDB id)	Number Restraints	Average Violation	Largest Violation	Number violations
1NAJ	146	*0.0001*	*0.01*	*1*
		0.003	2	1
2LPW	938	*0.0006*	*0.1*	*12*
		0.07[b]	7.0	45
1PQT	94	*0.01*	*0.1*	*3*
		0.01	0.1	2
1G14	218	*0.01*	*0.2*	*33*
		0.05	0.9	44
1RVH	446	*0.02*	*0.3*	*50*
		0.03	0.8	56
2LWG	415	*0.01*	*0.5*	*28*
		0.03	1.4	38
2K0V	634	*0.05*	*1.9*	*83*
		0.12	2.5	129
2L8Q	172	*0.0005*	*0.09*	*1*
		0.001	0.26	1
2M2C	296	*0.15*	*3.3*	*54*
		0.13	3.1	50
2NQ1	870	*0.02*	*1.3*	*111*
		0.09	3.9	162

[a] For each PDB entry we show the number of experimental restraints, the average deviation (A), the maximum deviation (A), and the number of restraint violations. In each cell NMR results are reported in italic, *i.e.*, the values obtained when experimental restraints were enforced to solve the structure; while the MD results obtained using parmbsc1 simulations are reported with normal characters. [b] Since the NOE deviations were larger than usual for this hairpin, calculations were repeated using parmbsc0 and CHARMM36 force-fields, finding 73 and 64 violations respectively.

Supplementary Table 7. Quality factor (Q-factor), $q = \sqrt{\sum\left(RDC_{calc} - RDC_{exp}\right)^2} \Big/ \sqrt{\sum RDC_{exp}^2}$, for the agreement between observed and predicted residual dipolar couplings (RDCs), using both experimental NMR structures and parmbsc1 MD simulations. [a]

Structure	**Alignment Method**	**Number RDCs**	**Q-factor (NMR)**	**Q-factor (MD)**
1NAJ	Bicelles	204	0.23	0.34
2LPW	Bicelles	57	0.25	0.54
1PQT	Pf1	29	0.11	0.41
1RVH	Pf1	72	0.13	0.27
2LWG	Pf1	46	0.18	0.29

[a] Note that lower Q-factor indicates better agreement. Typically data sets include both C-H and N-H dipolar couplings. The alignment media used to record NMR RDCs is indicated in all the cases. RDCs were back-calculated from the MD simulations using PALES.

Supplementary Table 8. Statistic of NOE violations for different nucleic acids, for oligomers solved after parmbsc1 development. NOE restraints here are determined using the full matrix relaxation and are more accurate than those typically found in the literature (rough data available upon request). [a]

Duplex	**Number restraints**	**Average violation**	**Largest violation**	**Number violations**[b]	**Rfactor$_{2\alpha}$**[c]
GG mismatch	246	*0.004*	*0.090*	*73|15|0*	*0.204*
		0.012	0.302	64|36|7	0.172
AA mismatch	230	*0.003*	*0.160*	*64|6|1*	*0.290*
		0.006	0.083	51|27|0	0.292
A***C***GT control	208	*0.006*	*0.046*	*85| 29|0*	*0.261*
		0.022	0.123	106|79|12	0.250
A***5mC***GT[d]	102	*0.034*	*0.205*	*57|49|14*	*0.197*
		0.035	0.189	60|45|18	0.243
A***5hmC***GT[e]	216	*0.004*	*0.045*	*63|18|0*	*0.232*
		0.014	0.218	86|57|2	0.236

[a] Note that the comparisons are made between metrics obtained for the NMR ensemble (the set of structures refined by imposing NMR restraints) in italics, and those coming from the unbiased MD trajectory in roman. [b] To define "number of violations" we used three criteria: i) the distances given by the flat well limits (left value in the cell), ii) the boundaries of the "contact" are extended by ±0.2 Å (middle value), and finally iii) the upper-limit is multiplied by 1.25 (right value in the cell). [c] The global quality factor Rfactor$_{2\alpha}$[1, 2] take values around 0.6 and 0.7 for B and A-DNA respectively. The sequences considered here are reported in **Supplementary Table 1**.[d] 5mC stands for 5-methyl-cytosine. [e] 5hmC stands for 5-hydroxymethyl-cytosine.

1. Gonzalez, C., Rullmann, J.A.C., Bonvin, A., Boelens, R. & Kaptein, R. *J. Magn. Reson.***91,** 659–664 (1991).
2. Gronwald, W. *et al. J. Biomol. NMR***17,** 137–151 (2000).

Supplementary Table 9. Different metrics of DNA flexibility in the Cartesian space for the Drew-Dickerson dodecamer simulation using parmbsc0 and parmbsc1 force-fields.

Metrics	**Parmbsc1**	**Parmbsc0**
Entropy all heavy [a]	2.14	2.14
	2.00	*2.00*
Entropy backbone	1.16	1.15
	1.11	*1.10*
First three eigenvalues [b]	176,127,102	204,135,104
Eigenvalues 10, 20 and 30	20,8,4	23,9,4
Self-similarity (10 eigenvalues)[c]	0.89	0.94
Similarity parmbsc1/parmbsc0[d]	0.81	
Relative similarity[e]	0.89	
Energy weighted similarity	0.88	
Relative weighted similarity	0.93	

[a] Entropies in kcal mol^{-1} K^{-1} are determined using Schlitter (roman) and Andrioacei-Karplus (italics) for the entire 1.2 µs simulations. [b] Eigenvalues (in Å^2) are computed by diagonalization of the covariance matrix and ordered according to their contribution to the total variance. [c] Self-similarity is computed by comparing the first and second halves of the same trajectory. [d] Similarity and weighted similarity indexes are computed using the Hess matrix[1], or following reference[2]. [e] Relative similarities are computed from absolute similarities and self-similarities as described elsewhere[3].

1. Hess, B. *Phys. Rev. E***62,** 8438 (2000).
2. Pérez, A. *et al. J. Chem. Theory Comput.***1,** 790–800 (2005).
3. Orozco, M., Pérez, A., Noy, A. & Luque, F.J. *Chem. Soc. Rev.***32,** 350–364 (2003).

Supplementary Table 10. Sequence-dependent dinucleotide force constants associated with the deformation of a single helical degree of freedom.[a]

bps	Twist	Tilt	Roll	Shift	Slide	Rise
AA	0.028 *0.036* **0.043** (0.092)	0.037 *0.045* **0.044** (0.100)	0.020 *0.023* **0.022** (0.049)	1.72 *1.68* **2.45** (3.98)	2.13 *2.91* **3.56** (6.16)	7.64 *9.33* **9.47** (21.75)
AC	0.036 *0.047* **0.034** (0.073)	0.038 *0.045* **0.034** (0.111)	0.023 *0.027* **0.025** (0.080)	1.28 *1.54* **1.55** (2.94)	2.98 *3.67* **3.33** (6.37)	8.83 *10.44* **8.31** (23.86)
AG	0.028 *0.031* **0.036** (0.064)	0.037 *0.049* **0.045** (0.149)	0.019 *0.025* **0.022** (0.096)	1.40 *1.54* **2.00** (3.21)	1.78 *2.78* **2.82** (7.19)	7.04 *9.73* **9.35** (29.50)
AT	0.031 *0.031* **0.032** (0.070)	0.035 *0.033* **0.032** (0.166)	0.022 *0.024* **0.023** (0.055)	1.05 *1.24* **1.21** (3.17)	3.77 *4.10* **3.49** (10.69)	9.34 *9.23* **7.32** (25.55)
CA	0.015 *0.028* **0.032** (0.043)	0.025 *0.028* **0.027** (0.082)	0.016 *0.016* **0.018** (0.048)	1.05 *0.77* **1.60** (3.73)	1.80 *2.69* **2.19** (2.40)	6.30 *7.66* **6.71** (18.24)
CC	0.026 *0.032* **0.030** (0.041)	0.042 *0.049* **0.043** (0.119)	0.020 *0.021* **0.021** (0.064)	1.43 *1.50* **1.53** (2.43)	1.57 1.78 **1.74** (3.54)	7.86 *9.59* **8.96** (30.31)
CG	0.014 *0.024* **0.032** (0.047)	0.026 *0.032* **0.024** (0.068)	0.016 *0.016* **0.017** (0.050)	1.05 *1.10* **1.82** (1.59)	1.91 *2.47* **2.48** (3.30)	6.11 *7.61* **6.64** (14.16)
GA	0.024 *0.034* **0.040** (0.071)	0.038 *0.045* **0.041** (0.087)	0.020 *0.023* **0.024** (0.046)	1.32 *1.40* **2.27** (6.54)	1.88 *2.66* **3.40** (2.78)	8.48 *10.08* **10.12** (22.82)
GC	0.022 *0.031* **0.027** (0.055)	0.036 *0.043* **0.031** (0.082)	0.026 *0.025* **0.028** (0.082)	1.18 *1.32* **1.70** (3.35)	2.59 *3.19* **4.79** (6.24)	9.47 *11.16* **9.43** (25.86)
TA	0.018 *0.028* **0.036** (0.052)	0.019 *0.025* **0.021** (0.148)	0.015 *0.015* **0.015** (0.029)	0.64 *0.50* **0.93** (3.86)	1.25 *2.16* **1.52** (2.35)	6.08 *7.47* **6.61** (21.91)

[a] Parmbsc0 (roman)[1], parmbsc1 (italics), and CHARMM27 (bold) force-fields are compared with stiffness values derived from inspection of the X-Ray structural variability of the different base pair steps (in brackets)[2]. Note that values for a particular base pair step are diagonal entries of its stiffness matrix. Values reported in the table are averages over all the equivalent steps. The rotational values are in kcal mol^{-1} deg^{-2} and translational ones are in kcal mol^{-1} Å^{-2}.

1. Perez, A., Lankas, F., Luque, F. J. & Orozco, M. *Nucleic Acids Res.*36, 2379–2394 (2008).
2. Olson, W.K., Gorin, A.A., Lu, X.-J., Hock, L.M. & Zhurkin, V.B. *Proc. Natl. Acad. Sci.***95,** 11163–11168 (1998).

Supplementary Table 11. Elastic properties derived from atomistic MD simulations of three sequences of DNA.[a]

Persistence length							**Other stiffness descriptors**	
DNA	*Roll*	*Tilt*	*Isotropic*	*Dynamics*	*Static*	*Total*	*Torsion module*	*Stretch Module*
Seq3[b]	41±10	63±16	49±11	*63±1*	*566±150*	*57±2* 41±20 49±20	48±19 *101±9*	1,373±195 *1,857±22*
Seq4a	41±8	64±14	50±9	*71±1*	*608±150*	*64±2* 42±23 50±23	49±13 *102±10*	1,430±210 *1,567±42*
Seq4b	41±7	65±15	50±9	*71±1*	*310±44*	*57±2* 39±20 48±21	46±13 *107±12*	1,476±185 *1,832±45*
Avg.	41±14	64±26	50±17	*68±2*	*495±211*	*59±4* 41±30 49±30	47±26 *104±18*	1,426±341 *1,752±65*

[a] Persistence lengths and torsion modules are in nm, and stretch module are in pN. Values in roman correspond to 2 bp windows, while values in italic correspond approximately to one DNA turn windows[1]: (i) persistence lengths are calculated by linearly fitting the directional decay from 2 bp until 11 bp sub-fragments, and the static contributions come from the distribution of sequence-dependent static bends obtained through the MD average structure; (ii) stretch modulus are obtained by linearly fitting end-to-end variances of all central sub-fragments containing from 8 bp up to 16 bp to avoid the very long end-effect; (iii) torsional modulus is evaluated by averaging the 38 central sub-fragments containing 11 bp. Only the central 48-mer of the 56-mers was considered to minimize end-effects. Underlined values were obtained using a local implementation of Olson's Monte Carlo procedure[2], without additional corrections, or including (underlined with a curved line) partial variance corrections as discussed in Noy and Golestanian 2012[1].[b] See **Supplementary Table 5** for the definition of the sequences. As reference experimental estimates for persistence lengths are around 50 nm[3], for static persistence lengths are in the range of 200-1,500 nm[4, 5], for stretch modulus are around 1,100-1,500 pN[6, 7] and for torsion (twist) constants are in the range 80-120 nm[8, 9].

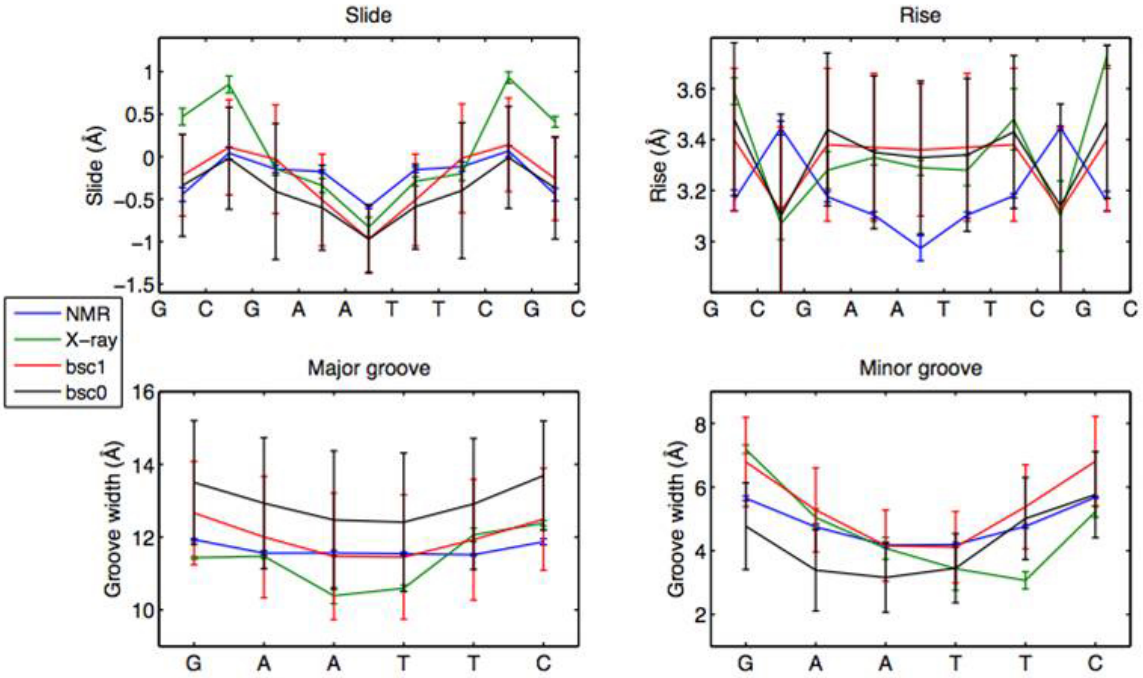
1. Noy, A. & Golestanian, R. *Phys. Rev. Lett.***109,** 228101 (2012).
2. Zheng, G., Czapla, L., Srinivasan, A.R. & Olson, W.K. *Phys. Chem. Chem. Phys.***12,** 1399–1406 (2010).
3. Mazur, A.K. & Maaloum, M. *Nucleic Acids Res.***42,** 14006-14012 (2014).
4. Smith, S.B., Finzi, L. & Bustamante, C. *Science***258,** 1122–1126 (1992).
5. Moukhtar, J. *et al. J. Phys. Chem. B***114,** 5125–5143 (2010).
6. Smith, S.B., Cui, Y. & Bustamante, C. *Science***271,** 795–799 (1996).
7. Gross, P. *et al. Nat. Phys.***7,** 731–736 (2011).
8. Strick, T.R., Allemand, J.-F., Bensimon, D., Bensimon, A. & Croquette, V. *Science***271,** 1835–1837 (1996).
9. Moroz, J.D. & Nelson, P. *Proc. Natl. Acad. Sci.***94,** 14418–14422 (1997).

Supplementary Table 12. Differences between QM and force-field estimates for the parameterized systems. Values refer to calculations performed in water.

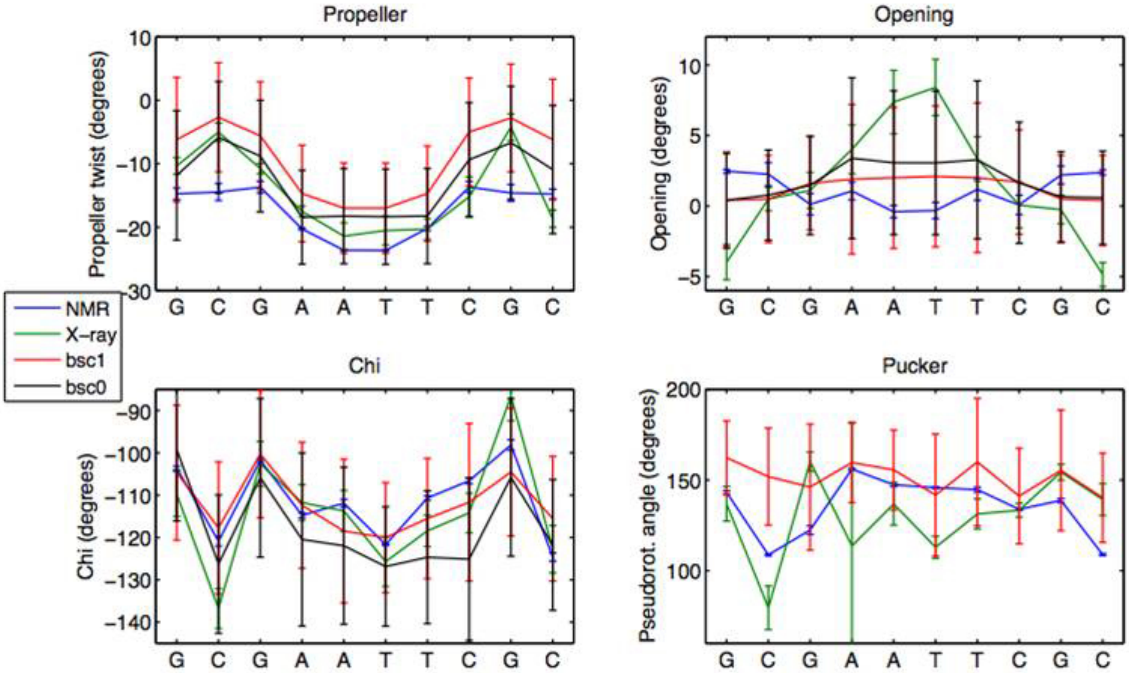
Torsion	Adenosine	Guanosine	Cytosine	Thymidine
		Glycosidic torsion (χ)		
		Geometries (°) [a]		
Anti	14 / 40	9 / 40	2.5 / 1	2.5 / 1
Barrier	1.5 / 11	2.5 / 15	13 / 10	11 / 11
Syn	7 / 32	2.5 / 30	12 / 30	–12 / 30
		Energies (kcal mol^{-1}) [b]		
Anti/Syn	0.0 / –0.3	–0.4 / –0.6	–1.1 / 1.3	–0.8 / 1.7
Barrier [c]	0.3 / –2.0	0.0 / –2.1	–0.6 / –0.7	–0.9 / –1.2
Profile	0.3 / 2.5	1.2 / 2.8	0.9 / 4.0	0.9 / 3.9
		Phase angle (P)		
		Geometries (°) [a]		
North	10 / 30	10 / 10	10 / 40	0 / 10
East	0 / 10	0 / 0	10 / 10	0 / 10
South	0 / 0	10 / 10	0 / 0	0 / 0
		Energies (kcal mol^{-1}) [b]		
North/South	–0.1 / –1.5	0.0 / –1.0	–0.6 / –1.6	0.5 / –0.5
East Barrier	–0.2 / 0.4	–0.5 /0.7	–0.1 / 1.2	–0.8 / 0.0
Profile	0.4 / 0.6	0.5 / 0.4	0.4/ 0.7	0.2 / 0.5

[a] Errors in the position of the minima and transition state when parmbsc1 (first number in the cell) or parmbsc0 (second number in the cell) values are compared with MP2 geometries. [b] Errors in the estimates of the relative stability and transition barrier when parmbs1 (first number in the cell) or parmbsc0 (second number in the cell) values are compared with single-point CCSD(T)/CBS results. [C] Energy values refer to barrier at χ around 120 degrees, note that the large barrier located at χ around 0 is very well reproduced at the parmbsc1 level, but very poorly at the parmbsc0 one (**Supplementary Fig. 26**).

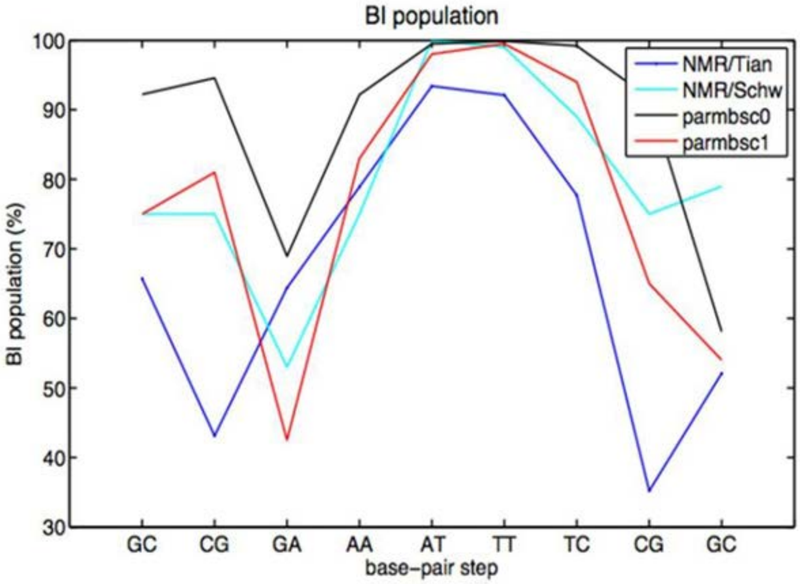
SUPPLEMENTARY FIGURES



Supplementary Figure 1| Helical parameters of DDD: Slide, Rise and grooves' width. Comparison of slide, rise, major and minor groove width average values per base-pair step coming from NMR structure pdb: 1NAJ (blue), X-ray structure pdb: 1BNA (green), 1 µs run using parmbsc0 force-field (black) and 1.2 µs run using parmbsc1 force-field.



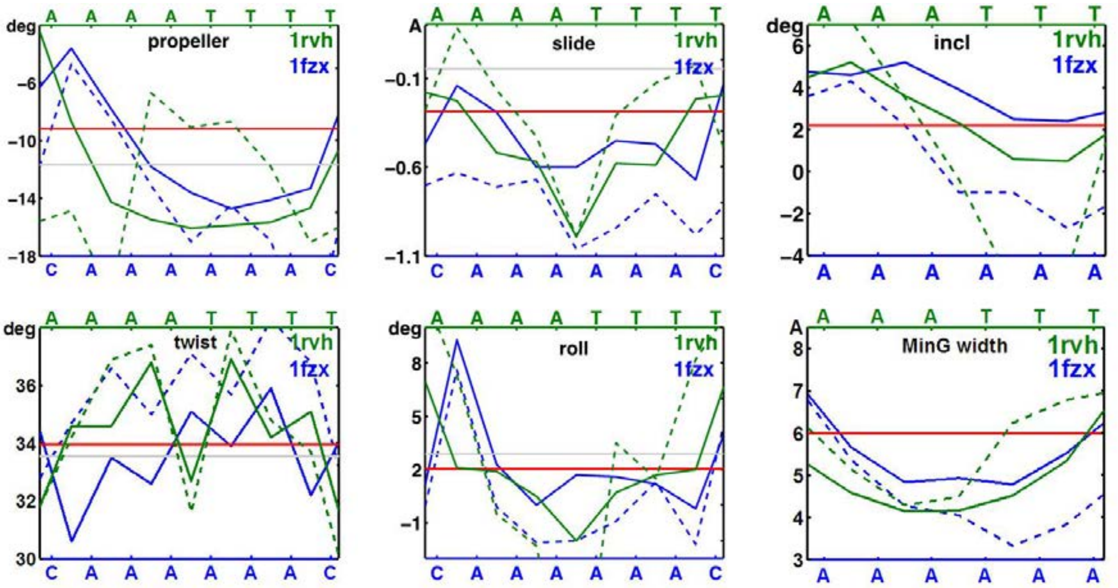
Supplementary Figure 2| Helical parameters per base-pair of DDD. Comparison of propeller twist, base opening, χ (chi) and pseudo-rotational angle (pucker) average values per base-pair step coming from NMR structure pdb:1NAJ (blue), X-ray structure pdb:1BNA (green), 1 μs run using parmbsc0 force-field (black), and 1.2 μs run using parmbsc1 force-field.



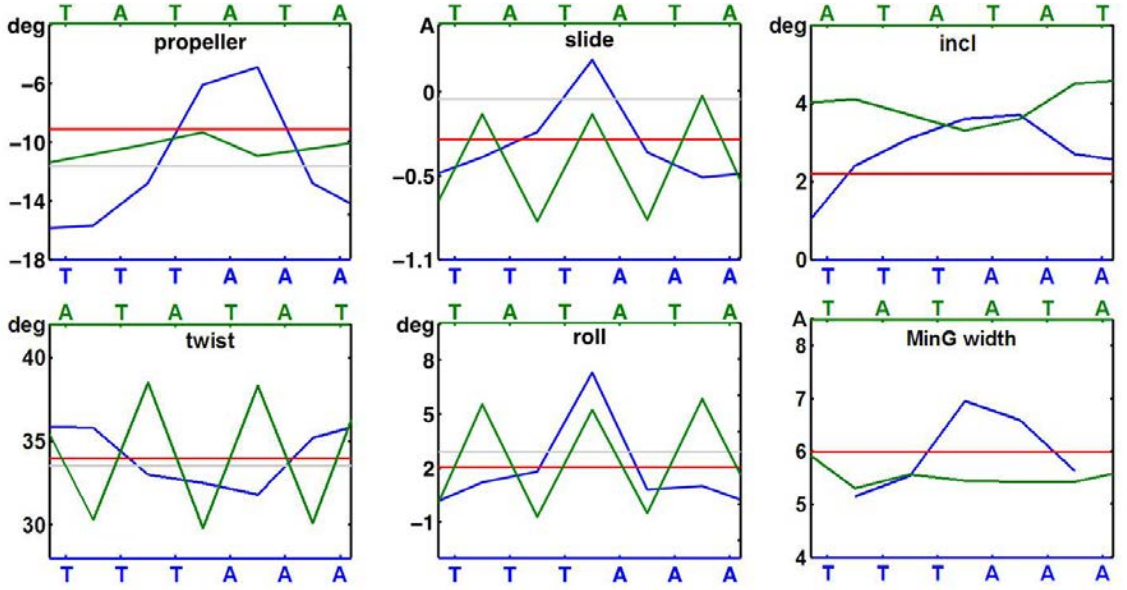
Supplementary Figure 3| BI/BII populations of DDD. Comparison of BI population percentage per base-pair step for DDD. Values coming from NMR/Tian *et al.*[1] (blue), NMR/ Schwieters *et al.*[2] (light blue), 1 µs run using parmbsc0 force-field (black) and 1.2 µs run using parmbsc1 force-field (red).

1. Tian, Y., Kayatta, M., Shultis, K., Gonzalez, A., Mueller, L.J., & Hatcher, M.E. *J. Phys. Chem. B***113**, 2596–2603 (2008).
2. Schwieters, C.D. & Clore, G.M., *Biochemistry***46**, 1152–1166 (2007).

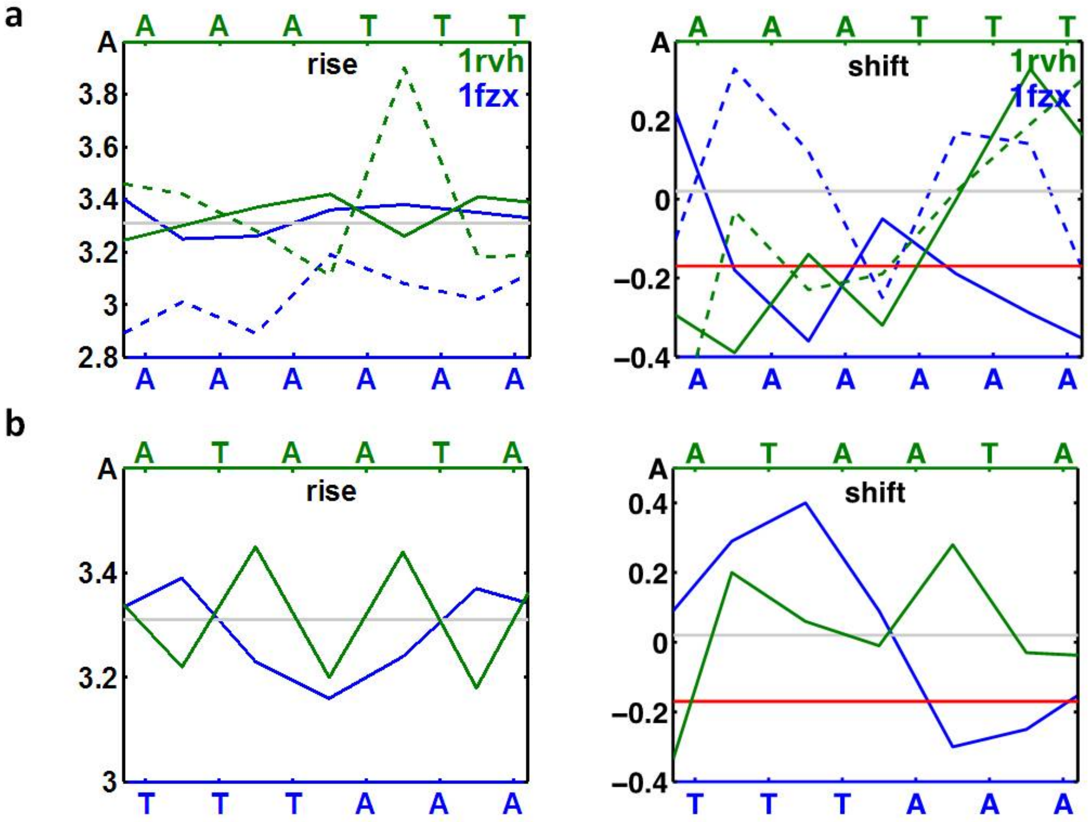
,



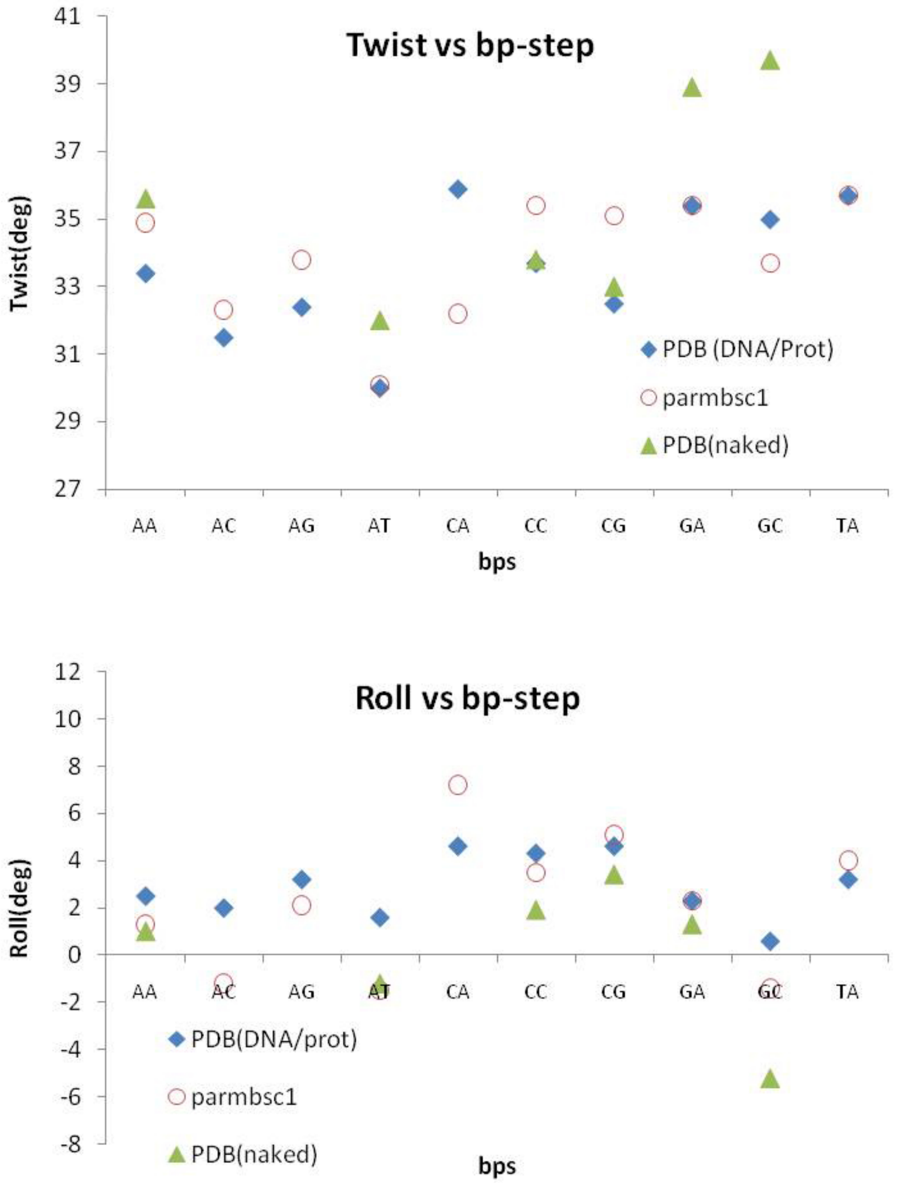
Supplementary Figure 4| Helical parameters of A-tract sequences: AATT and AAAA. Comparison in structural characteristics such as propeller twist, slide, inclination, twist, roll and minor groove width of values obtained using parmbsc1 force-field (full line) and experimental values (dashed lines) for AATT (pdb code:1RVH) (green) and AAAA (pdb code: 1FZX) (blue) sequences. Experimental average is represented with a grey line, while parmbsc1 average is represented with a red line.



Supplementary Figure 5| Helical parameters of A-tract sequences: ATAT and TTAA. Comparison in structural characteristics such as propeller twist, slide, inclination, twist, roll and minor groove width of values obtained using parmbsc1 force-field (full line) and experimental values (dashed lines) for ATAT (green) and TTAA (blue) sequences. Experimental average is represented with a grey line, while parmbsc1 average is represented with a red line.

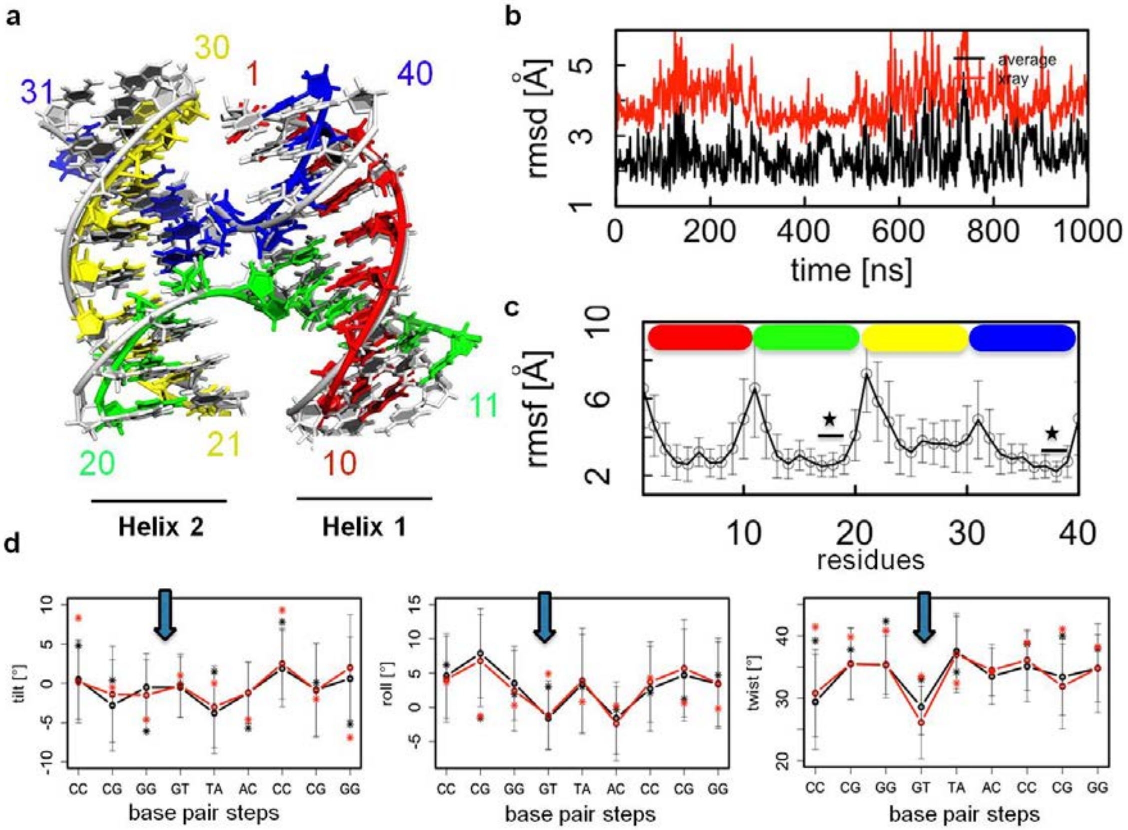


Supplementary Figure 6| Base-pair step helical parameters of A-tract sequences. Comparison in rise and shift of values obtained using parmbsc1 force-field (full line) and experimental values (dashed lines) for **(a)** AATT (pdb code:1RVH) (green) and AAAA (pdb code: 1FZX) (blue) and **(b)** ATAT (green) and TTAA (blue) sequences. Experimental average is represented with a grey line, while parmbsc1 average is represented with a red line.



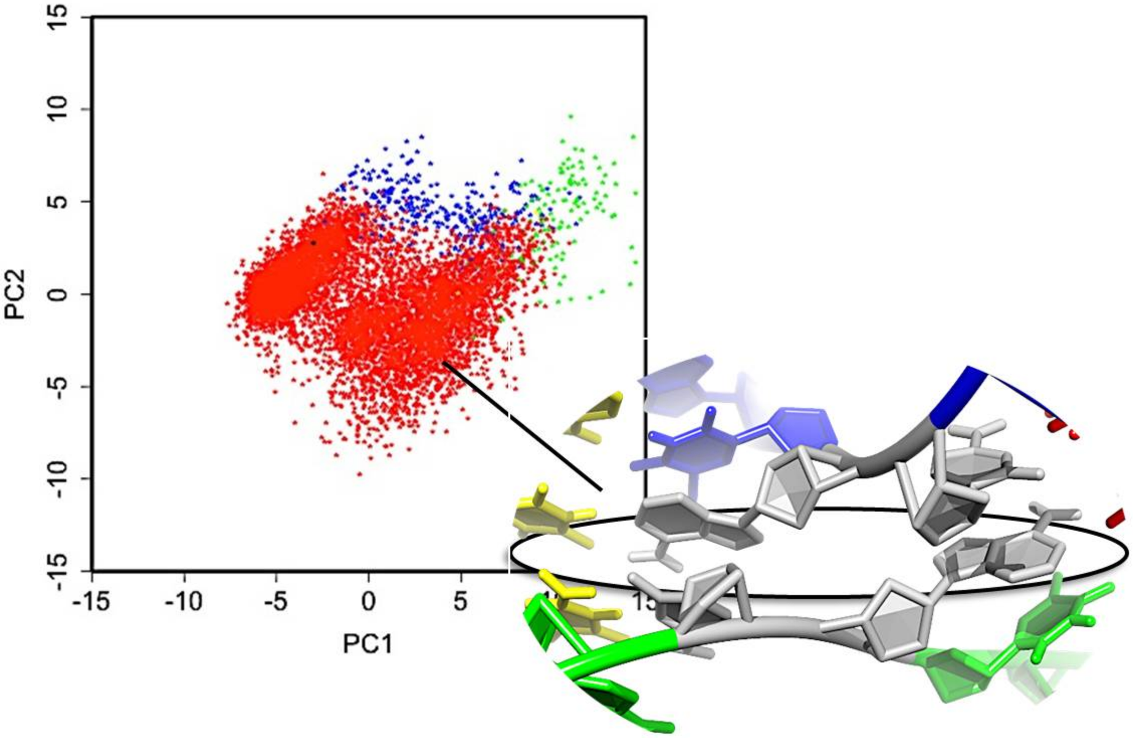
Supplementary Figure 7| Sequence-dependent variability of twist and roll. Comparison of DNA-protein complexes (blue), naked DNA (green) and parmbsc1 (red) values for twist (top) and roll (bottom) values per base-pair step. Values of DNA-protein complex come from analysis of 636 structures from PDB, while values of naked DNA come from analysis of 103 structures from PDB[1].

[1] Dans, P.D., Pérez, A., Faustino, I., Lavery, R. & Orozco, M. *Nucleic Acids Res.*40, 10668–10678 (2012).

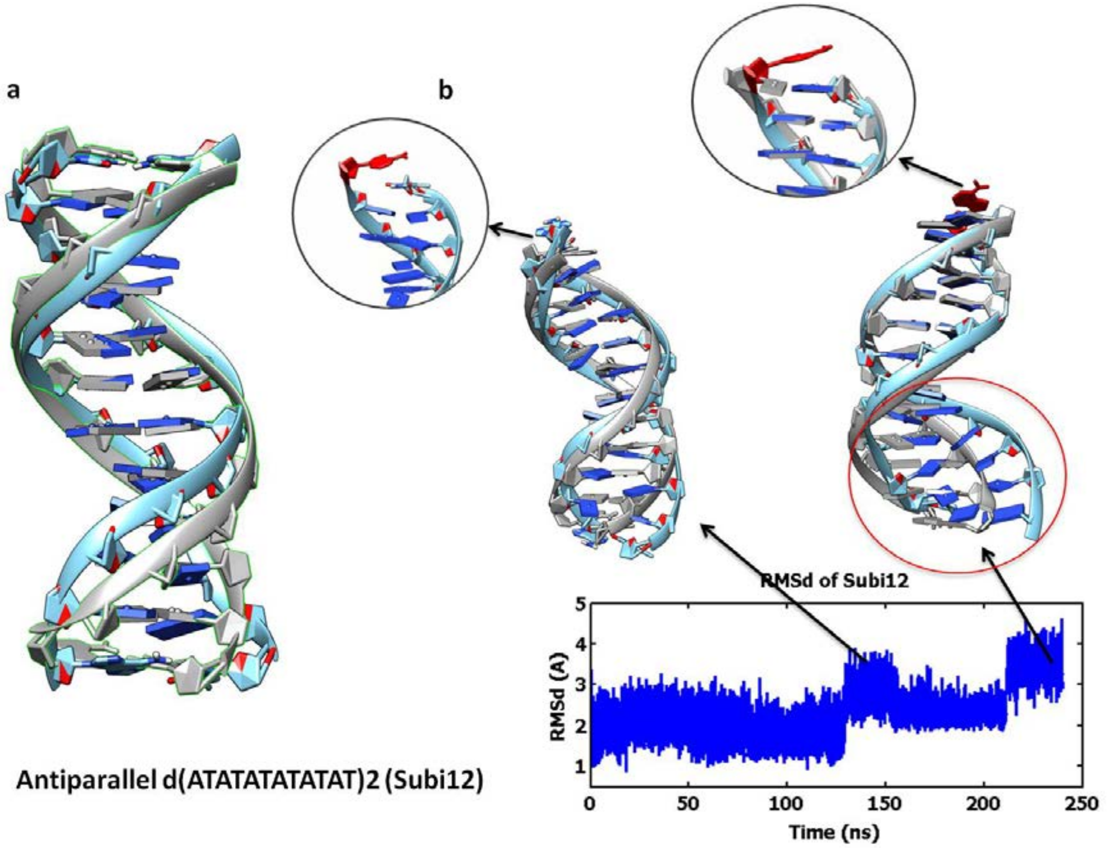


Supplementary Figure 8| Holliday junction structural features are close to x-ray (1DCW) structure. (**a**) Structural comparison of the time-averaged structure (in colors) with the x-ray reference structure (grey). (**b**) All heavy atoms RMSD and (**c**) per-residue RMSD from 1 µs MD simulation. X-ray structure was also taken as reference in the per-residue RMSD calculation. Note the higher RMSD values correspond to end strand bases. Starred residues are placed in the junction between helices. (**d**) Selected time-averaged helical parameters for the symmetric helices I and II. For experimental reference structures see ref. 1.

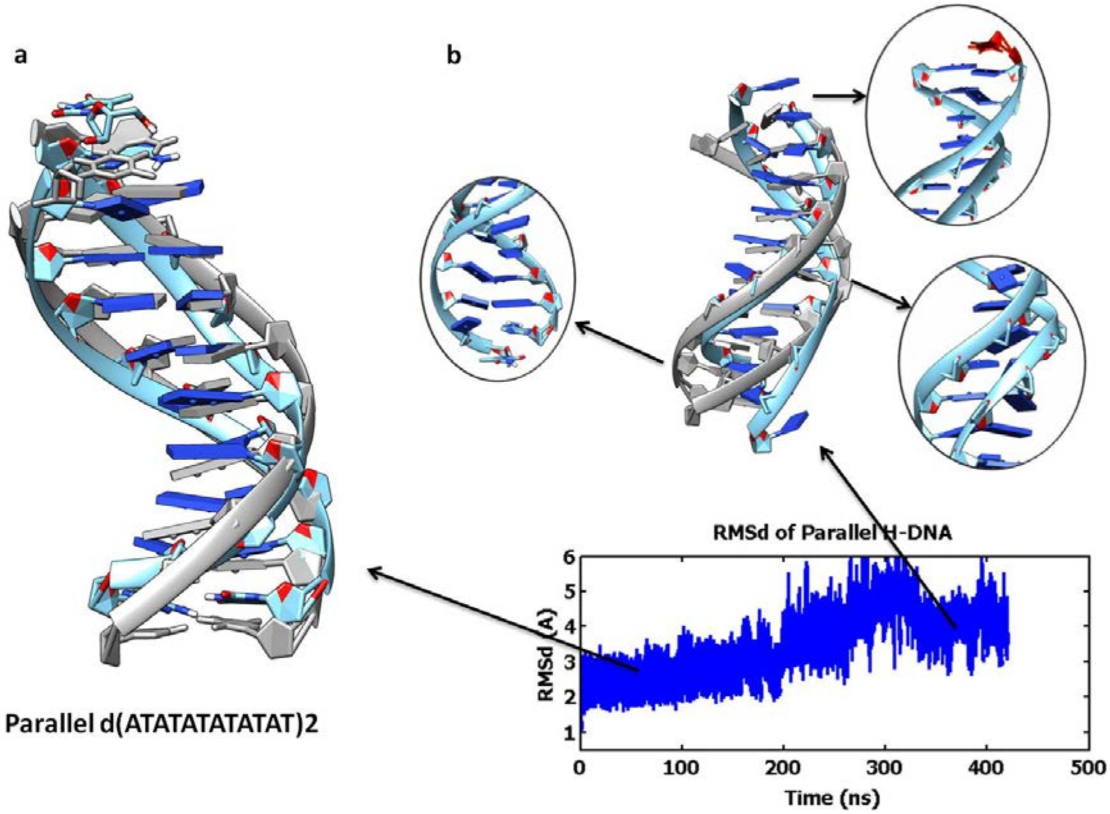
1. McKinney, S.A., Déclais, A.-C., Lilley, D.M.J. & Ha, T. *Nat. Struct. Mol. Biol.***10,** 93–97 (2003).



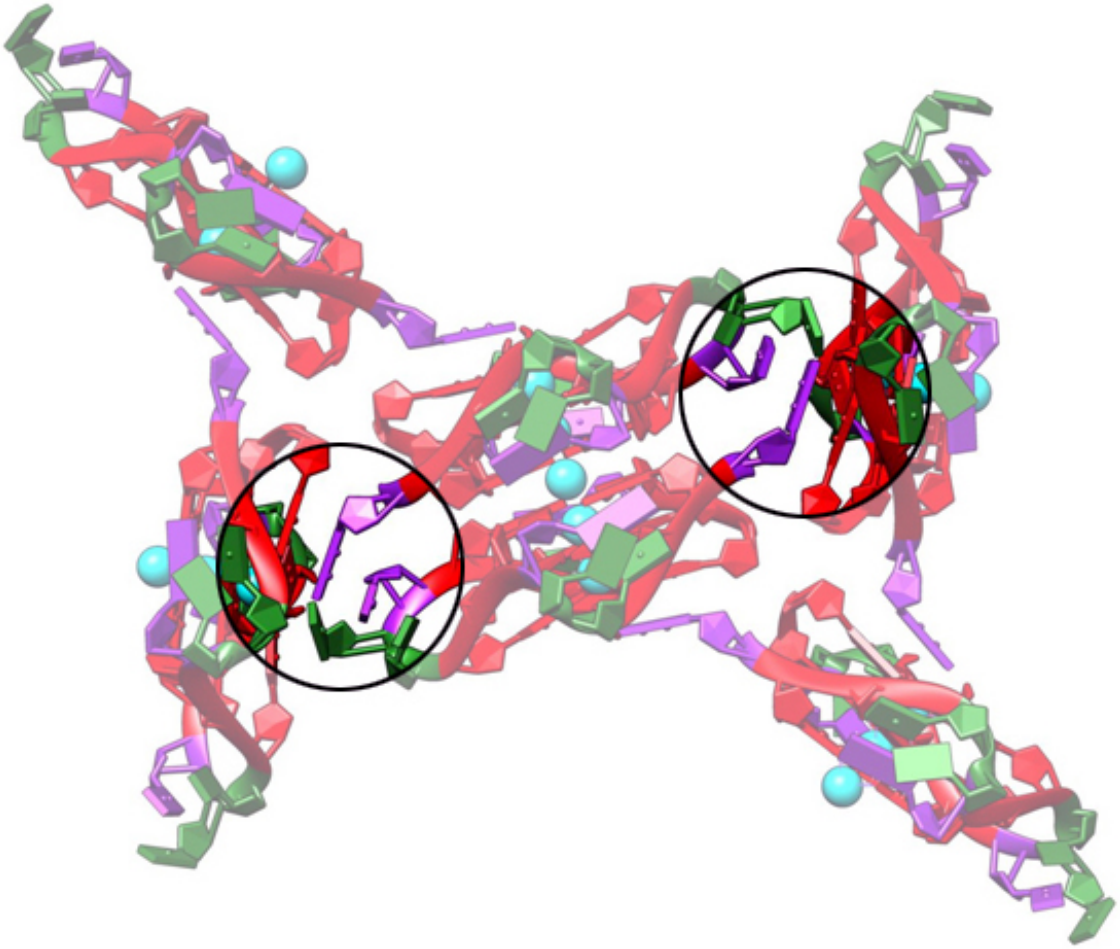
Supplementary Figure 9| Holliday junction PCA results. Projection to the first two PCA-eigenvectors based on the heavy atoms of junction bases (residues 16, 17, 36, and 37). The major conformation (in red) is present over ~95% of the simulation.



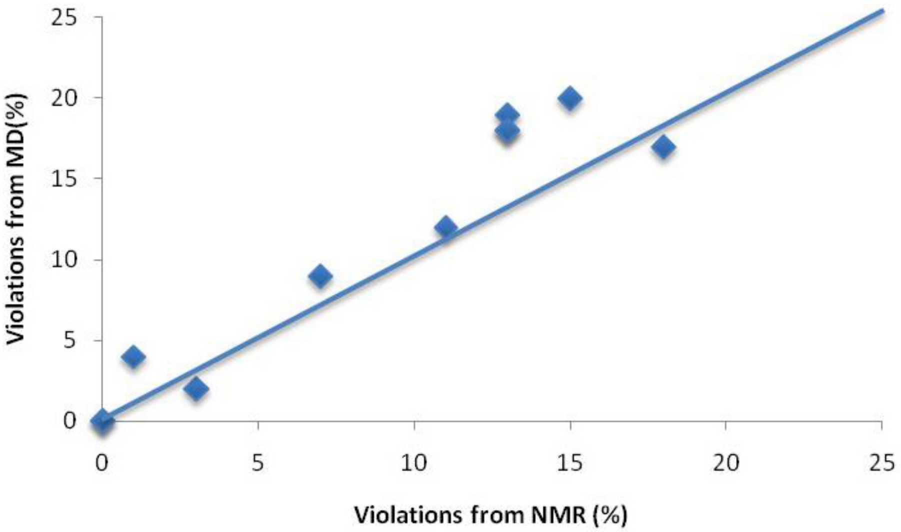
Supplementary Figure 10| Simulation antiparallel of H-DNA. (a) Comparison of experimental structure (made from pdb code: 1GQU) (grey) with the last snapshot of a 250 ns run using parmbsc1 (light blue). Bellow is an illustration of the duplex sequence. **(b)** RMSd of the 250 ns run with several snapshots plotted along the trajectory (light blue) compared with the experimental structure (grey) with highlighted distortions in the duplex.



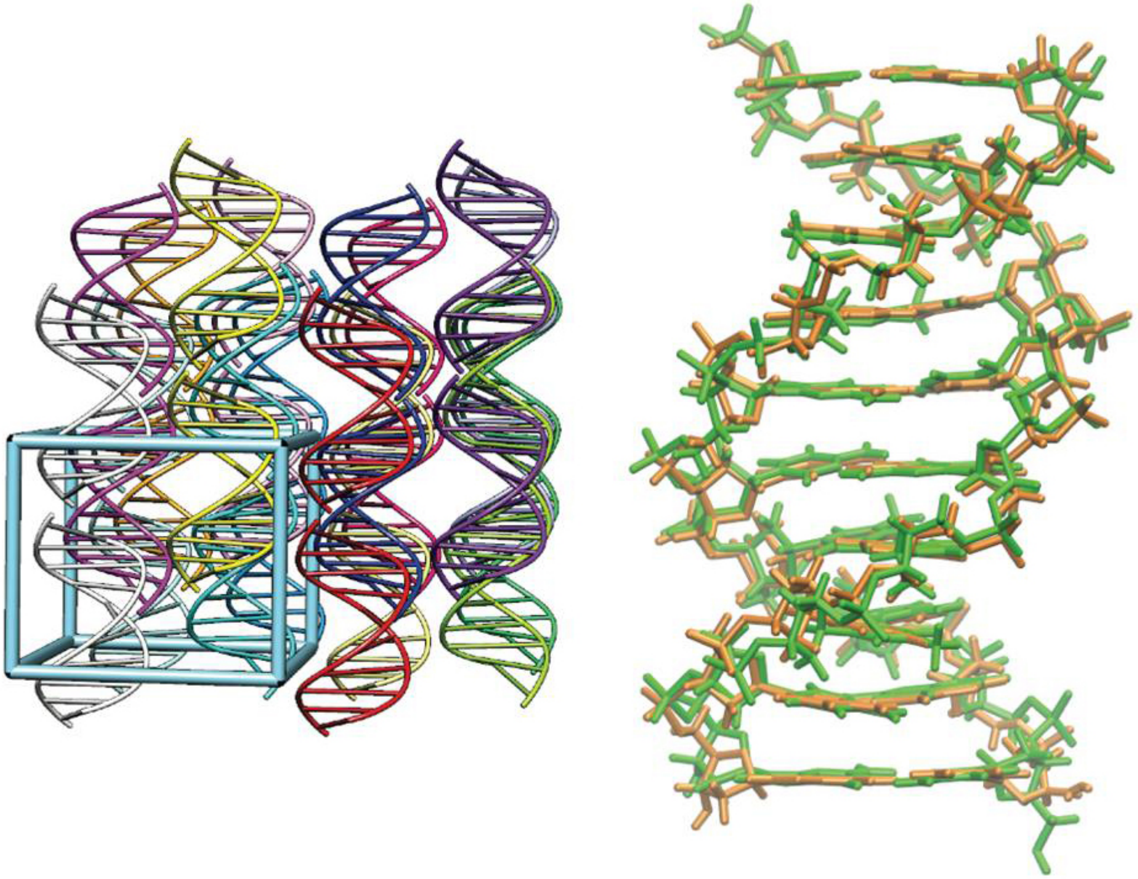
Supplementary Figure 11| Simulation of parallel H-DNA. (a) Comparison of experimental structure (grey) with a snapshot from a 400 ns run using parmbsc1 (light blue). **(b)** RMSd of the 400 ns run with several snapshots plotted along the trajectory (light blue) compared with the experimental structure (grey) with highlighted sever distortions in the duplex.



Supplementary Figure 12| Crystal packing of Human Talomeric Quadruplex (HTQ). Crystal packing of HTQ quadruplex (pdb code: 1KF1) showing interactions between loops' bases and other crystal units. Loop residues stacked to the neighboring units are highlighted in the circles.



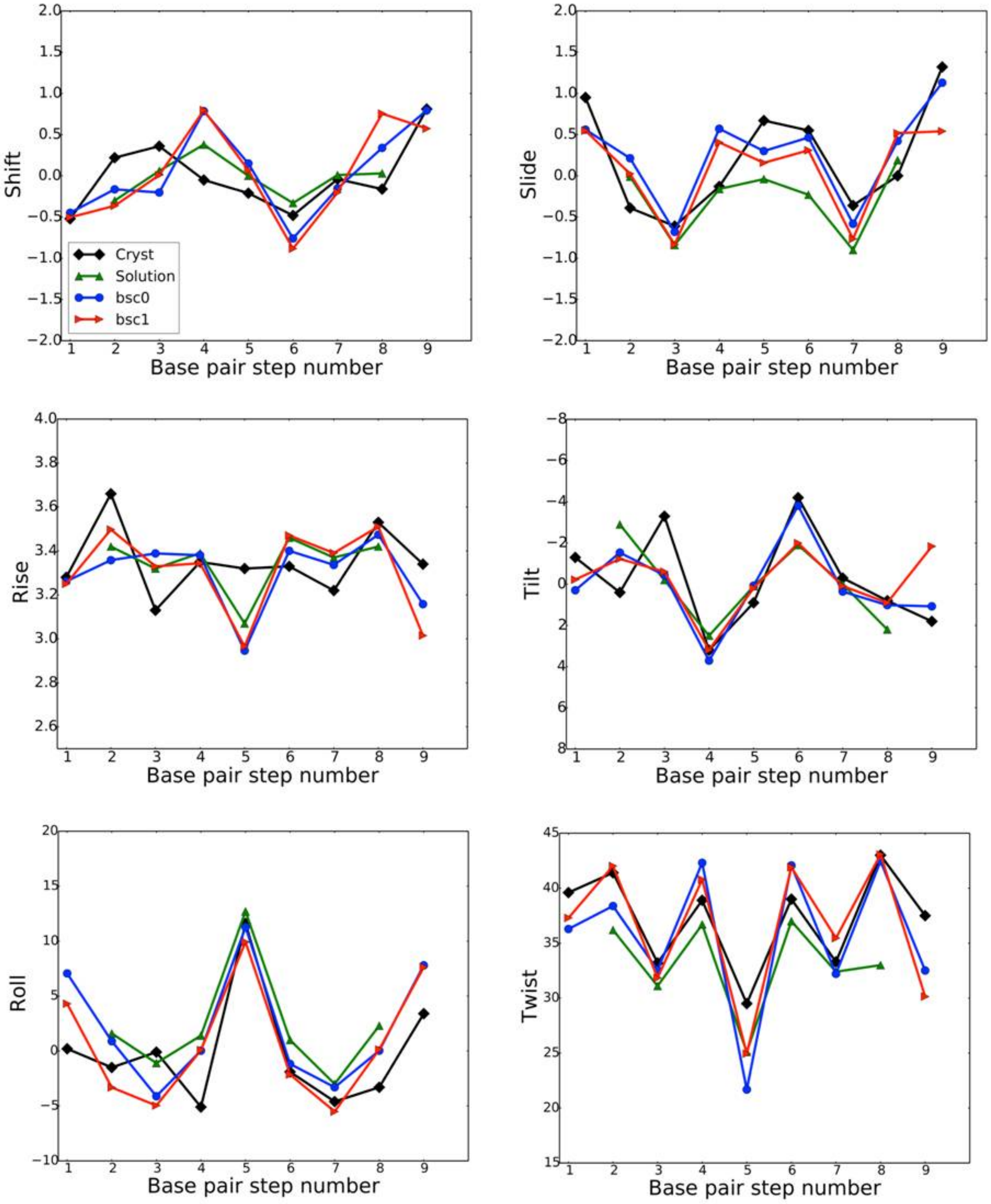
Supplementary Figure 13| Correlation between the number of violations in NOE restraints found in MD-parmbsc1 trajectories and corresponding NMR models. See **Supplementary Table 7** for details on structures.



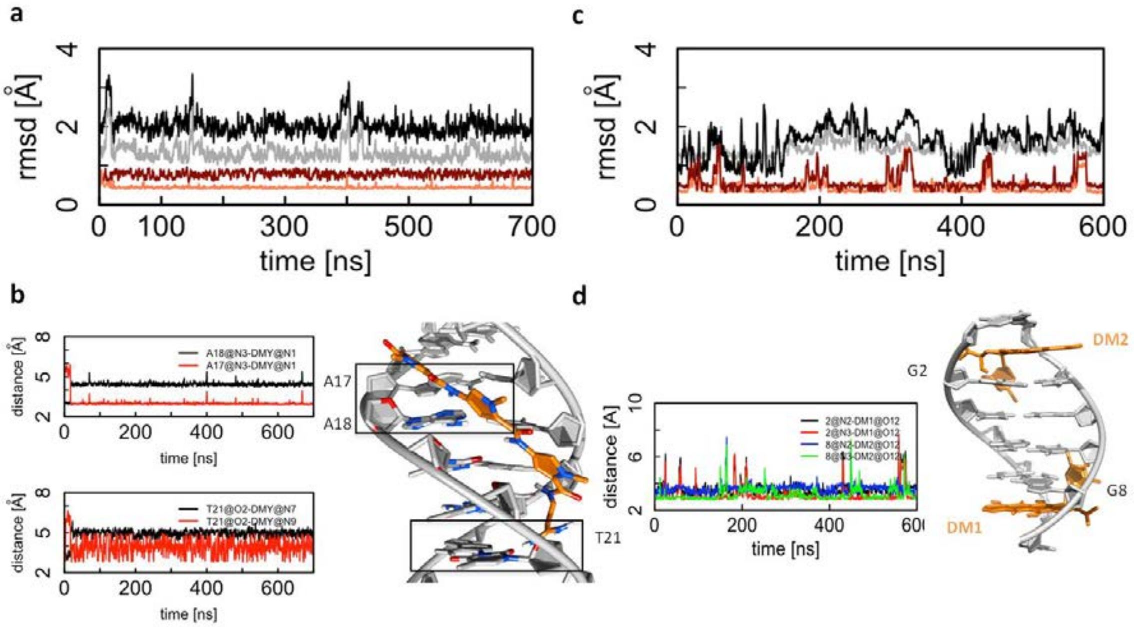
Supplementary Figure 14| Representation of the crystal structure simulation of a B-DNA duplex (PDB: 1D23). The **s**imulation box used in the crystal simulations is shown on the left, while comparison between the best-fit average structure from parmbsc1 simulations (orange) and the crystal structure (green) are shown on the right. Note that the RMS deviation for all DNA heavy atoms of the simulation average structure (compared to the PDB structure) is 0.70 Å. This can be compared to 0.77 Å for a crystal simulation using parmbsc0, and 1.83 Å for a solution simulation also using parmbsc0[1].

1. Liu, C., Janowski, P.A. & Case, D.A. *Biochim. Biophys. Acta (BBA)-General Subj.***1850**, 1059–1071 (2014).

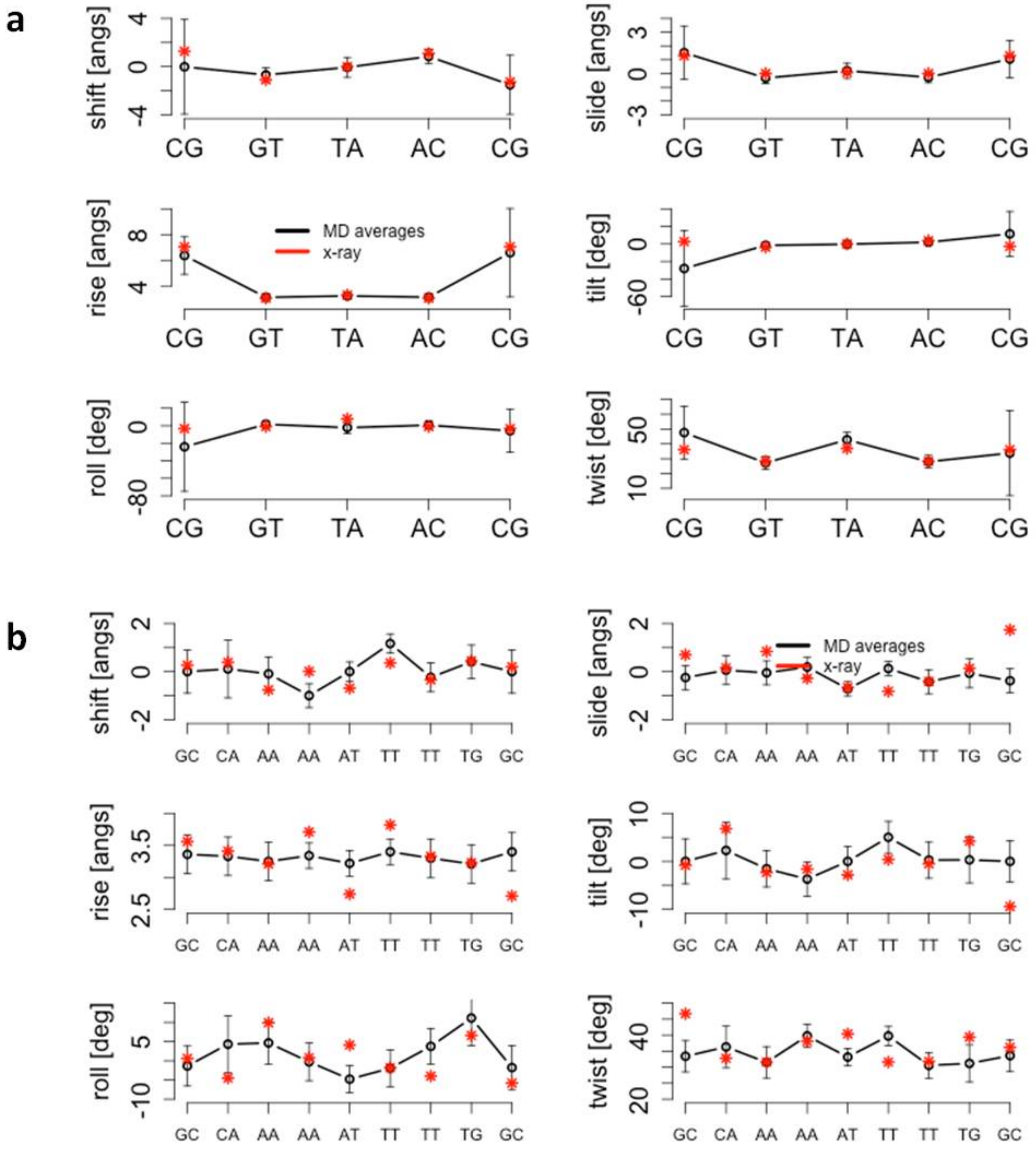
.



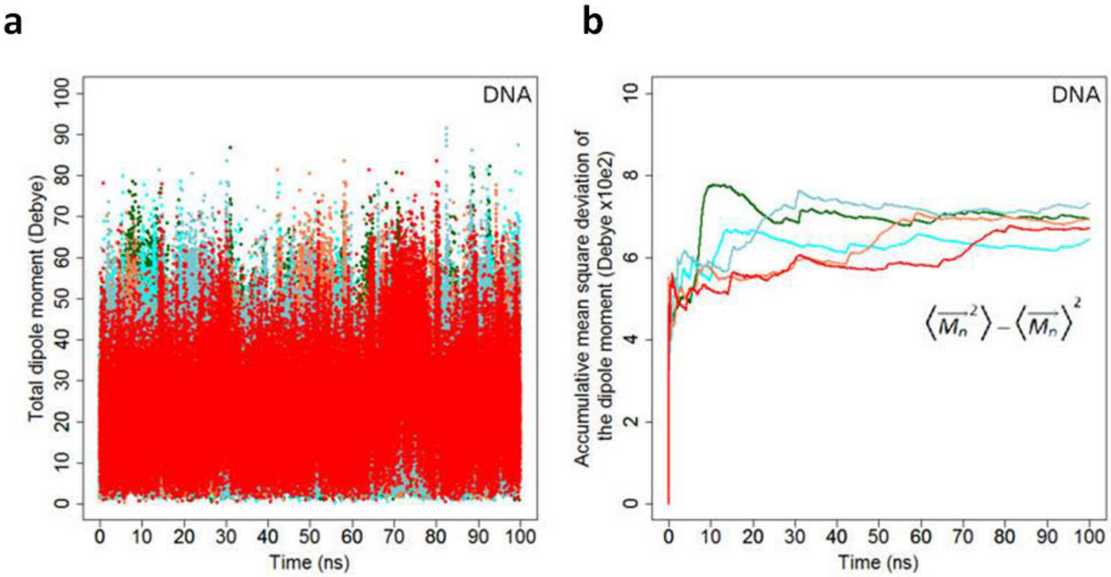
Supplementary Figure 15| Helicoidal analysis of a simulation of a B-DNA duplex (PDB: 1D23) within crystal environment. Helical parameters comparing results from simulation using parmbsc0 (blue) and parmbsc1 (red) force-fields, a simulation in solution (green) and the crystal structure (black).



Supplementary Figure 16| Representative stability properties in drug-DNA complexes with parmbsc1. RMSD (**a**) and representative distance between the distamycin A and the closest residues. (**b**) RMSD plots relative to x-ray (PDB id: 2DND), and MD-average structures for DNA (black and grey respectively) and distamycin A (red and orange respectively). Original contacts with the DNA are rapidly replaced by neighboring atoms keeping distamycin A within the minor groove. RMSD (**c**) and representative distances between the first daunomycin (PDB id: 1D11) and the closest guanine. (**d**) Second daunomycin's RMSd values are similar. Stabilizing interactions (h-bonds) between the N3 of guanine (residues 2 and 8 respectively) and a hydroxyl group in the daunomycin were stable along time.



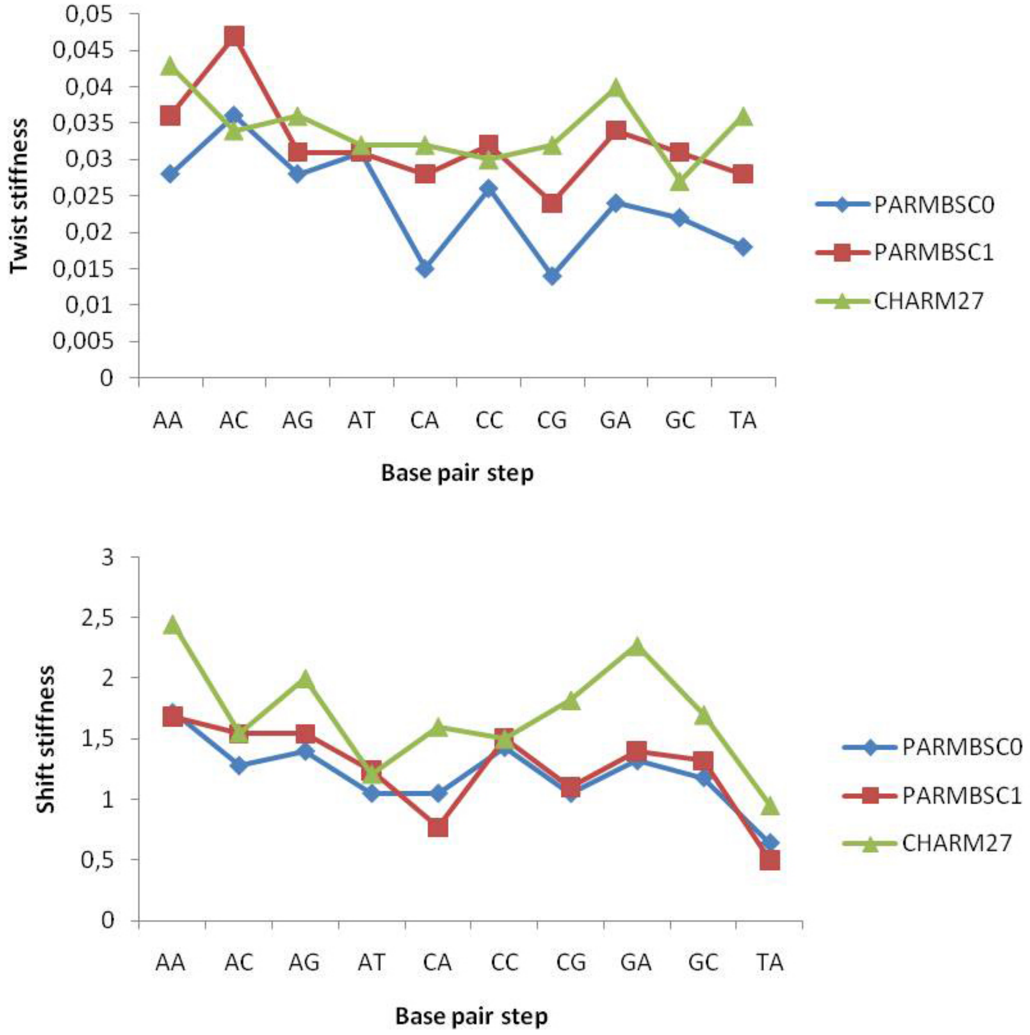
Supplementary Figure 17| Representative helical base pair step parameters in drug-DNA complexes. Time-averaged values associated to the DNA in complex with daunomycin (**a**) and distamycin A (**b**) in black compared with the original values from the X-ray structures (red, PDB id: 1D11 and 2DND for daunomycin and distamycin respectively).



Component	Volume (nm^3)	Dielectric constant parmbsc1 (ε_r)	Dielectric constant AFM (ε_r)
DNA	10.1 ± 0.2	8.0 ± 0.3	8.5 ± 1.4
DNA$_{\text{sugar–phosphate}}$	3.4 ± 0.2	19.9 ± 1.1	
DNA$_{\text{sugar–base}}$	7.6 ± 0.2	2.1 ± 0.2	

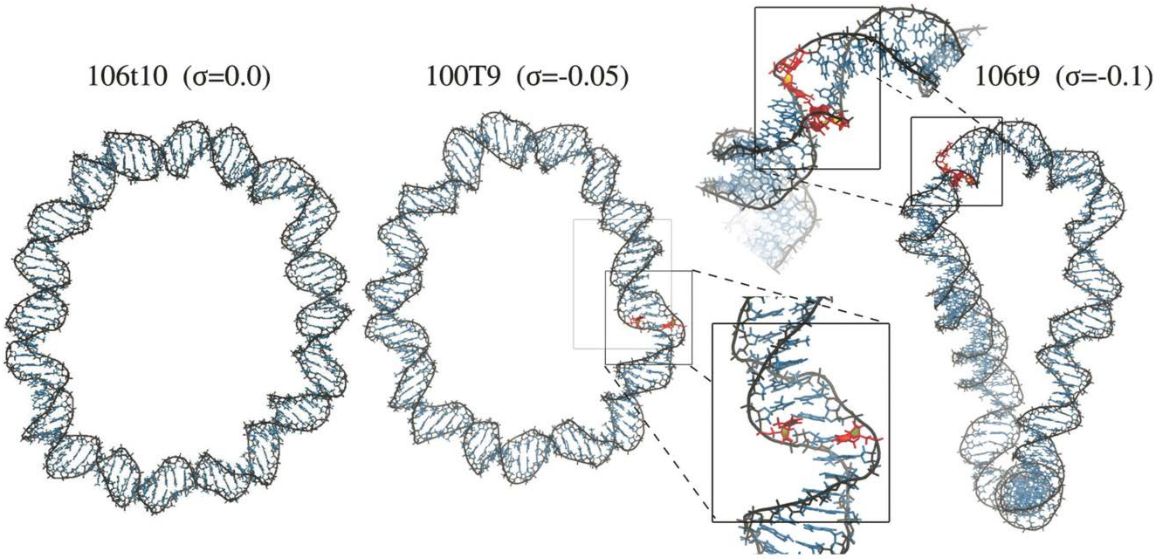
Figure 18| DNA dielectric constant. (**a**) Total dipole moment over time for 5 different replicas (100 ns each) taken from the microsecond long DDD simulation. (**b**) Accumulative mean square deviation of the dipole moment for the five replicas showing fairly good convergence after 30–40 ns. Values of whole DNA, sugar and phosphate groups, and sugar and base contributions are shown in the table below. See ref. 1 for the detailed procedure followed herein.

1. Cuervo, A., Dans, P. D.*et al. Proc. Natl. Acad. Sci.***111,** E3624–E3630 (2014).

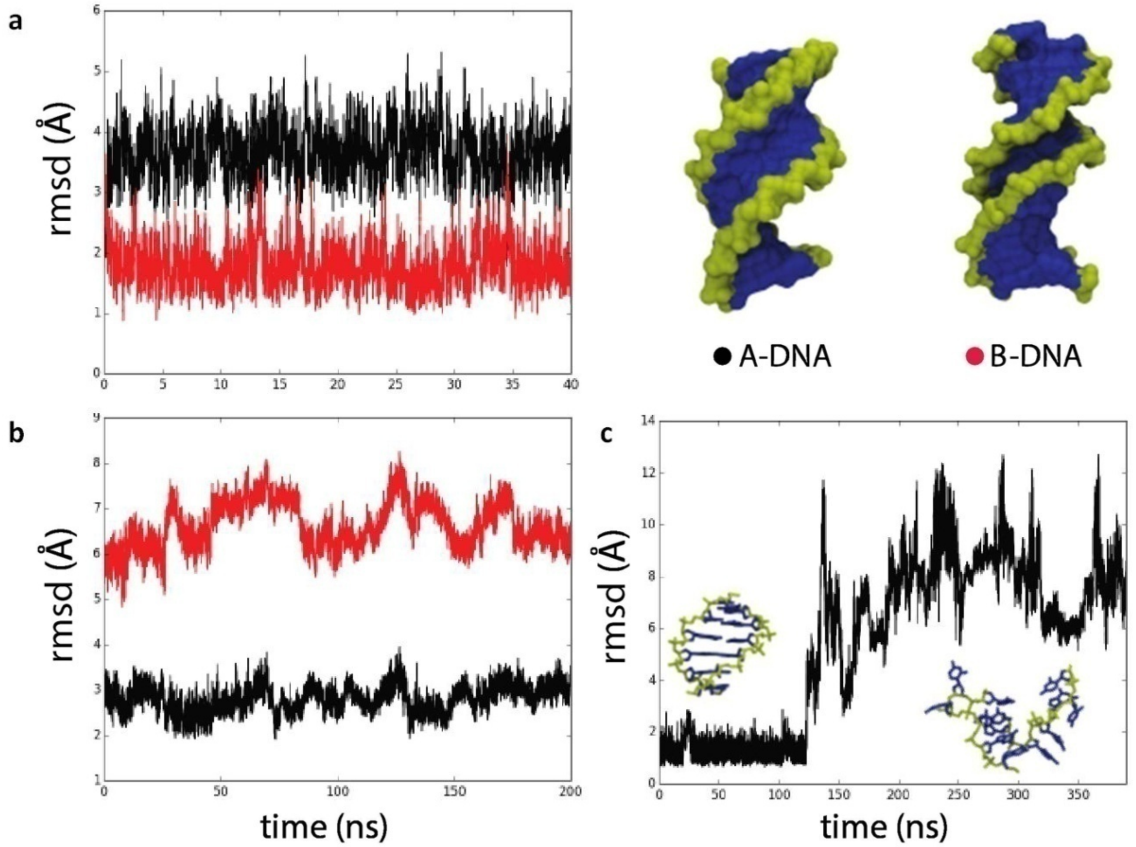


Supplementary Figure 19| Sequence dependent helical deformability. Variability of Twist (top) and Shift (bottom) stiffness constants for 10 unique base-steps. Parmbsc0 and CHARMM27 values are taken from ref 1.

1. Perez, A., Lankas, F., Luque, F.J. & Orozco, M. *Nucleic Acids Res.***36,** 2379–2394 (2008).

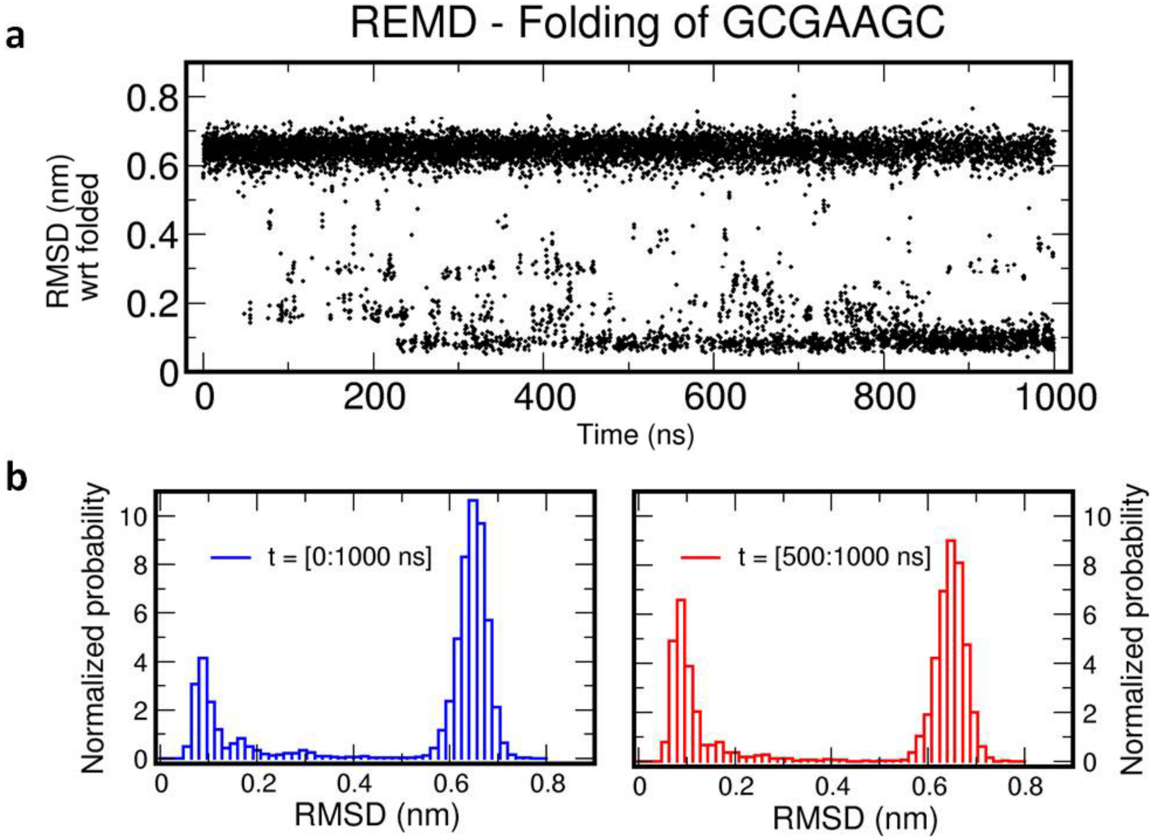


Supplementary Figure 20| Analysis of DNA minicircles. Final frames of the minicircles MD simulations. The secondary structure of the relaxed loop with 106 bp and 10 helical turns (106t10) remains intact, while the 2 negatively supercoiled circles show significant denaturalization. The 100 bp circle with 9 turns (100t9) presents 2 adjacent pyrimidine base-flipping towards the major groove, and the 106 bp circle with 9 turns (106t9), denature over multiple consecutive base pairs.



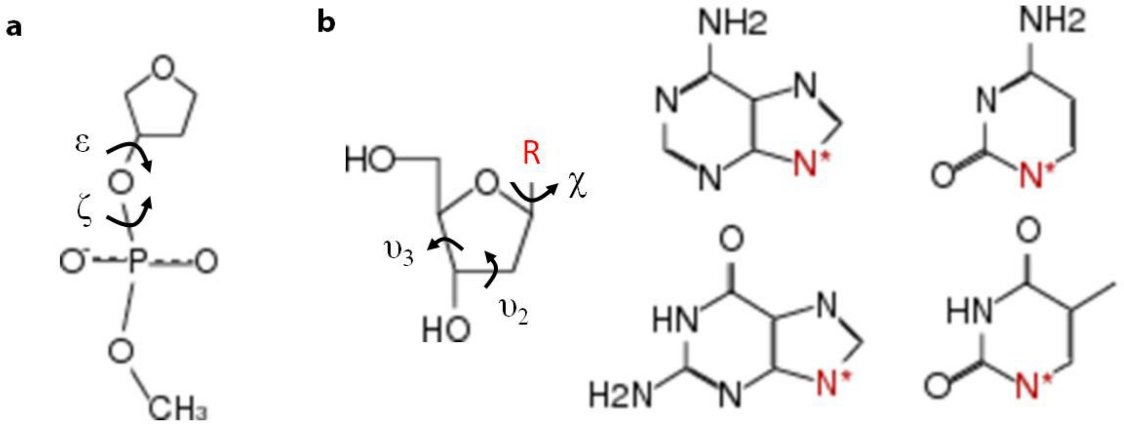
Supplementary Figure 21| MD simulations of conformational changes. (**a**) A to B transition simulation of DDD, where A-DNA form is presented in black with B-DNA in red. (**b**) Simulation of DDD in mixture of water and ethanol (see refs. 1 y 2 for additional discussion). (**c**) Unfolding of d(GGCGGC)$_2$ in 4 M pyridine water solution[3].

1. Soliva, R., Luque, F.J., Alhambra, C. & Orozco, M. *J. Biomol. Struct. Dyn.***17,** 89–99 (1999).
2. Ivanov, V.I., Minchenkova, L.E., Minyat, E.E., Frank-Kamenetskii, M.D. & Schyolkina, A.K. *J. Mol. Biol.***87,** 817–833 (1974).
3. Perez, A. & Orozco, M. *Angew. Chemie Int. Ed.***49,** 4805–4808 (2010).

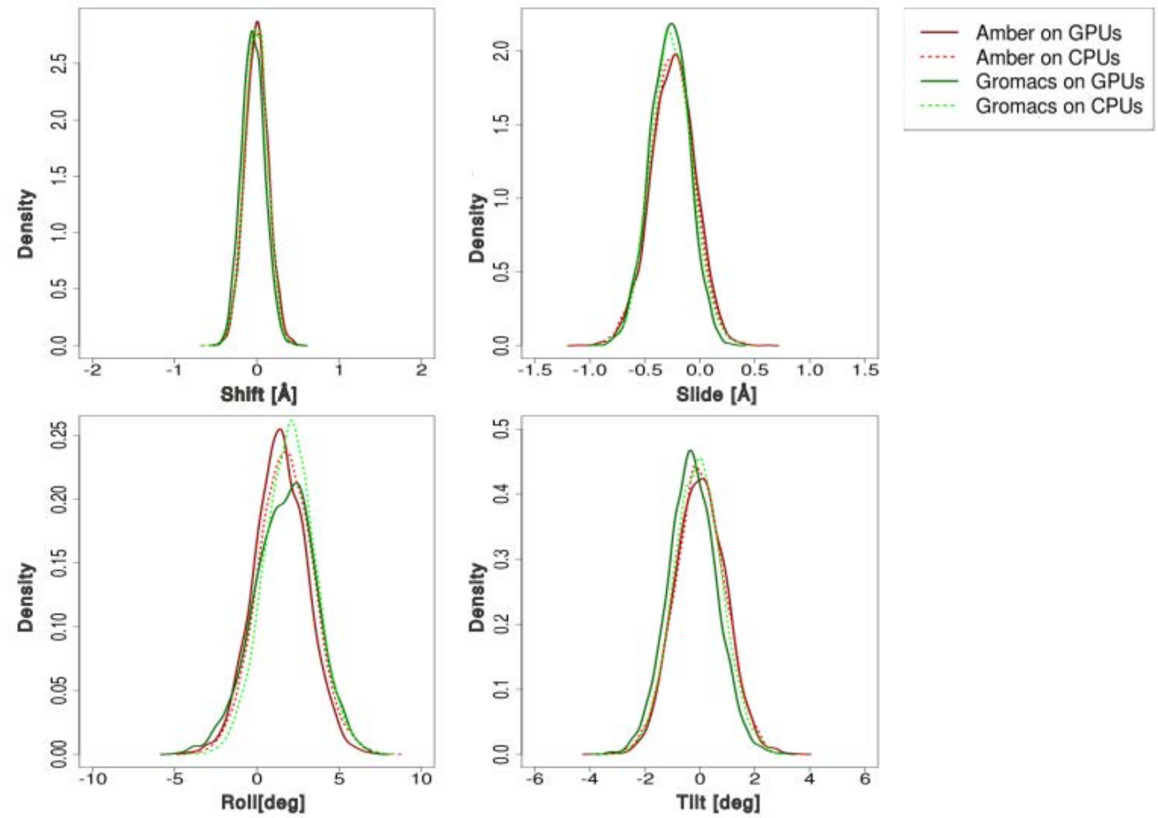


Supplementary Figure 22| Hairpin folding. Replica exchange MD (REMD) simulations of the folding of the small hairpin d(GCGAAGC) in water using parmbsc1 force-field. (**a**) RMSD with the respect to the folded state. (**b**) Probabilities of RMSDs in whole (blue) and second part (red) of microsecond runs of REMD. Structures are clearly recognizing the folded conformation and keeping it. For technical details see reference 1.

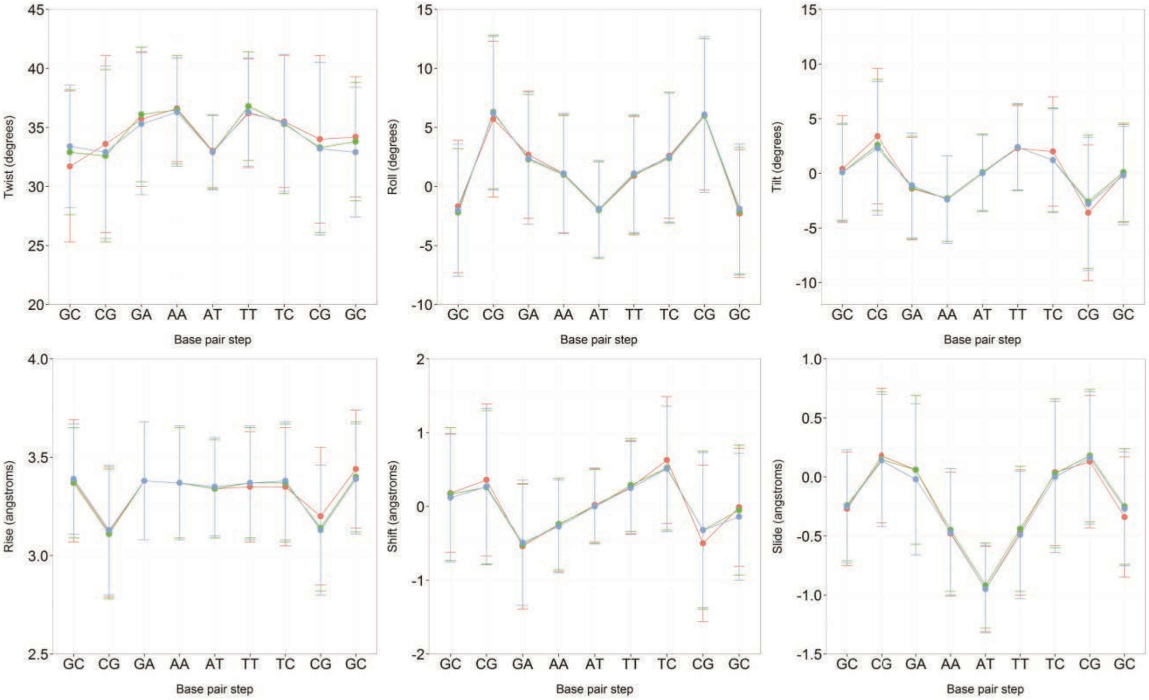
1. Portella, G., Orozco, M. *Angewandte chemie Int. Ed.*<b>49</b>, 7673–7676 (2010).



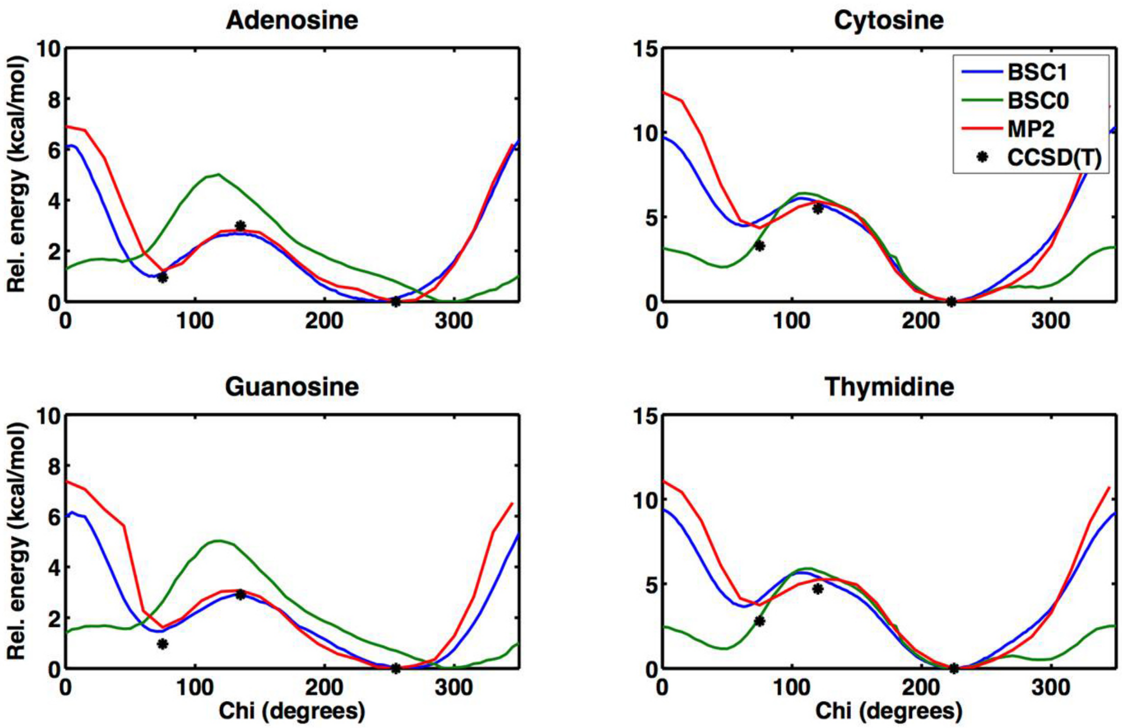
Supplementary Figure 23| Model compounds used in QM optimization. (**a**) Compound used for ε/ζ parameterization. (**b**) Compounds used for χ and sugar puckering parameterizations, where R represents the base, shown on the right.



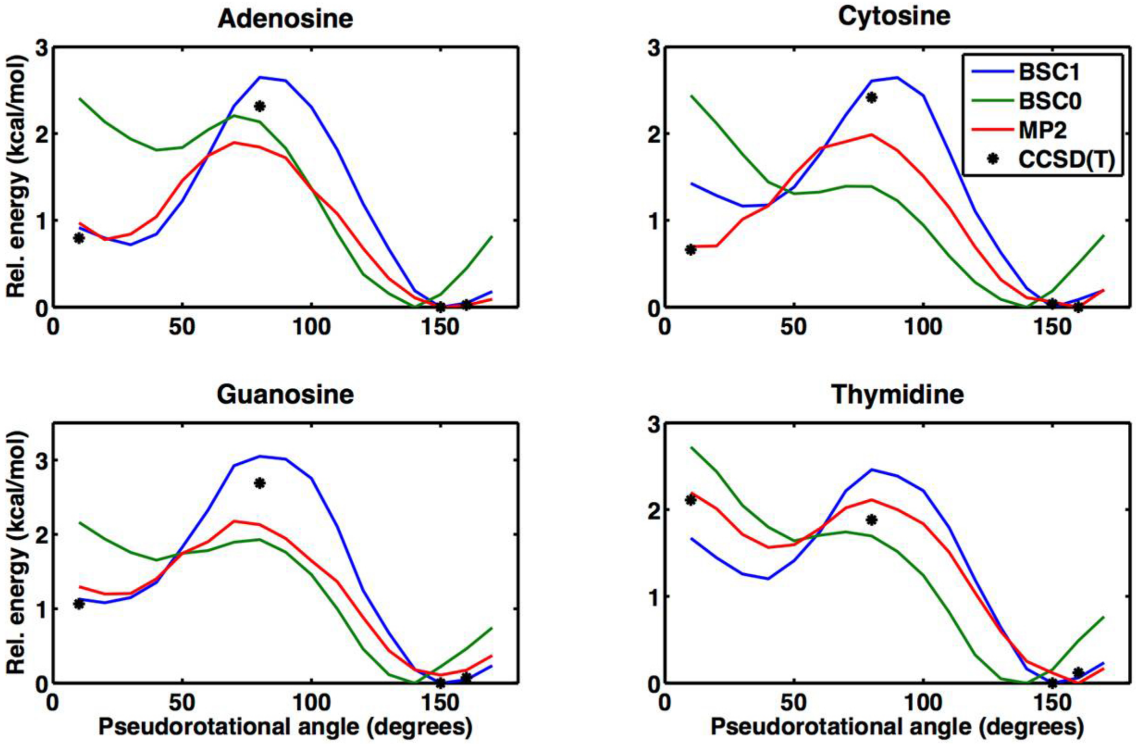
Supplementary Figure 24| Using DDD to compare different simulation engines. Normalized distributions of the helical parameters shift, slide, roll and tilt are shown for the four MD simulations (AMBER vs GROMACS, and GPU vs CPU codes). Due to the shortness of the simulation runs (100 ns), slight differences in roll angle can be detected using different MD engines.



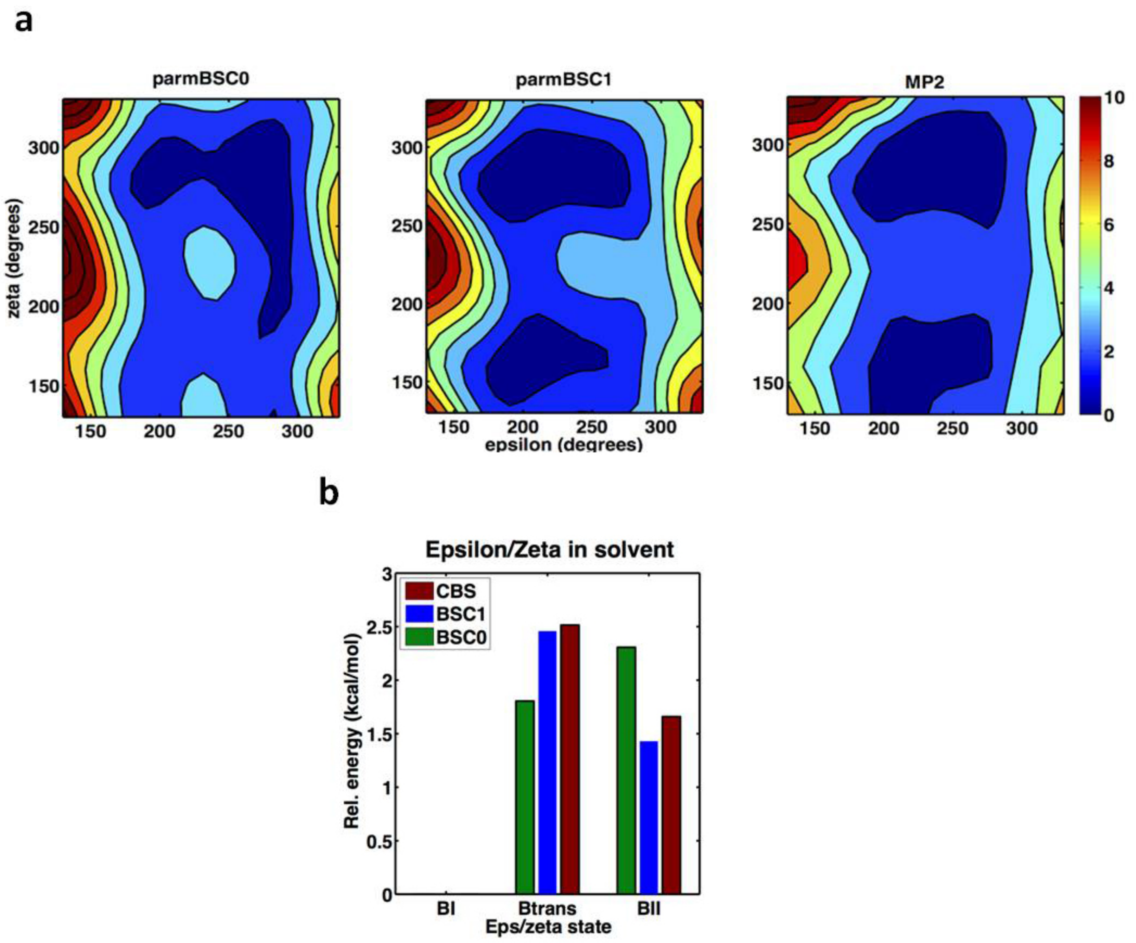
Supplementary Figure 25| Variation of helical parameters along the sequence for 2 µs of MD simulation of DDD with added salt (NaCl) concentrations: minimum Na^+ for neutrality (green), 150 mM (red) and 500 mM (blue). PME was used in all the cases.



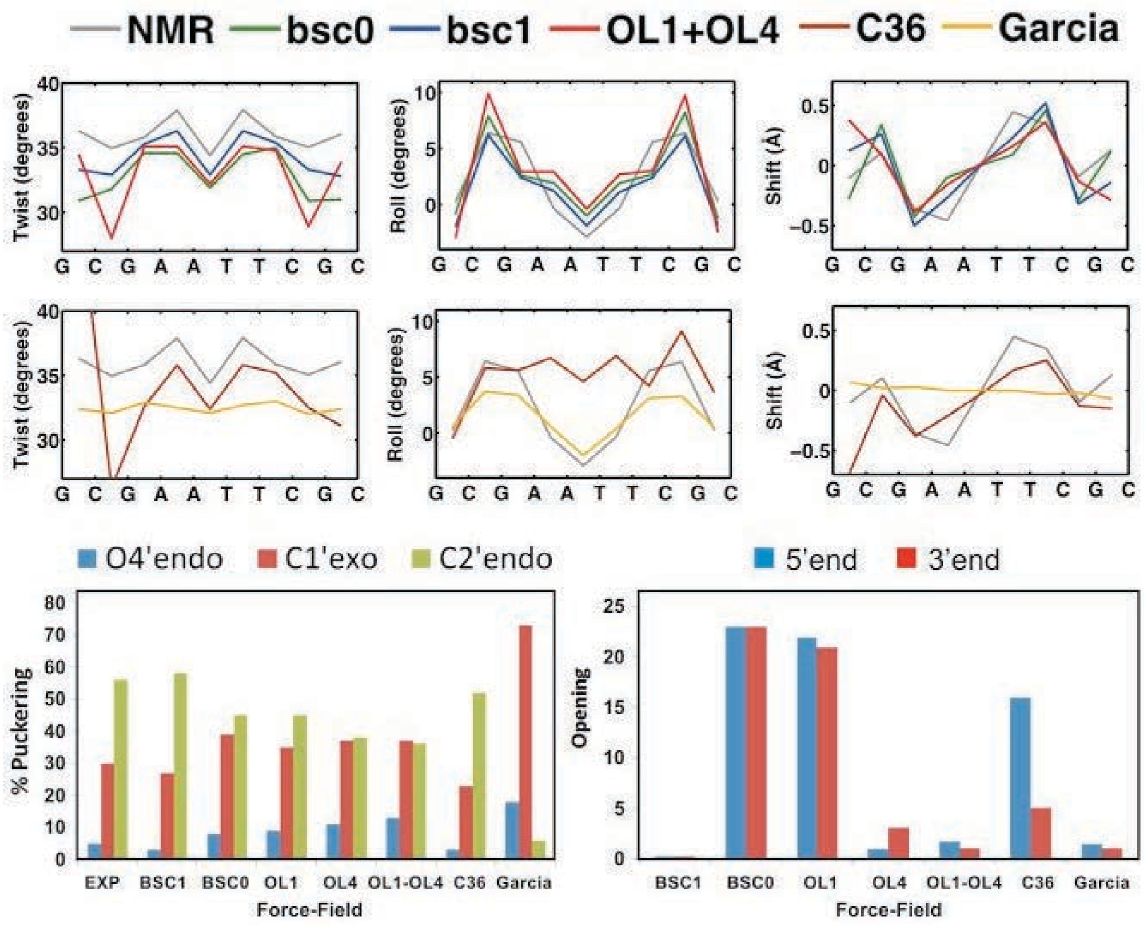
Supplementary Figure 26| Profiles of χ (chi) dihedral for 4 DNA bases in solution. Comparison of profiles obtained from QM using MP2/aug-cc-pVDZ (red) method with solvent corrections (Supplementary Notes), and PMF profiles using parmbsc0 (green) and parmbsc1 (blue) force-fields. Complete basis set (CBS) values for specific points are represented with a black dot.



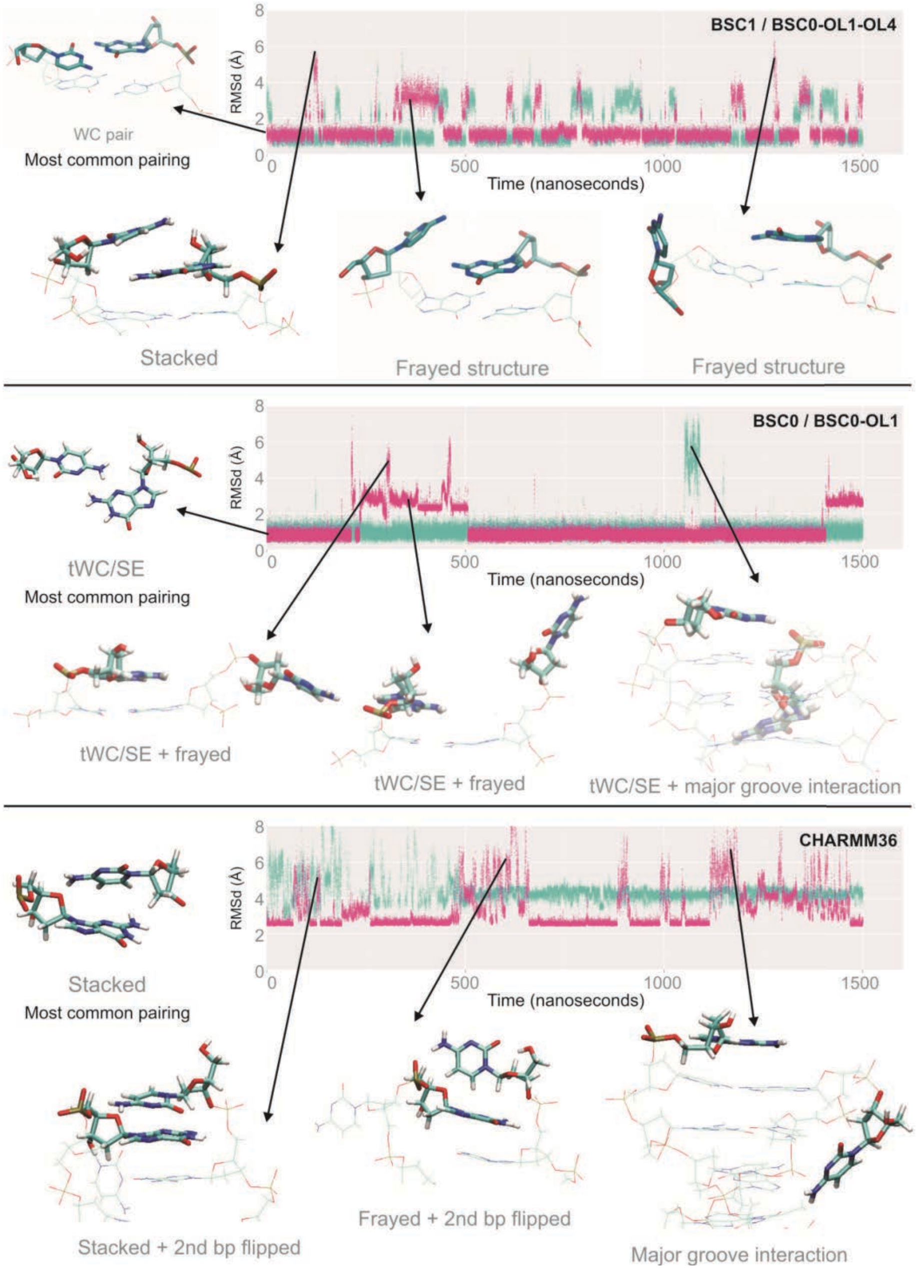
Supplementary Figure 27| Profiles of pseudorotational angle for 4 DNA bases in solution. Comparison of profiles obtained from QM using MP2/aug-cc-pVDZ (red) method with solvent corrections (Supplementary Notes), and PMF profiles using parmbsc0 (green) and parmbsc1 (blue) force-fields. Complete basis set (CBS) values for specific points are represented with a black dot.



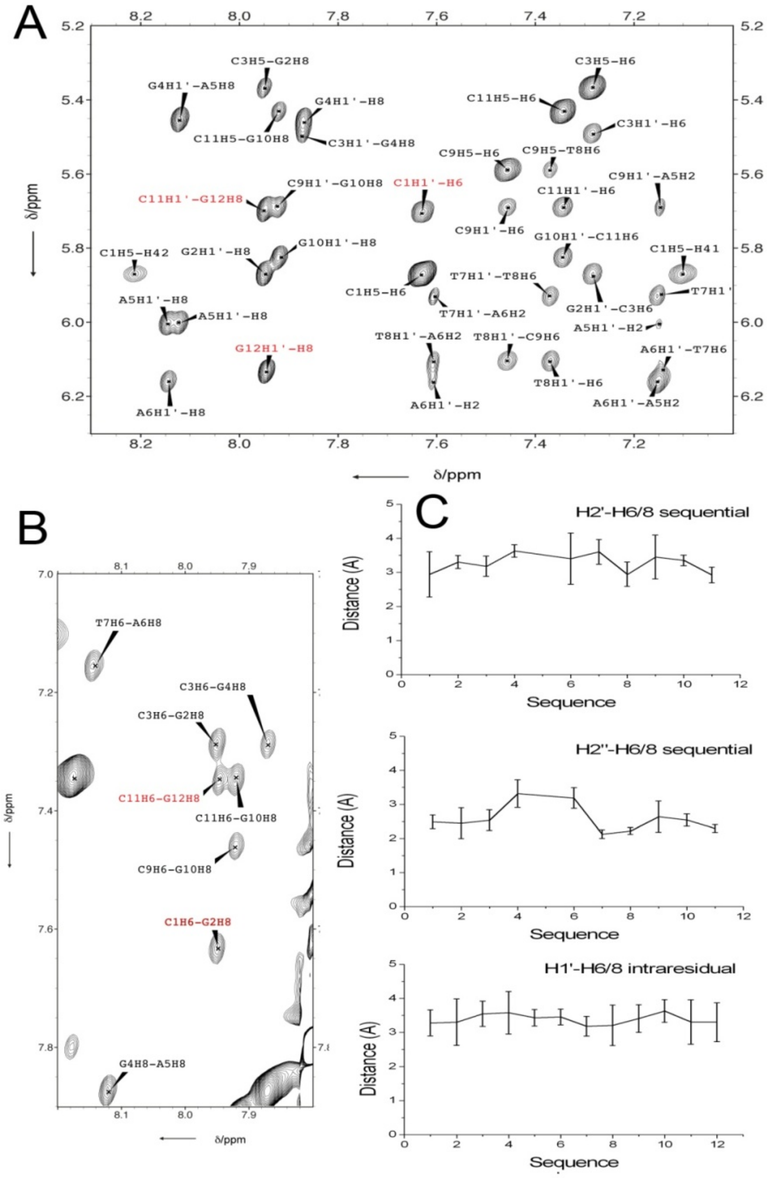
Supplementary Figure 28| ε/ζ (epsilon/zeta) profiles in solution. (**a**) Contour profiles of epsilon/zeta from QM calculations using MP2/aug-cc-pVDZ method (right), and PMF profiles using parmbsc0 (left) and parmbsc1 (middle) force-fields. Energies are given in kcal mol^{-1} and the color bar goes from blue (0 kcal mol^{-1}) to red (10 kcal mol^{-1}). (**b**) Values at key points of the profile comparing parmbsc0 (green), parmbsc1 (blue) and complete basis set (CBS) (dark red) values.



Supplementary Figure 29| Structural characteristics of DDD in MD simulations with different force-fields. First row variation of key helical coordinates along sequence in parmbsc0, parmbsc1 and parmbsc0-OL1+OL4 (those force-fields providing the best average parameters in **Supplementary Table 2**). Second raw correspond to force-fields providing less accurate average values in **Supplementary Table 2** (CHARMM36 and parmbsc0-Cheng-Garcia). In these two rows only the 10 mer segment is shown (to avoid dramatic scale bias in case of fraying of terminal bases), and only NMR results are used as reference (to make more clear the plots; note that nearly identical profiles are obtained from X-Ray (see **Fig. 1**)). The third row corresponds to the distribution of sugar puckering (taking as experimental reference the average of NMR and X-Ray structures) and the average opening at the terminal basis. The superior behavior of parmbsc1 is evident in all plots, as well the prevalence of fraying artifacts for some of the force-field, and the presence of non-negligible distortions in CHARMM36 and parmbsc0-CG trajectories, even for the central portion of the helix.



Supplementary Figure 30| Details of the evolution of the terminal base pairs. RMSd of the terminal base pairs (C1:G24 in pink and G12:C13 in cyan) along 1.5 μs of MD trajectories. Firs row: profiles for a force-field showing no fraying artifacts (but indeed frequent short-living openings) such a parmbsc1 (parmbsc0-OL1+OL4 and parmbsc0-OL4 provide similar profiles, while parmbsc0-CG (Cheng-Garcia) shows completely frozen terminal base pairs). Second row: profile for a force-field like parmbsc0 which suggest fraying and the formation of unusual contacts (parmbsc0-OL1 provides identical profiles) with tWC pairing and *syn* nucleotides. Third row: profiles obtained for CHARMM36, where despite the center of the duplex is well conserved terminal Watson-Crick pairings are mostly lost and substituted by a myriad of alternative contacts. In all cases structures sampled along specific time frames are shown.



Supplementary Figure 31| NOE data on the terminal base steps of DDD. A) H1´-aromatic region of the NOESY spectra of DDD (mixing time 200 ms, buffer conditions 125 mM NaCl, 25 mM sodium phosphate, pH 7, T = 25 º C). Some relevant cross-peaks involving terminal residues are labelled in red colour. B) Aromatic-aromatic region of the NOESY spectra (same experimental conditions). Note that NOE intensities involving terminal residues (i.e. C1H6-G2H8, C11H6-G12H8 in red) are not significantly lower than those involving central residues, indicating that the terminal bases remain stacked on top of their neighbours. C) Some experimental distances obtained from a full relaxation matrix analysis of the NOE data *vs* sequence. Sequential H2´-H6/8 and H2”-H6/8 do not exhibit dramatic changes for the terminal base steps, indicating that the fraying effect in these residues is not significant under these experimental conditions. All intra-residual H1´-H6/8 distances, including the terminal base residues, are around 3-4 Å, characteristic of glycosidic angle conformation in *anti*.

2 The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA

Supporting Data

The role of unconventional hydrogen bonds in determining BII propensities in B-DNA

Alexandra Balaceanu[1,2,&], Marco Pasi[3,4,&], Pablo D. Dans[1,2,&],
Adam Hospital[1,2], Richard Lavery[3,*], Modesto Orozco[1,2,5,*]

[1] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology.BaldiriReixac 10-12, Barcelona 08028, Spain.

[2] Joint BSC-IRB Program in Computational Biology, Institute for Research in Biomedicine. BaldiriReixac 10-12, Barcelona 08028, Spain.

[3]MMSB, Univ. Lyon I/CNRS UMR 5086, Institut de Biologie et Chimie des Protéines, 7 Passage du Vercors, Lyon 69367, France.

[4]School of Pharmacy and Centre for Biomolecular Sciences, University of Nottingham, University Park NG7 2RD, UK.

[5]Department of Biochemistry and Biomedicine, Faculty of Biology, University of Barcelona. Diagonal 643, Barcelona 08028, Spain.

[&] Equally contributing authors.

[*] Correspondence to: Prof. Richard Lavery (richard.lavery@ibcp.fr) or Prof. Modesto Orozco (modesto.orozco@irbbarcelona.org).

SUPPORTING TABLES

Table S1. PDB codes of the experimental structures used to analyze the presence of the C6/C8···O3' interaction in RpY, RpR, YpR and YpY steps.[a]

RpY steps

100D, 104D, 109D, 111D, 112D, 113D, 117D, 118D, 121D, 126D, 129D, 131D, 132D, 137D, 138D, 140D, 141D, 142D, 157D, 158D, 160D, 161D, 169D, 171D, 172D, 173D, 181D, 188D, 189D, 194D, 1AFZ, 1AHD, 1AL5, 1BDN, 1BDZ, 1BNA, 1BUT, 1BWT, 1CGC, 1CQO, 1CS2, 1D02, 1D13, 1D18, 1D30, 1D32, 1D46, 1D48, 1D63, 1D65, 1D68, 1D77, 1D80, 1D88, 1D89, 1D96, 1D98, 1DA4, 1DA5, 1DC0, 1DCV, 1DI2, 1DN9, 1DNM, 1DNO, 1DNT, 1DNX, 1DOU, 1DUF, 1DXN, 1F69, 1F6C, 1FHY, 1FIX, 1FKY, 1FKZ, 1FMQ, 1FMS, 1FQ2, 1FTD, 1FZX, 1G00, 1G14, 1G7Z, 1G80, 1GIP, 1H0M, 1HQ7, 1HRZ, 1HX4, 1HZ0, 1I0T, 1IH1, 1K9H, 1K9L, 1LAI, 1LEY, 1LP7, 1M18, 1MDY, 1MTG, 1MXJ, 1MXK, 1N1K, 1NAJ, 1NEV, 1ONM, 1OUP, 1P24, 1P25, 1P26, 1P3I, 1PBM, 1PG9, 1PGC, 1PRP, 1PT3, 1PYI, 1QAI, 1QCH, 1QP5, 1QXB, 1RC7, 1RVH, 1S2R, 1S32, 1SFU, 1SNH, 1T4X, 1TQR, 1UBD, 1UQB, 1UQF, 1UQG, 1V9G, 1VAQ, 1VT6, 1VT8, 1VTC, 1VTD, 1VTU, 1VZK, 1WOE, 1X2X, 1X2Y, 1X2Z, 1X30, 1XAM, 1YTB, 1YYK, 1YYO, 1Z7F, 1ZEX, 1ZEY, 1ZF1, 1ZF2, 1ZF5, 1ZFE, 1ZFG, 1ZFM, 1ZPH, 1ZPI, 203D, 222D, 227D, 237D, 240D, 250D, 252D, 257D, 260D, 271D, 281D, 287D, 289D, 295D, 2ANA, 2AOQ, 2B0K, 2B1C, 2B1D, 2B2B, 2B3E, 2BNA, 2D47, 2D94, 2D95, 2DAU, 2DBE, 2DND, 2DP7, 2DPB, 2DYW, 2FKC, 2FL3, 2GIG, 2GIH, 2GII, 2GVR, 2GYX, 2HKB, 2HKC, 2I2I, 2I5A, 2JXQ, 2JYK, 2K8T, 2K8U, 2KAR, 2KAS, 2KNK, 2KNL, 2KUZ, 2KYD, 2L7D, 2L8C, 2L8Q, 2L8U, 2L8W, 2LEV, 2LG2, 2LG3, 2M2C, 2MKN, 2MNB, 2MND, 2MO7, 2NLM, 2NPW, 2NQ0, 2NQB, 2QK9, 2R22, 2RT8, 2RVE, 2STT, 2STW, 2WCC, 2YKG, 302D, 303D, 321D, 328D, 330D, 334D, 348D, 349D, 353D, 355D, 360D, 368D, 369D, 370D, 371D, 372D, 393D, 394D, 395D, 396D, 398D, 3ADL, 3C25, 3D0P, 3E3Y, 3E40, 3E41, 3E43, 3E44, 3EBC, 3EQT, 3F8O, 3FL6, 3FQB, 3GCY, 3GDA, 3GNK, 3GOJ, 3IGM, 3IKT, 3KBD, 3KXB, 3L25, 3M9E, 3NJ7, 3ODA, 3OG8, 3OIE, 3QMB, 3QMC, 3QMD, 3QMG, 3QMH, 3QMI, 3SLP, 3TED, 3U05, 3U08, 3U0U, 3U2N, 3VYX, 3VYY, 3W97, 3ZD6, 3ZD7, 3ZQL, 401D, 404D, 405D, 414D, 420D, 428D, 434D, 435D, 440D, 442D, 443D, 445D, 453D, 455D, 466D, 4AGZ, 4AH0, 4AH1, 4ATK, 4BPB, 4BZT, 4BZU, 4BZV, 4C64, 4DY8, 4E60, 4EVV, 4F2X, 4GHL, 4HIV, 4HLY, 4HP3, 4IBU, 4IZQ, 4KBD, 4KWX, 4L24, 4MSB, 4MSR, 4NFO, 4NW3, 4PZI, 4RKG, 4U37, 4U8A, 4U8B, 4U8C, 5BNA, 7BNA, 8DRH, 9BNA, 9DNA, 116D, 119D, 124D, 197D, 1A84, 1AL9, 1AU5, 1B0S, 1BC7, 1BJD, 1BN9, 1BP8, 1D19, 1D28, 1D29, 1D42, 1D70, 1D78, 1D79, 1D82, 1DCW, 1DK9, 1EVP, 1FYM, 1ILC, 1JE8, 1K2J, 1K2K, 1L4J, 1LU5, 1LWA, 1M6G, 1NT8, 1NVN, 1NVY, 1QPH, 1RMX, 1RN9, 1RVI, 1SAA, 1SS7, 1SSV, 1SXQ, 1UQC, 1UQE, 1VFC, 1VJ4, 1VT9, 1VTA, 1VTB, 1X2O, 1X2S, 1X2U, 1X2V, 1ZEZ, 1ZF0, 1ZF7, 1ZG1, 1ZYF, 1ZYG, 1ZYH, 206D, 243D, 258D, 2ADY, 2AHI, 2BZF, 2DES, 2E1C, 2F8W, 2GB9, 2K0V, 2KTT, 2KY7, 2MF8, 2MNE, 2MNF, 2OG0, 2P7C, 2PKV, 2PL4, 2PLO, 2R2R, 2R2T, 2Z9O, 308D, 378D, 3AAF, 3DVO, 3DW9, 3G00, 3HS1, 3I1D, 3IXN, 3JXB, 3JXD, 3LPV, 3MKW, 3MKY, 3MKZ, 3R86, 3ZVN, 407D, 408D, 423D, 424D, 425D, 4D8J, 4IRI, 4LEY, 4LEZ, 4LLL, 4M95, 4NDH, 4OKL, 4R49, 4R4A, 4R4D, 1AGH, 1C4L, 1D20, 1DHH, 1DRN, 1K8J, 1KBD, 1N1N, 1R4I, 1ZBL, 2C5R, 2KBD, 2KDZ, 2ORH, 2QKK, 367D, 377D, 3ZQC, 4EZ2, 4GHA, 4I6Z, 4LG2, 167D, 182D, 198D, 1BUF, 1CVX, 1CVY, 1D23, 1D56, 1D57, 1D62, 1DN6, 1EOO, 1OSL, 1RSB, 1SY8, 1UQA, 1UQD, 1VRR, 1WQY, 1WQZ, 1Z3F, 1ZF6, 1ZFC, 1ZFF, 1ZFH, 1ZNS, 1ZTW, 224D, 245D, 2FJW, 2FJX, 2G1Z, 2IVH, 2LWG, 2LWH, 2R1J, 2WIW, 3ANA, 3EY0, 3FBD, 3FSI, 3FT6, 3JXC, 3MLN, 3MLO, 3MLP, 3UXW, 3ZPL, 4HW1, 4J2I, 4JBK, 4LNQ, 4OCD, 4RVE

RpR steps

100D, 111D, 116D, 126D, 132D, 137D, 138D, 158D, 160D, 161D, 167D, 169D, 187D, 188D, 189D, 1A84, 1AFZ, 1AGH, 1AU5, 1B0S, 1BC7, 1BCE, 1BN9, 1BP8, 1BUT, 1C4L, 1CGC, 1CVX, 1CVY, 1D02, 1D13, 1D80, 1DC0, 1DCV, 1DCW, 1DHH, 1DI2, 1DN6, 1DNM, 1DRN, 1EVP, 1F69, 1F6C, 1FHY, 1FIX, 1FYM, 1FZX, 1G00, 1G14, 1IH1, 1IKK, 1ILC, 1JE8, 1K8J, 1KBD, 1L4J, 1LAI, 1LU5, 1LWA, 1M18, 1M6G, 1MXJ, 1MXK, 1N1K, 1NEV, 1NT8, 1NVN, 1NVY, 1P24, 1P25, 1P26, 1P3I, 1PG9, 1PGC, 1PYI, 1QCU, 1QP5, 1QPH, 1RC7, 1RXB, 1S32, 1SDR, 1SS7, 1SSV, 1TQR, 1UBD, 1UQF, 1VAQ, 1VFC, 1VJ4, 1VT5, 1VT6, 1VT8, 1VT9, 1VTA, 1VTC, 1VTD, 1WQY, 1WQZ, 1YTB, 1ZEX, 1ZEY, 1ZEZ, 1ZF0, 1ZF1, 1ZF2, 1ZF5, 1ZF6, 1ZF7, 1ZF9, 1ZFC, 1ZFE, 1ZFF, 1ZFG, 1ZFH, 1ZFM, 1ZG1, 1ZNS, 1ZYF, 1ZYG, 1ZYH, 206D, 240D, 257D, 259D, 281D, 282D, 2A7E, 2ADY, 2AHI, 2ANA, 2B1C, 2B2B, 2BZF, 2C5R, 2D47, 2D94, 2D95, 2FKC, 2FL3, 2GIG, 2GIH, 2GII, 2HKB, 2HKC, 2IVH, 2K0V, 2KBD, 2KNK, 2KNL, 2L8C, 2L8U, 2L8W, 2LEV, 2LWG, 2LWH, 2MKN, 2NPW, 2NQ0, 2NQB, 2ORH, 2PKV, 2PL4, 2PLO, 2QHB, 2STT, 2STW, 2WIW, 2Z3X, 317D, 321D, 330D, 334D, 348D, 349D, 368D, 369D, 370D, 371D, 372D, 393D, 394D, 396D, 398D, 3AAF, 3ANA, 3C25, 3DVO, 3DW9, 3E3Y, 3E40, 3E41, 3E43, 3E44, 3EBC, 3FBD, 3GNK, 3GOJ, 3HS1, 3IXN, 3KBD, 3KXB, 3LPV, 3M9E, 3MKW, 3MKY, 3MKZ, 3MLN, 3MLO, 3MLP, 3NJ7, 3ODA, 3OG8, 3QMB, 3QMC, 3QMD, 3QMG, 3QMH, 3QMI, 3R86, 3SSF, 3W97, 401D, 407D, 408D, 414D, 423D, 424D, 425D, 440D, 4EVV, 4EZ2, 4HIV, 4HP3, 4IBU, 4IRI, 4IZQ, 4JBK, 4KBD, 4KWX, 4LNQ, 4MSB, 4MSR, 4NW3, 4OKL, 4PZI, 4R49, 4R4A, 4R4D, 4RVE, 8DRH, 9DNA, 109D, 112D, 113D, 119D, 129D, 140D, 141D, 142D, 157D, 171D, 172D, 173D, 179D, 182D, 196D, 198D, 1AHD, 1BCB, 1BJD, 1BNA, 1BWT, 1D23, 1D28, 1D29, 1D30, 1D46, 1D56, 1D57, 1D82, 1D89, 1D93, 1DK9, 1DOU, 1DUF, 1EOO, 1FMQ, 1FMS, 1FQ2, 1FTD, 1GIP, 1H0M, 1K9H, 1K9L, 1LEY, 1MDY, 1MTG, 1N1N, 1NAJ, 1OSL, 1OUP, 1PRP, 1PT3, 1QXB, 1R4I, 1RMX, 1RN9, 1SAA, 1SY8, 1UQD, 1VRR, 1VZK, 1X2O, 1X2S, 1X2U, 1X2V, 1X2X, 1X2Y, 1X2Z, 1X30, 1YYK, 1YYO, 1Z3F, 1Z7F, 1ZBL, 1ZPH, 1ZPI, 224D, 227D, 245D, 251D, 271D, 287D, 289D, 2AOQ, 2B0K, 2B1D, 2B3E, 2BNA, 2DAU, 2DBE, 2DP7, 2DPB, 2DYW, 2E1C, 2FJW, 2FJX, 2GVR, 2GYX, 2I2I, 2I5A, 2JXQ, 2JYK, 2K8T, 2K8U, 2KAR, 2KAS, 2KDZ, 2L7D, 2LG2, 2LG3, 2MF8, 2MNB, 2MND, 2MNE, 2MNF, 2NLM, 2QK9, 2QKK, 2R1J, 2R22, 2RVE, 2Z9O, 302D, 303D, 307D, 328D, 355D, 360D, 3D0P, 3FT6, 3IKT, 3JXB, 3JXC, 3JXD, 3OIE, 3SLP, 3U05, 3U08, 3U0U, 3U2N, 3UXW, 3ZPL, 3ZQC, 3ZQL, 404D, 405D, 420D, 428D, 442D, 443D, 445D, 453D, 455D, 4AGZ, 4C64, 4D8J, 4GHA, 4GHL, 4HLY, 4I6Z, 4L24, 4LEY, 4LEZ, 4LG2, 4LLL, 4NDH, 4NFO, 4U8A, 4U8B, 4U8C, 5BNA, 7BNA, 9BNA, 1BDZ, 1CS2, 1D98, 1DA4, 1DA5, 1FKY, 1FKZ, 1G7Z, 1ONM, 1SK5, 1SXQ, 1UQB, 250D, 2KYD, 2OG0, 2P7C, 2R2R, 2R2T, 2WCC, 434D, 435D, 466D, 4ATK, 4U37, 121D, 194D, 1AL5, 1BDN, 1BUF, 1CQO, 1D62, 1D63, 1D65, 1D70, 1D77, 1DXN, 1HQ7, 1HRZ, 1PLY, 1RNA, 1RVH, 1RVI, 1S2R, 1YYW, 1ZTW, 237D, 252D, 2DND, 2G1Z, 2KTT, 2KUZ, 2KY7, 2MO7, 3FSI, 4AH0, 4AH1, 4HW1, 4J2I, 4OCD

YpR steps
100D, 104D, 109D, 111D, 112D, 113D, 116D, 117D, 118D, 119D, 121D, 129D, 131D, 132D, 137D, 138D, 157D, 160D, 161D, 169D, 171D, 179D, 181D, 187D, 188D, 189D, 194D, 196D, 197D, 198D, 1AGH, 1AL5, 1AL9, 1B0S, 1BDN, 1BJD, 1BNA, 1BP8, 1BWT, 1C4L, 1CGC, 1CQO, 1D13, 1D19, 1D20, 1D23, 1D28, 1D29, 1D30, 1D32, 1D46, 1D48, 1D56, 1D57, 1D63, 1D65, 1D68, 1D77, 1D80, 1D88, 1D89, 1D96, 1D98, 1DCV, 1DCW, 1DI2, 1DN9, 1DNM, 1DNO, 1DNT, 1DNX, 1DOU, 1DUF, 1DXN, 1F69, 1F6C, 1FHY, 1FIX, 1FMQ, 1FMS, 1FQ2, 1FTD, 1FZX, 1G00, 1G14, 1G7Z, 1G80, 1GIP, 1HQ7, 1HX4, 1HZ0, 1I0T, 1IH1, 1ILC, 1JE8, 1K2J, 1K2K, 1K9H, 1K9L, 1L4J, 1LAI, 1LEY, 1LP7, 1M18, 1M6G, 1N1K, 1N1N, 1NAJ, 1NEV, 1NT8, 1NVN, 1NVY, 1ONM, 1OSL, 1OUP, 1P24, 1P25, 1P26, 1P3I, 1PBM, 1PRP, 1PT3, 1PYI, 1QP5, 1QPH, 1QXB, 1RC7, 1RMX, 1RN9, 1RVI, 1RXB, 1S2R, 1S32, 1SAA, 1SFU, 1SNH, 1SS7, 1SSV, 1T4X, 1UBD, 1UQC, 1UQD, 1UQE, 1UQF, 1UQG, 1V9G, 1VT5, 1VT6, 1VT8, 1VTC, 1VTD, 1VTU, 1VZK, 1WOE, 1XAM, 1YTB, 1YYK, 1YYO, 1Z3F, 1ZEX, 1ZEY, 1ZEZ, 1ZF0, 1ZF1, 1ZF2, 1ZF5, 1ZF7, 1ZF9, 1ZFC, 1ZFE, 1ZFF, 1ZFG, 1ZFM, 1ZG1, 1ZNS, 1ZPH, 1ZPI, 1ZYG, 203D, 206D, 222D, 227D, 237D, 240D, 243D, 250D, 251D, 252D, 257D, 259D, 260D, 271D, 282D, 287D, 289D, 295D, 2ADY, 2AHI, 2B0K, 2B1C, 2B1D, 2B2B, 2B3E, 2BNA, 2C5R, 2D47, 2D94, 2D95, 2DAU, 2DBE, 2DES, 2DND, 2DP7, 2DPB, 2DYW, 2F8W, 2FKC, 2FL3, 2GB9, 2GIG, 2GIH, 2GII, 2GVR, 2GYX, 2HKB, 2HKC, 2I2I, 2I5A, 2IVH, 2JXQ, 2JYK, 2K8T, 2K8U, 2KAR, 2KAS, 2KDZ, 2KNK, 2KNL, 2KUZ, 2L7D, 2L8C, 2L8Q, 2L8U, 2L8W, 2LEV, 2LG2, 2LG3, 2M2C, 2MF8, 2MKN, 2MNB, 2MND, 2MNE, 2MNF, 2MO7, 2NLM, 2NQB, 2ORH, 2QK9, 2QKK, 2R22, 2RT8, 2RVE, 2STT, 2STW, 2WIW, 2YKG, 302D, 303D, 308D, 321D, 328D, 330D, 348D, 349D, 353D, 355D, 360D, 367D, 368D, 369D, 370D, 371D, 372D, 377D, 393D, 394D, 395D, 396D, 3ADL, 3C25, 3D0P, 3DVO, 3DW9, 3E3Y, 3E40, 3E41, 3E43, 3E44, 3EBC, 3EQT, 3F8O, 3FBD, 3FL6, 3FQB, 3FT6, 3G00, 3GCY, 3GDA, 3GNK, 3GOJ, 3HS1, 3I1D, 3IKT, 3IXN, 3JXB, 3JXD, 3KXB, 3L25, 3OG8, 3OIE, 3QMB, 3QMC, 3QMD, 3QMG, 3QMH, 3QMI, 3R86, 3SLP, 3U05, 3U08, 3U0U, 3U2N, 3UXW, 3VYX, 3VYY, 3ZD6, 3ZD7, 3ZQC, 3ZQL, 401D, 414D, 423D, 424D, 425D, 428D, 442D, 443D, 445D, 453D, 455D, 4AGZ, 4AH0, 4AH1, 4ATK, 4BPB, 4BZV, 4C64, 4D8J, 4DY8, 4E60, 4EVV, 4EZ2, 4F2X, 4GHA, 4HIV, 4HLY, 4HP3, 4IBU, 4IRI, 4KWX, 4L24, 4LG2, 4M95, 4MSB, 4MSR, 4NW3, 4OKL, 4PZI, 4R49, 4R4A, 4R4D, 4RKG, 4U8A, 4U8B, 4U8C, 5BNA, 7BNA, 8DRH, 9BNA, 9DNA, 124D, 126D, 158D, 167D, 182D, 1AFZ, 1AHD, 1BC7, 1BDZ, 1BN9, 1BUF, 1BUT, 1CVX, 1CVY, 1D02, 1D18, 1D62, 1D78, 1D79, 1DA4, 1DA5, 1DC0, 1DK9, 1EVP, 1FKY, 1FKZ, 1H0M, 1HRZ, 1K8J, 1KBD, 1LWA, 1MDY, 1MXJ, 1PG9, 1PGC, 1QAI, 1QCH, 1R4I, 1RVH, 1SY8, 1TQR, 1UQA, 1UQB, 1VAQ, 1VTB, 1WQY, 1WQZ, 1X2O, 1X2S, 1X2U, 1X2V, 1X2X, 1X2Y, 1X2Z, 1X30, 1Z7F, 1ZBL, 1ZF6, 1ZYF, 1ZYH, 224D, 245D, 258D, 281D, 2AOQ, 2E1C, 2KBD, 2KTT, 2KY7, 2NPW, 2NQ0, 2P7C, 2R1J, 2WCC, 2Z9O, 307D, 334D, 398D, 3IGM, 3JXC, 3KBD, 3M9E, 3MKW, 3MKY, 3MKZ, 3MLN, 3MLO, 3MLP, 3NJ7, 3ODA, 3TED, 3W97, 3ZPL, 3ZVN, 405D, 407D, 408D, 420D, 4BZT, 4BZU, 4GHL, 4I6Z, 4IZQ, 4KBD, 4LEY, 4LEZ, 4LNQ, 4NFO, 140D, 141D, 142D, 1A84, 1AU5, 1CS2, 1D70, 1DN6, 1LU5, 2BZF, 2FJW, 2FJX, 2K0V, 2OG0, 3LPV, 1D42, 1D82, 1D93, 1EOO, 1FYM, 1IKK, 1RSB, 1SK5, 1VFC, 1VJ4, 1VRR, 1VT9, 1VTA, 1ZFH, 1ZTW, 2A7E, 2LWG, 2LWH, 2PKV, 2PL4, 2PLO, 2QHB, 2R2R, 2R2T, 317D, 378D, 3AAF, 3EY0, 3FSI, 4HW1, 4J2I, 4JBK, 4LLL, 4NDH, 4RVE

YpY steps
100D, 116D, 126D, 132D, 137D, 138D, 140D, 141D, 142D, 158D, 160D, 161D, 167D, 187D, 188D, 189D, 1A84, 1AHD, 1AU5, 1B0S, 1BP8, 1BUT, 1CGC, 1CVX, 1CVY, 1D02, 1D13, 1D20, 1D62, 1DC0, 1DCV, 1DCW, 1DI2, 1EVP, 1F69, 1F6C, 1FHY, 1FIX, 1FKY, 1FKZ, 1FYM, 1G00, 1HX4, 1HZ0, 1IH1, 1IKK, 1ILC, 1JE8, 1K8J, 1KBD, 1L4J, 1LAI, 1LU5, 1M18, 1M6G, 1MXK, 1N1K, 1NT8, 1NVN, 1NVY, 1P24, 1P25, 1P26, 1P3I, 1PG9, 1PGC, 1PYI, 1QP5, 1QPH, 1R4I, 1RC7, 1RXB, 1S32, 1SS7, 1SSV, 1UBD, 1UQF, 1VAQ, 1VJ4, 1VT5, 1VT6, 1VT8, 1VT9, 1VTA, 1VTC, 1VTD, 1WQY, 1WQZ, 1ZBL, 1ZEX, 1ZEY, 1ZEZ, 1ZF0, 1ZF1, 1ZF2, 1ZF5, 1ZF6, 1ZF7, 1ZF9, 1ZFC, 1ZFE, 1ZFF, 1ZFG, 1ZFH, 1ZFM, 1ZG1, 1ZNS, 1ZYF, 1ZYG, 1ZYH, 240D, 257D, 259D, 281D, 282D, 2A7E, 2ADY, 2AHI, 2ANA, 2B1C, 2B2B, 2C5R, 2D47, 2D94, 2D95, 2FKC, 2FL3, 2GIG, 2GIH, 2GII, 2HKB, 2HKC, 2IVH, 2K0V, 2K8T, 2K8U, 2KAR, 2KAS, 2KBD, 2KNK, 2KNL, 2L8C, 2L8U, 2L8W, 2LG2, 2LG3, 2LWG, 2LWH, 2MF8, 2MKN, 2NPW, 2NQ0, 2NQB, 2ORH, 2PKV, 2PL4, 2PLO, 2QK9, 2QKK, 2RT8, 2STT, 2STW, 2WIW, 317D, 321D, 330D, 334D, 348D, 349D, 353D, 368D, 369D, 370D, 371D, 372D, 393D, 394D, 396D, 3AAF, 3ANA, 3C25, 3DVO, 3DW9, 3E3Y, 3E40, 3E41, 3E43, 3E44, 3EBC, 3FBD, 3GNK, 3GOJ, 3HS1, 3IXN, 3KBD, 3KXB, 3LPV, 3M9E, 3MKW, 3MKY, 3MKZ, 3MLN, 3MLO, 3MLP, 3NJ7, 3ODA, 3OG8, 3QMB, 3QMC, 3QMD, 3QMG, 3QMH, 3QMI, 3R86, 3TED, 3W97, 3ZQL, 401D, 407D, 408D, 414D, 423D, 424D, 425D, 434D, 435D, 440D, 466D, 4EZ2, 4HP3, 4IBU, 4IRI, 4IZQ, 4JBK, 4KBD, 4KWX, 4LNQ, 4M95, 4MSB, 4MSR, 4NW3, 4OKL, 4PZI, 4R49, 4R4A, 4R4D, 4RVE, 4U37, 9DNA, 119D, 172D, 173D, 196D, 1BDZ, 1CS2, 1D82, 1D93, 1DA4, 1DA5, 1DNM, 1EOO, 1G7Z, 1H0M, 1K9H, 1K9L, 1LP7, 1MDY, 1MTG, 1N1N, 1ONM, 1OSL, 1RMX, 1RN9, 1SK5, 1TQR, 1UQB, 1VRR, 1X2O, 1X2S, 1X2U, 1X2V, 1X2X, 1X2Y, 1X2Z, 1X30, 1ZTW, 250D, 251D, 2B1D, 2FJW, 2FJX, 2JYK, 2KUZ, 2L8Q, 2M2C, 2MNE, 2MNF, 2OG0, 2R1J, 2RVE, 2Z9O, 3FSI, 3IKT, 3JXB, 3JXC, 3JXD, 3SLP, 3ZPL, 479D, 4ATK, 4I6Z, 4LEY, 4LEZ, 4LLL, 4NDH, 109D, 124D, 129D, 171D, 179D, 182D, 198D, 1BN9, 1BNA, 1BWT, 1D23, 1D28, 1D29, 1D30, 1D46, 1D56, 1D57, 1D77, 1DOU, 1DUF, 1FMQ, 1FMS, 1FQ2, 1FTD, 1GIP, 1LEY, 1LWA, 1NAJ, 1OUP, 1PRP, 1PT3, 1SAA, 1SY8, 1UQD, 1VZK, 1Z3F, 1ZPH, 1ZPI, 224D, 227D, 245D, 287D, 289D, 2AOQ, 2B0K, 2B3E, 2BNA, 2DAU, 2DBE, 2DYW, 2E1C, 2GVR, 2GYX, 2I2I, 2I5A, 2L7D, 2MNB, 2MND, 2MO7, 2NLM, 2WCC, 302D, 303D, 328D, 355D, 360D, 3D0P, 3FT6, 3OIE, 3U05, 3U08, 3U0U, 3U2N, 3UXW, 3ZVN, 428D, 442D, 443D, 445D, 453D, 455D, 4AGZ, 4C64, 4D8J, 4EVV, 4HLY, 4L24, 4U8A, 4U8B, 4U8C, 5BNA, 7BNA, 8DRH, 9BNA, 111D, 112D, 113D, 121D, 194D, 1BJD, 1BUF, 1CQO, 1D63, 1D65, 1D80, 1DXN, 1HQ7, 1QXB, 1RVH, 1RVI, 1S2R, 1SNH, 1SXQ, 1VFC, 1YTB, 237D, 252D, 2DND, 2G1Z, 2KDZ, 2KTT, 2KY7, 2LEV, 2P7C, 2QHB, 2R2R, 2R2T, 3G00, 3ZQC, 4AH0, 4AH1, 4HW1, 4J2I, 4OCD

[a] For RpY steps we analyzed 554 experimental structures containing information about 3,991 dinucleotides (37% GpC, 18% GpT, 19% ApC, and 26% ApT). Similarly, for RpR steps the total number of nucleotide analyzed was 3,649 (26% GpG, 21% ApG, 21% GpA, and 32% ApA) coming from 484 structures of the PDB, for YpR steps we analyzed 560 structures and 4,161 dinuleotides (15% TpG, 44% CpG, 19% TpA, and 22% CpA) and for YpY steps results are obtained from 465 structures and 3,162 dinucloetides (30% TpT, 16% CpT, 27% TpC, and 27% CpC).

Table S2. BI/BII percentages and the C6/C8···O3' average distance for all the bps calculated from crystal structures.[a]

bps	state	%	mean C6/C8···O3' distance (Å)	s.d. C6/C8···O3' distance (Å)
			RpR steps	
GG	BI	91.6	5.24	0.48
	BII	8.4	3.63	0.39
GA	BI	91.1	5.11	0.40
	BII	8.9	3.64	0.40
AG	BI	96.3	5.29	0.44
	BII	3.7	4.03	0.58
AA	BI	97.5	5.15	0.37
	BII	2.5	3.91	0.45
			RpY steps	
GC	BI	90.0	5.03	0.39
	BII	10.0	3.48	0.38
GT	BI	97.6	5.32	0.37
	BII	2.4	4.15	0.31
AC	BI	97.1	5.15	0.40
	BII	2.9	3.80	0.35
AT	BI	99.2	5.30	0.38
	BII	0.8	3.95	0.40
			YpR steps	
CG	BI	93.3	5.28	0.57
	BII	6.7	3.73	0.45
CA	BI	88.9	5.19	0.55
	BII	11.1	3.48	0.34
TG	BI	88.5	5.21	0.56
	BII	11.5	3.50	0.39
TA	BI	91.7	5.20	0.46
	BII	8.3	3.59	0.43
			YpY steps	
CC	BI	95.4	5.10	0.39
	BII	4.6	3.72	0.41
CT	BI	95.7	5.25	0.35
	BII	4.3	3.83	0.37
TC	BI	97.8	5.10	0.35
	BII	2.2	3.88	0.36
TT	BI	99.4	5.20	0.29
	BII	0.6	3.64	0.58

SUPPORTING FIGURES

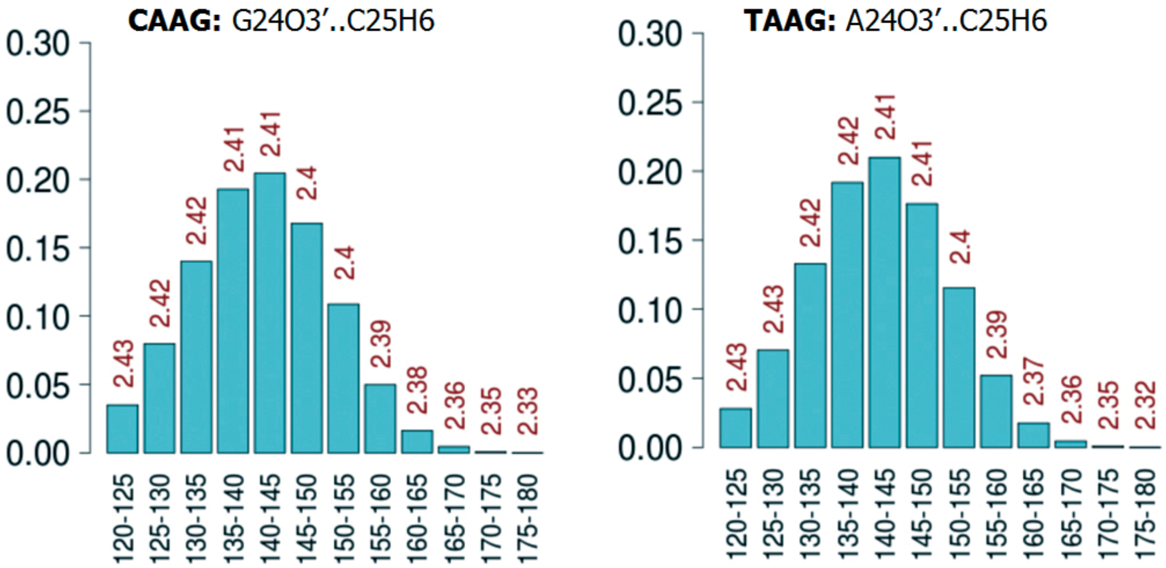


Figure S1. Distribution of C6-H6···O3’ H-bond angles in RpY interactions for systems simulated with the new parmBSC1 force field for nucleic acids. Structures with H6···O3’ bond distances < 2.5 Å and C6-H6···O3’ angles > 120° were selected and a histogram was built. The mean distance of the corresponding set is given above each bin bar.

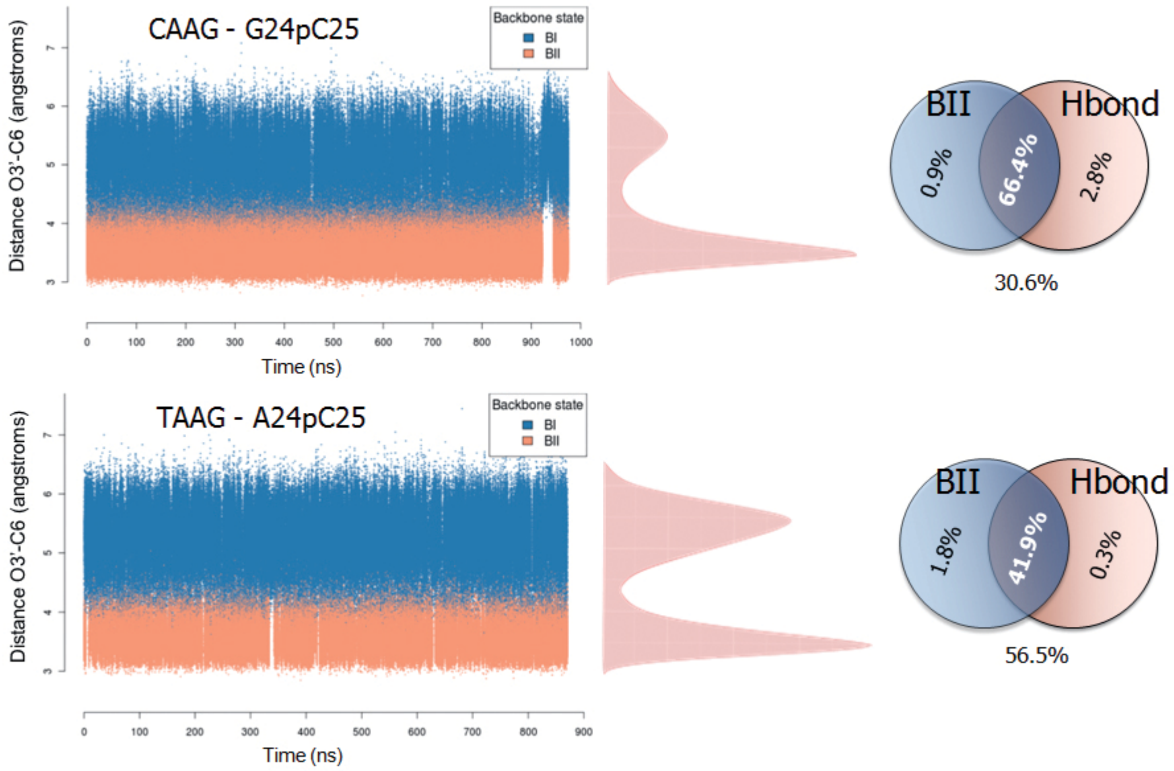


Figure S2. Left: Time evolution of C6···O3’ distance in two RpY steps (GpC and ApC) from simulations performed with the new parmBSC1 force field, colored by the backbone conformation at the step junction and the corresponding distribution. Right: Venn Diagrams of occurrences of the BII state and C6-H6···O3’ H-bond at the same RpY dinucleotide steps.

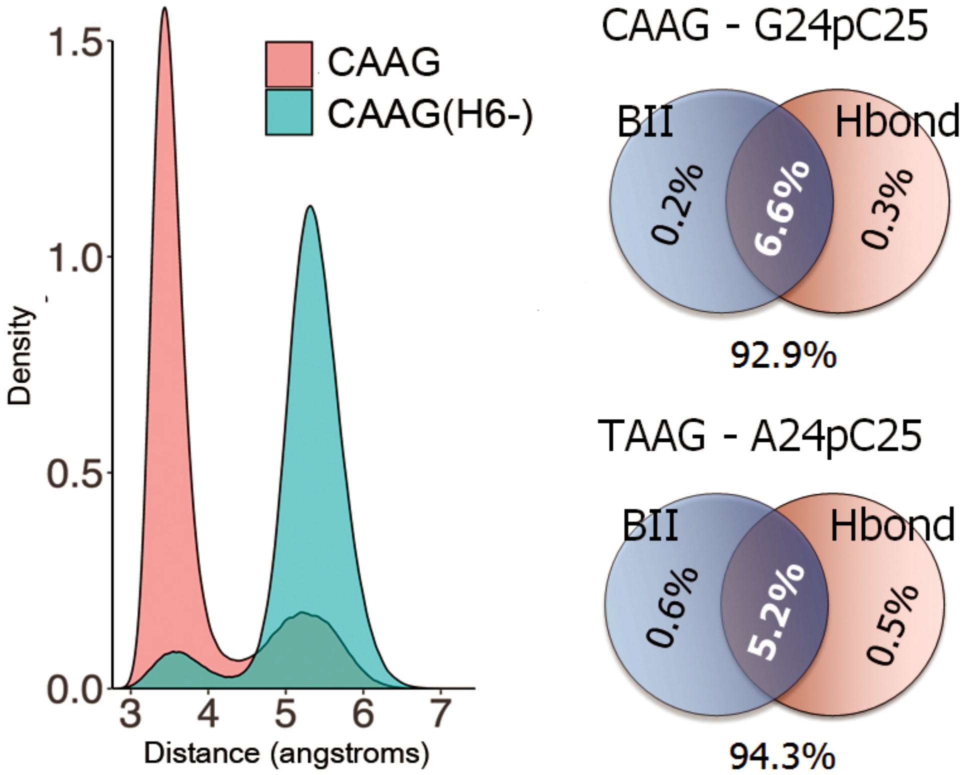


Figure S3. Left: Distribution of C6···O3' H-bond distance in the GpC step either from the conventional simulation (CAAG) or from trajectories with the modified pyrimidine base where the H6 atom was removed (CAAG(H6-)). Right: Venn Diagrams of occurrences of the BII state and C6-H6···O3' hydrogen bonds at two such modified (H6-) RpY steps (GpC and ApC).

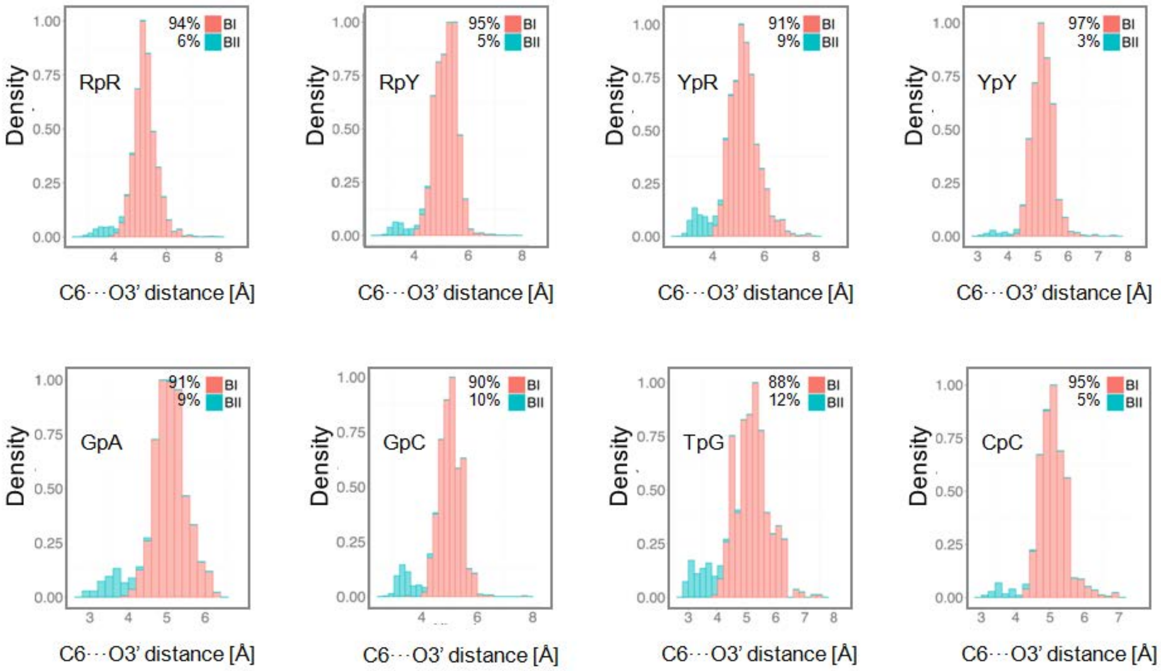


Figure S4. Distribution of distances between donor (C8 or C6) and acceptor O3' atoms of the H-bond as determined from an analysis of isolated B-DNA structures (resolution <2.5 Å) in the PDB database. Global distribution of all RpR, RpY, YpR and YpY steps (first row) and the most represented steps GpA, GpC, TpG and CpC respectively (second row). Bars are colored by the backbone state at the junction between the two bases.

SUPPLEMENTARY MATERIAL

THE PHYSICAL PROPERTIES OF B-DNA BEYOND CALLADINE'S RULES

Pablo D. Dans[a,b,1], Alexandra Balaceanu[a,b,2], Marco Pasi[c,d,2], Alessandro S. Patelli[e,2],
Daiva Petkevičiūtė[e,f,2], Jürgen Walther[a,b,2], Adam Hospital[a,b], Richard Lavery[d],
John H. Maddocks[e,1], and Modesto Orozco[a,b,g,1]

[a]Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology. Baldiri Reixac 10-12, 08028 Barcelona, Spain.
[b]Joint BSC-IRB Research Program in Computational Biology. Baldiri Reixac 10-12, 08028 Barcelona, Spain.
[c]LBPA, École normale supérieure Paris-Saclay, 61 Av. du Pdt Wilson, Cachan 94235, France.
[d]Bases Moléculaires et Structurales des Systèmes Infectieux, Univ. Lyon I/CNRS UMR 5086, IBCP, 7 Passage du Vercors, Lyon 69367, France.
[e]Institute of Mathematics, Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland.
[f]Faculty of Mathematics and Natural Sciences, Kaunas University of Technology, Studentų g. 50,
51368 Kaunas, Lithuania.
[g]Department of Biochemistry and Molecular Biology. University of Barcelona, 08028 Barcelona, Spain.

[1]To whom correspondence should be addressed:
Dr. Pablo D. Dans, Tel: +34 934039073, Email: pablo.dans@irbbarcelona.org; or
Prof. John H. Maddocks, Tel: +41 216932762, Email: john.maddocks@epfl.ch; or
Prof. Modesto Orozco, Tel: +34 934037155, Email: modesto.orozco@irbbarcelona.org.

[2]These co-authors equally contributed to this work and were alphabetically sorted.

SUPPLEMENTARY METHODS

Simulation details. Canonical duplexes were generated using Arnott B-DNA fiber parameters(1), and solvated by a truncated octahedral box of SPC/E(2) water molecules with a minimum distance of 10 Å between DNA and the closest face of the box. Systems were neutralized with K^+ or Na^+ ions adding additional 150 mM of K^+Cl^-

(or Na^+Cl^-). PARMBSC0(3) and PARMBSC1(4) force fields were used to describe DNA, while Dang's parameters were used for ions(5). Systems were optimized and equilibrated as described elsewhere(6), and simulated for 1 µs in the NPT ensemble, using Particle-Mesh Ewald corrections(7) and periodic boundary conditions. SHAKE was used to constrain bonds involving hydrogen(8), allowing 2 fs integration step. Typically, analyses presented here correspond to the second part of the trajectory (last 500 ns).

Bayesian Information Criterion (BIC), Bayes Factors, and the Helguerro's theorem. We used the BIC methodology to determine the optimal number of Gaussian functions needed to fit a given distribution. This is done by finding the set of parameters that minimizes the BIC values (the model with the lower BIC is chosen) according to(9):

$$-2lnp(x|k) \approx BIC = -2\ln(L) + kln(n)$$

Where x are the observed data, k is the number of free parameters to be estimated, and *p(x|k)* is the probability of the observed data given the number of parameters, or, in other words, the likelihood of the parameters given the dataset. L is the maximized value of the likelihood function for the estimated model, and n is the number of data points in x (the number of observations). In this work we limit the BIC to considering a maximum of two Gaussians, leading to the classification of each distribution as uninormal (fitted with one Gaussian) or binormal (fitted with a combination of two Gaussians).

The Bayes Factors that can be extracted from the BIC analysis were used to determine the strength of the evidence in favour of the model chosen by BIC(10, 11). This leaded to a third classification labelled as "insufficient evidence", when either of the two models determined with BIC (uninormal or binormal) couldn't be statistically supported.

Finally, when there was sufficient evidence to favour a binormal fitting, we used an extension of the Helguerro's theorem(12, 13) to define the modality of the distribution and distinguish the cases where the two peaks of the fitted Gaussians are close together from those where they are significantly separated. This is the most important distinction in terms of understanding DNA dynamics. In the first case, for practical purposes, the use of a single Gaussian distribution may often be justified to represent the data (the overall distribution may be interpreted as binormal-unimodal), while it cannot be ued to estimate higher moments in the second multi-peaked case (binormal-bimodal distributions). For a given parameter, we defined a base, base pair, or base-pair step as polymorphic from the structural point of view, when a given distribution was classified using these three approaches as binormal-bimodal.

Correlations between sub-states. For each tetranucleotide we calculated the correlation between the backbone state at the central step (base-pair step i) and the helical parameters shift, slide and twist at three consecutive levels around the central dinucleotide (i-1, i, and i+1). The substates of the torsion angles of the backbone were categorized following the standard definition: *gauche positive* (g+) = 60 ± 40 degrees; *trans* (t) = 180 ± 40 degrees; and *gauche negative* (g–) = 300 ± 40 degrees. For the correlations with BI/BII, we assigned to the backbone one of two possible discrete values, either BI or BII, according to the sub-state of the ζ torsion (g- or t respectively) at the central bps junction. All frames where the ζ torsion didn't fall inside the ranges defined by g- and t were not considered in the analysis. This leads to a strong reduction of the noise that comes from specific tetranucleotides, when trying to find patterns by grouping them (e.g. the "noise" arising from the individual behavior of the GAGA, GGGG, and AAGA tetranucleotides when considering the RRRR family). The point-biserial(14) correlation coefficient, mathematically equivalent to the Pearson correlation(15), was used as a measure of the correlation between these discrete sub-states of the backbone and the continuous values of the helical base-pair step parameters. The obtained correlation values were divided in five categories: i) ≥ -0.6, strong negative correlation; ii) < -0.6 and ≥ -0.4, mild negative correlation; iii) >-0.4 and< 0.4, no correlation; iv) ≥ 0.4 and < 0.6, mild positive correlation; and finally v) ≥ 0.6, strong positive correlation. We then group each of these categorized correlation matrices according to the 10 non-redundant tetranucleotide combinations of Y/R bases, and for each entry selected the dominant mode to describe the subset (i.e. the most common situation shared by the individual tetranucleotides within a family). In the same way, correlations between sum and differences of helical parameters have been computed, as previously done in Calladine's works(16, 17).

Kullback-Leibler (KL) divergence between configuration distributions. For each MD simulation we fit a Gaussian or multi-variate normal distribution on the helical coordinates by estimating a mean shape vector $\hat{w}$ and a stiffness, or inverse covariance matrix K, from the MD time series. (This Gaussian is in dimension 12N-6 for a fragment with N base pairs, so dimension 210 for the case N=18 considered here.) The KL divergence(18) is a convenient way to quantify the difference between two probability distributions. When both distributions are Gaussian with mean vectors $\hat{w}_1$, $\hat{w}_2$ and inverse covariance matrices $K_1$and K_2, then the divergence can be explicitly evaluated as:

$$D_{12} = \frac{1}{2}\left[K_1^{-1}\colon K_2 - \ln\left(\frac{\det K_2}{\det K_1}\right) - I\colon I\right] + \frac{1}{2}(\hat{w}_1 - \hat{w}_2)\cdot K_2(\hat{w}_1 - \hat{w}_2),$$

Where a colon denotes the standard Euclidean inner product for square matrices and *I* denotes the identity matrix of the same dimension as K_1 and K_2. The second term of this expression is interesting to look at separately: it quantifies the difference in expected shapes, weighted by one of the inverse covariance, and is equal to the square of the Mahalanobis distance:

$$M_{12} = \frac{1}{2}(\hat{w}_1 - \hat{w}_2)\cdot K_2(\hat{w}_1 - \hat{w}_2),$$

Both KL divergence and Mahalanobis distance are non-symmetric, but here we chose to report the symmetrized values: $D = \frac{1}{2}(D_{12} + D_{21})$ and $M = \frac{1}{2}(M_{12} + M_{21})$. To give a meaning to values of the KL divergence, the KL values were scaled by 12N-6 (being N the number of base-pairs in each oligomer), obtaining in this way a divergence per degree of freedom.

cgDNA calculation of DNA Persistence Length. The cgDNAmc code(19) allows efficient generation of ensembles of configurations over ensembles of sequences, so that the possible range of values of various expectations can be examined as the sequence of the DNA duplex varies. One standard set of expectations to compute is tangent-tangent correlations along the duplex in order to determine the associated decay rate or persistence length ℓ_p along a given fragment. The persistence length ℓ_p is often taken as an overall proxy for the stiffness of the duplex, with longer persistence length indicating greater stiffness. However it is known (see *e.g*. the discussion in ref 19) that the value of ℓ_p depends on both the stiffness of the duplex and on its intrinsic curvature, with bent sequences having lower persistence lengths. For this reason ℓ_p is sometimes called apparent persistence length. A sequence-dependent dynamic persistence length ℓ_d was introduced(19), which largely eliminates dependence on intrinsic curvature. Thus ℓ_d is a better proxy for an overall stiffness, while the difference (ℓ_d - ℓ_p) is an overall measure of how intrinsically bent the duplex is. Fig S2A provides spectra (or histograms) of possible values of both ℓ_p and ℓ_d for 10K sequences according to a cgDNA model parameter set fit to MD simulations of the miniABC library using the PARMBSC0 MD potentials. The range of variation in ℓ_d is small compared to that of ℓ_p, and it can be verified that all exceptionally low values of ℓ_p correspond to highly bent sequences. The same data for the same 10K sequences, but for a cgDNA model parameter set fit to MD simulations of the miniABC library using the PARMBSC1 MD potentials is shown in Fig S2B. The fact that the spectra of dynamic persistence lengths ℓ_d shifts to the right indicates that the PARMBSC1 potentials lead to duplexes that are slightly stiffer than for PARMBSC0, while the fact that the spectra of apparent persistence lengths has a smaller tail on the left indicates that PARMBSC1 leads to duplexes that have smaller intrinsic bends than for PARMBSC0. Figure S2 also provide the values of apparent and dynamic persistence lengths for the six independent dinucleotide tandem repeats poly(XZ). As such sequences are very straight, their apparent and dynamic persistence lengths are all very close. And for both the PARMBSC0 and PARMBSC1 parameter sets the sequence poly(AA) is the high outlier among all sequences, with poly(AT) being by far the low outlier for ℓ_d among all sequences.

Statistics, graphics and molecular plots. The statistical analysis, including the Bayesian Information Criterion (BIC), Bayes Factor analysis, Helguerro's theorem, Kullback-Lieber divergence, and correlations, as well as associated graphics, were obtained with R 3.0.1 statistical package(20), the MatLab R2016b package, numpy(21) and matplotlib(22). The molecular plots were generated using VMD 1.9(23).

SUPPLEMENTARY TABLES

Table S1. DNA sequences in the miniABC library.

Seq. number	Watson strand (5’-3’ direction)
1	GCAACGTGCTATGGAAGC
2	GCAATAAGTACCAGGAGC
3	GCAGAAACAGCTCTGCGC
4	GCAGGCGCAAGACTGAGC
5	GCATTGGGGACACTACGC
6	GCGAACTCAAAGGTTGGC
7	GCGACCGAATGTAATTGC
8	GCGGAGGGCCGGGTGGGC
9	GCGTTAGATTAAAATTGC
10	GCTACGCGGATCGAGAGC
11	GCTGATATACGATGCAGC
12	GCTGGCATGAAGCGACGC
13	GCTTGTGACGGCTAGGGC

Table S2. Sequence-averaged conformational parameters obtained from the different miniABC simulations.[a]

	miniABC$_{\mathrm{BSC0}}$-K		miniABC$_{\mathrm{BSC1}}$-K		miniABC$_{\mathrm{BSC1}}$-Na	
Parameter	Average	SD	Average	SD	Average	SD
Shear (Å)	0.02	0.30	0.02	0.30	0.02	0.30
Stretch (Å)	0.03	0.12	0.03	0.12	0.03	0.11
Stagger (Å)	0.06	0.40	0.10	0.38	0.10	0.38
Buckle (°)	0.8	10.8	1.5	9.9	1.6	9.7
Propeller (°)	-12.0	8.2	-9.0	8.1	-9.3	8.2
Opening (°)	2.2	4.5	1.8	4.3	1.8	4.2
Xdisp (Å)	-1.77	1.52	-0.88	1.36	-0.64	1.43
Ydisp (Å)	0.03	1.27	0.00	1.13	-0.01	1.17
Inclination (°)	8.2	7.1	4.0	6.6	2.8	7.0
Tip (°)	0.2	6.7	0.3	6.3	0.3	6.4
Shift (Å)	-0.03	0.69	-0.03	0.80	-0.04	0.83
Slide (Å)	-0.51	0.62	-0.29	0.55	-0.22	0.55
Rise (Å)	3.32	0.32	3.32	0.30	3.32	0.29
Tilt (°)	-0.3	4.3	-0.3	4.4	-0.3	4.5
Roll (°)	4.5	5.8	2.4	5.7	1.7	5.8
Twist (°)	32.1	5.6	34.4	5.5	34.7	5.3
α (°)	-71.1	13.9	-72.1	15.4	-72.3	15.4
β (°)	170.3	13.8	167.8	21.0	166.9	21.2
γ (°)	56.3	12.3	55.0	18.9	55.0	19.1
δ (°)	119.4	21.3	135.3	15.5	136.2	14.7
ε (°)	-167.4	25.4	-160.4	25.8	-158.6	27.1
ζ (°)	-94.1	33.5	-111.4	41.6	-113.8	43.8
χ (°)	-120.5	20.2	-112.1	17.0	-111.2	16.9
Phase (°)	128.3	37.6	151.4	26.5	152.3	25.0
Amplitude (°)	38.4	7.0	41.6	6.6	41.8	6.6

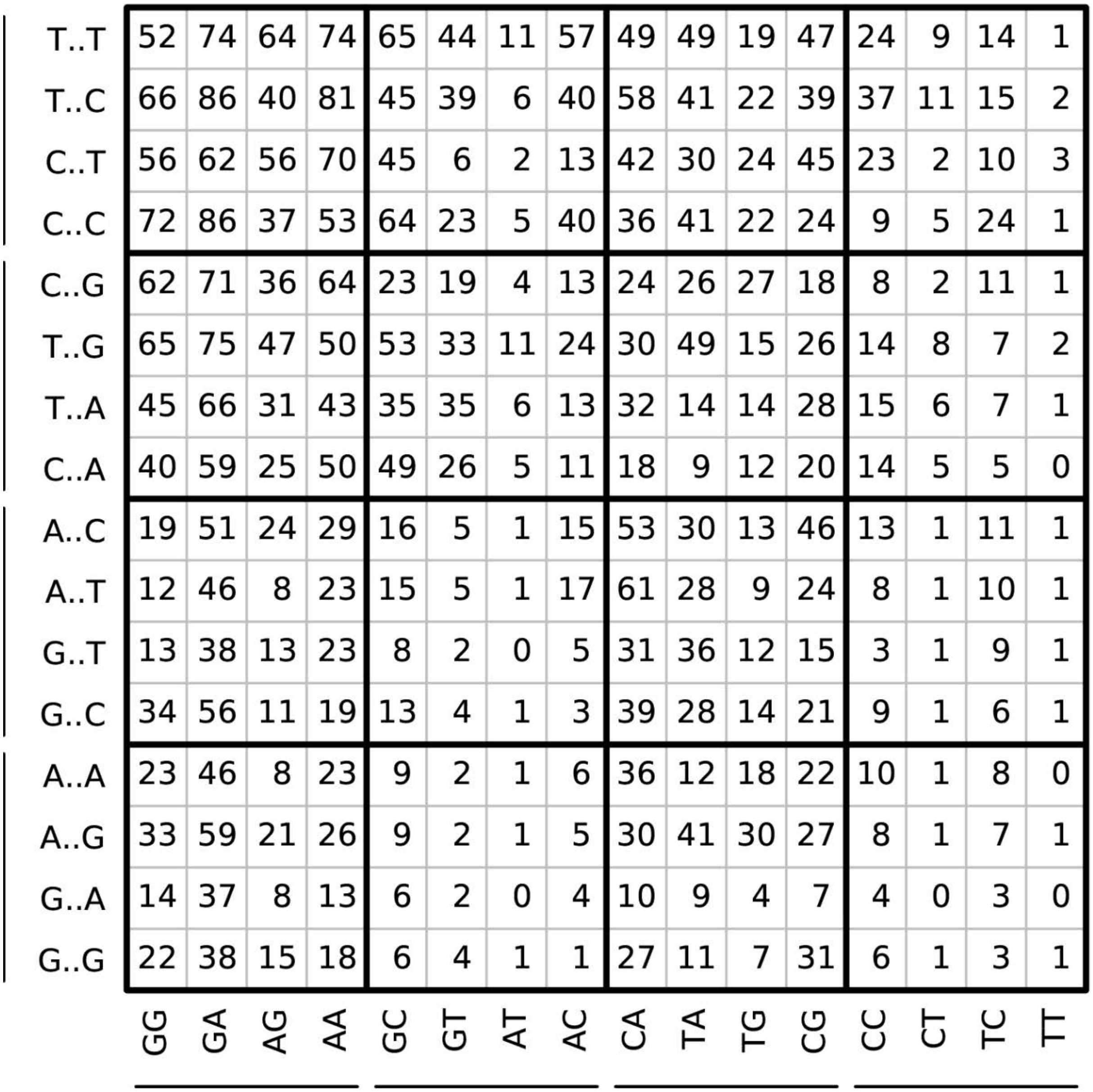
[a] Capping base pairs were removed from the analysis. For the dihedral angles only the Watson strand was considered.

Table S3. DNA breathing and fraying. Base opening statistics based on the analysis of the WC hydrogen bonds.

	Loss of one Hbond [a]		Loss of two Hbonds		Loss of three Hbonds		Solvent exchange[b]	
	Occ.[c] (%)	$\langle t_{1/2} \rangle$[d] (ns)	Occ. (%)	$\langle t_{1/2} \rangle$ (ns)	Occ. (%)	$\langle t_{1/2} \rangle$ (ns)	Occ. (%)	$\langle t_{1/2} \rangle$ (ns)
				K+Cl-				
C:G bp terminal	3.73	0.099	2.55	0.754	1.73	1.332	2.14	3.436
C:G bp terminal(-1)[e]	0.33	0.327	0.01	15.53	<0.01	---	<0.01	---
C:G bp central	0.45	0.251	0.03	10.47	0.01	315.2	0.01	149.5
A:T bp central	1.67	0.089	0.06	7.700	---	---	0.03	41.54
				Na+Cl-				
C:G bp terminal	2.81	0.095	1.57	0.761	0.87	2.209	1.20	3.552
C:G bp terminal(-1)	0.38	0.288	0.01	14.39	<0.01	---	<0.01	.---
C:G bp central	0.52	0.222	0.03	8.651	<0.01	---	<0.01	---
A:T bp central	1.59	0.094	0.04	8.963	---	---	0.01	62.49

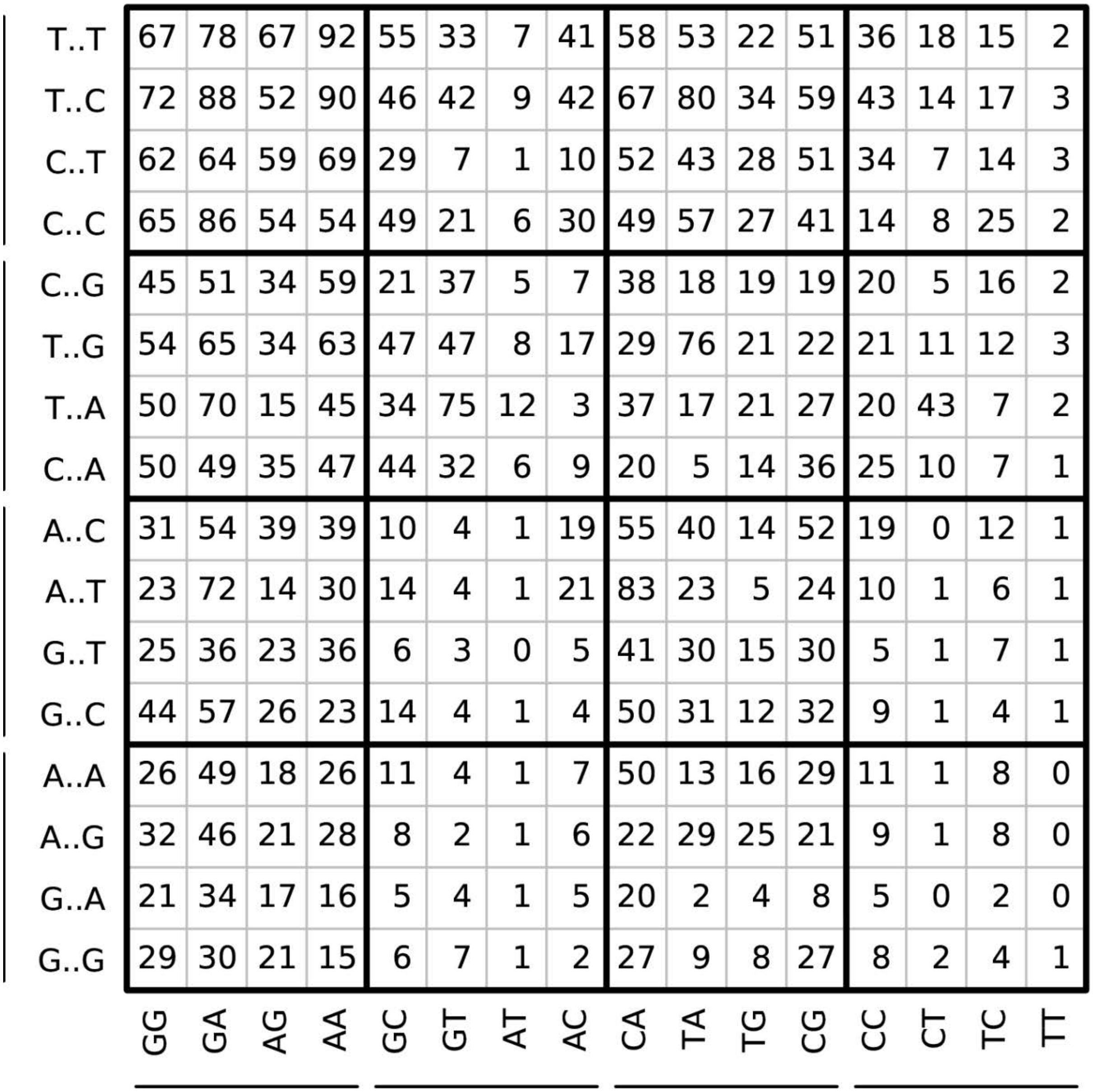
[a] We consider a hydrogen bond broken when the distance between the heavy atoms involved in the Watson-Crick interactions was greater than 3.5 Å. [b] Solvent exchange refers to base openings where at least one donor-acceptor distance of WC hbonds is larger than 6 Å. These large separations allow water molecules to interact directly with the base, and eventually exchange protons with imino groups of the bases. [c] Occ. stands for occurrence in %. [d] Average open base lifetime. [e] Refers to the C:G base-pair prior to last (residue numbers 2:35 and 17:20), see Table S1.

Table S4. BII percentages for all the 256 tetranucleotides obtained from miniABC$_{BSC1}$-K.



	GG	GA	AG	AA	GC	GT	AT	AC	CA	TA	TG	CG	CC	CT	TC	TT
T..T	52	74	64	74	65	44	11	57	49	49	19	47	24	9	14	1
T..C	66	86	40	81	45	39	6	40	58	41	22	39	37	11	15	2
C..T	56	62	56	70	45	6	2	13	42	30	24	45	23	2	10	3
C..C	72	86	37	53	64	23	5	40	36	41	22	24	9	5	24	1
C..G	62	71	36	64	23	19	4	13	24	26	27	18	8	2	11	1
T..G	65	75	47	50	53	33	11	24	30	49	15	26	14	8	7	2
T..A	45	66	31	43	35	35	6	13	32	14	14	28	15	6	7	1
C..A	40	59	25	50	49	26	5	11	18	9	12	20	14	5	5	0
A..C	19	51	24	29	16	5	1	15	53	30	13	46	13	1	11	1
A..T	12	46	8	23	15	5	1	17	61	28	9	24	8	1	10	1
G..T	13	38	13	23	8	2	0	5	31	36	12	15	3	1	9	1
G..C	34	56	11	19	13	4	1	3	39	28	14	21	9	1	6	1
A..A	23	46	8	23	9	2	1	6	36	12	18	22	10	1	8	0
A..G	33	59	21	26	9	2	1	5	30	41	30	27	8	1	7	1
G..A	14	37	8	13	6	2	0	4	10	9	4	7	4	0	3	0
G..G	22	38	15	18	6	4	1	1	27	11	7	31	6	1	3	1

Table S5. BII percentages for all the 256 tetranucleotides obtained from miniABC$_{BSC1}$-Na.

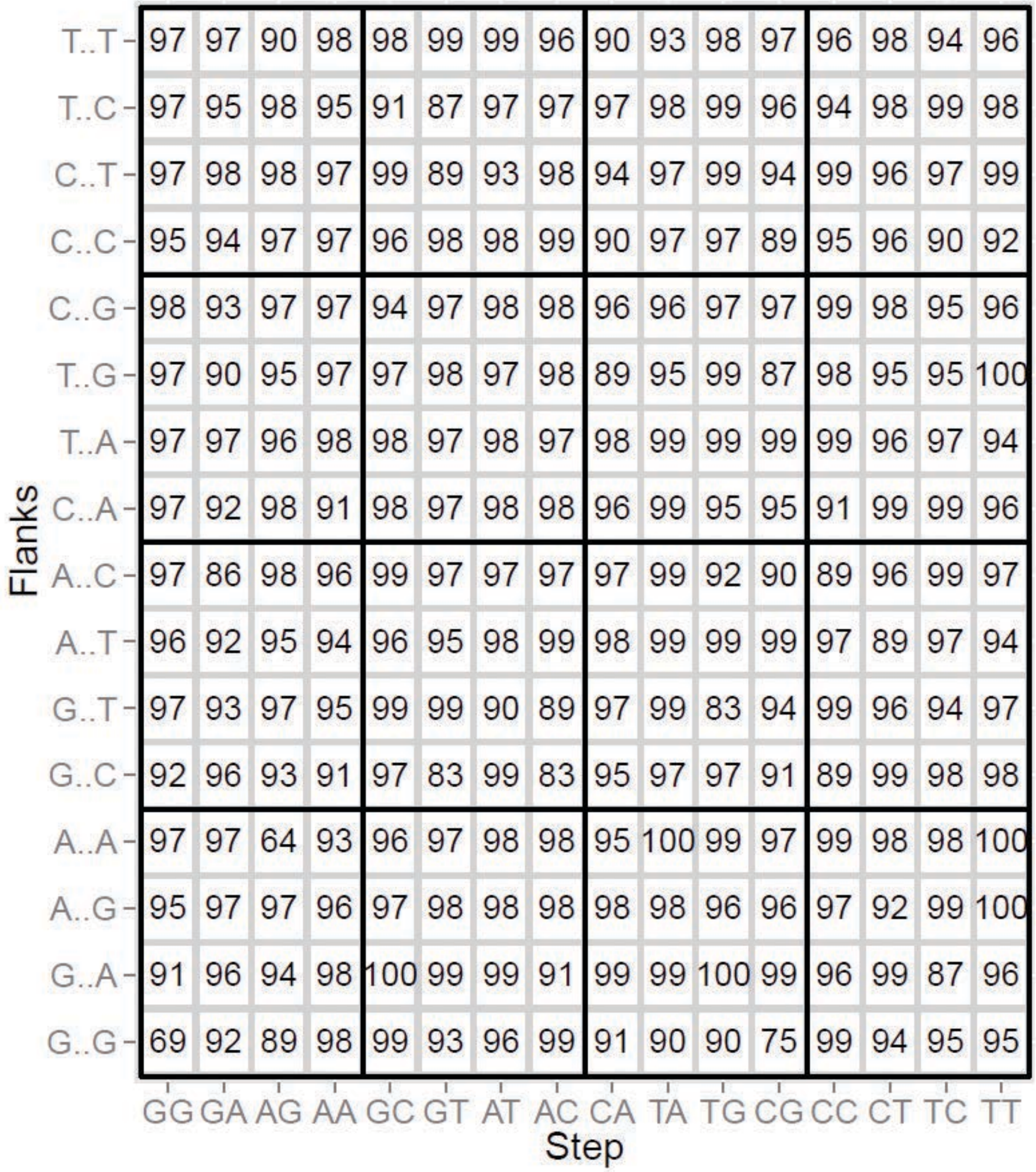


	GG	GA	AG	AA	GC	GT	AT	AC	CA	TA	TG	CG	CC	CT	TC	TT
T..T	67	78	67	92	55	33	7	41	58	53	22	51	36	18	15	2
T..C	72	88	52	90	46	42	9	42	67	80	34	59	43	14	17	3
C..T	62	64	59	69	29	7	1	10	52	43	28	51	34	7	14	3
C..C	65	86	54	54	49	21	6	30	49	57	27	41	14	8	25	2
C..G	45	51	34	59	21	37	5	7	38	18	19	19	20	5	16	2
T..G	54	65	34	63	47	47	8	17	29	76	21	22	21	11	12	3
T..A	50	70	15	45	34	75	12	3	37	17	21	27	20	43	7	2
C..A	50	49	35	47	44	32	6	9	20	5	14	36	25	10	7	1
A..C	31	54	39	39	10	4	1	19	55	40	14	52	19	0	12	1
A..T	23	72	14	30	14	4	1	21	83	23	5	24	10	1	6	1
G..T	25	36	23	36	6	3	0	5	41	30	15	30	5	1	7	1
G..C	44	57	26	23	14	4	1	4	50	31	12	32	9	1	4	1
A..A	26	49	18	26	11	4	1	7	50	13	16	29	11	1	8	0
A..G	32	46	21	28	8	2	1	6	22	29	25	21	9	1	8	0
G..A	21	34	17	16	5	4	1	5	20	2	4	8	5	0	2	0
G..G	29	30	21	15	6	7	1	2	27	9	8	27	8	2	4	1

Table S6. Pearson correlation coefficients between BII% and the formation of the C-H···O H-bond.

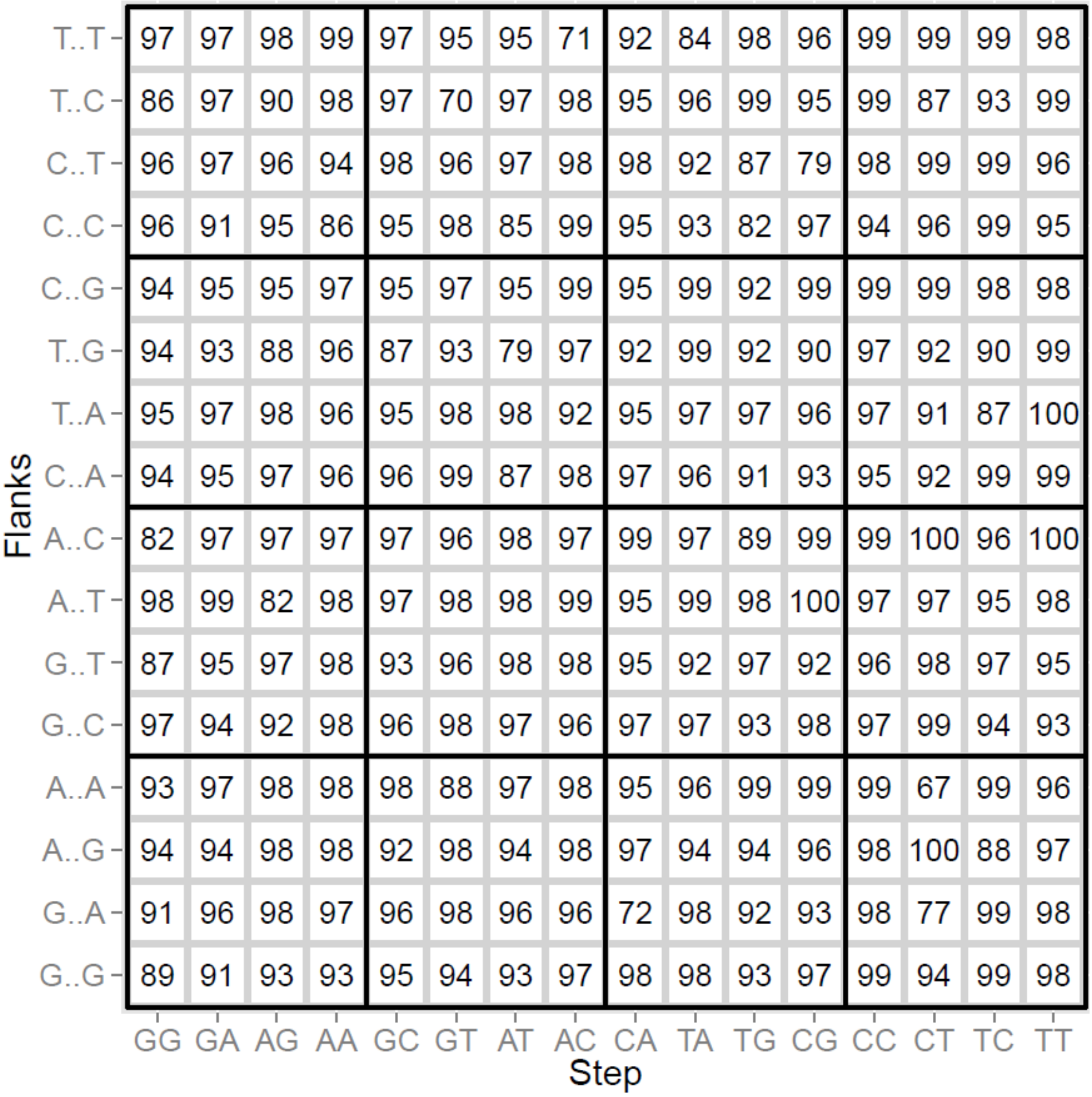
	BII% vs C8-H8···O3’		BII% vs C6-H6···O3’		
Set	RR	YR	RY	YY	Total
miniABC$_{BSC1}$-K	1.000	0.999	0.994	0.996	0.998
miniABC$_{BSC1}$-Na	1.000	0.999	0.995	0.997	0.998

Table S7. Percentages of α/γ torsions in the canonical sub-state (characterized by α in g- and γ in g+) for all the 256 tetranucleotides obtained from miniABC$_{BSC1}$-K.



Flanks	GG	GA	AG	AA	GC	GT	AT	AC	CA	TA	TG	CG	CC	CT	TC	TT
T..T	97	97	90	98	98	99	99	96	90	93	98	97	96	98	94	96
T..C	97	95	98	95	91	87	97	97	97	98	99	96	94	98	99	98
C..T	97	98	98	97	99	89	93	98	94	97	99	94	99	96	97	99
C..C	95	94	97	97	96	98	98	99	90	97	97	89	95	96	90	92
C..G	98	93	97	97	94	97	98	98	96	96	97	97	99	98	95	96
T..G	97	90	95	97	97	98	97	98	89	95	99	87	98	95	95	100
T..A	97	97	96	98	98	97	98	97	98	99	99	99	99	96	97	94
C..A	97	92	98	91	98	97	98	98	96	99	95	95	91	99	99	96
A..C	97	86	98	96	99	97	97	97	97	99	92	90	89	96	99	97
A..T	96	92	95	94	96	95	98	99	98	99	99	99	97	89	97	94
G..T	97	93	97	95	99	99	90	89	97	99	83	94	99	96	94	97
G..C	92	96	93	91	97	83	99	83	95	97	97	91	89	99	98	98
A..A	97	97	64	93	96	97	98	98	95	100	99	97	99	98	98	100
A..G	95	97	97	96	97	98	98	98	98	98	96	96	97	92	99	100
G..A	91	96	94	98	100	99	99	91	99	99	100	99	96	99	87	96
G..G	69	92	89	98	99	93	96	99	91	90	90	75	99	94	95	95

Table S8. Percentages of α/γ torsions in the canonical sub-state (characterized by α in g- and γ in g+) for all the 256 tetranucleotides obtained from miniABC$_{BSC1}$-Na.



Flanks	GG	GA	AG	AA	GC	GT	AT	AC	CA	TA	TG	CG	CC	CT	TC	TT
T..T	97	97	98	99	97	95	95	71	92	84	98	96	99	99	99	98
T..C	86	97	90	98	97	70	97	98	95	96	99	95	99	87	93	99
C..T	96	97	96	94	98	96	97	98	98	92	87	79	98	99	99	96
C..C	96	91	95	86	95	98	85	99	95	93	82	97	94	96	99	95
C..G	94	95	95	97	95	97	95	99	95	99	92	99	99	99	98	98
T..G	94	93	88	96	87	93	79	97	92	99	92	90	97	92	90	99
T..A	95	97	98	96	95	98	98	92	95	97	97	96	97	91	87	100
C..A	94	95	97	96	96	99	87	98	97	96	91	93	95	92	99	99
A..C	82	97	97	97	97	96	98	97	99	97	89	99	99	100	96	100
A..T	98	99	82	98	97	98	98	99	95	99	98	100	97	97	95	98
G..T	87	95	97	98	93	96	98	98	95	92	97	92	96	98	97	95
G..C	97	94	92	98	96	98	97	96	97	97	93	98	97	99	94	93
A..A	93	97	98	98	98	88	97	98	95	96	99	99	99	67	99	96
A..G	94	94	98	98	92	98	94	98	97	94	94	96	98	100	88	97
G..A	91	96	98	97	96	98	96	96	72	98	92	93	98	77	99	98
G..G	89	91	93	93	95	94	93	97	98	98	93	97	99	94	99	98

Step

SUPPLEMENTARY FIGURES

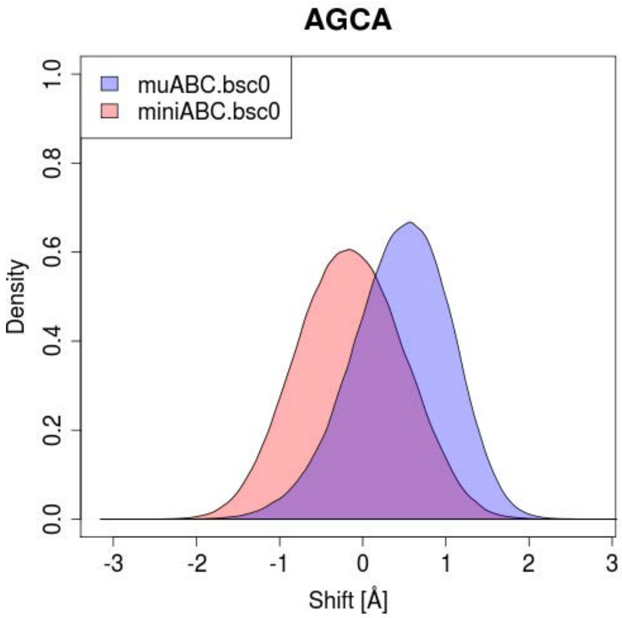


Figure S1. Shift distribution of the AGCA tetranucleotide obtained from µABC$_{BSC0}$-K and miniABC$_{BSC0}$-K. Both are bell-shaped Gaussian distributions, with a similar standard deviation, but different mean. All 1,631 pairs of other analogous marginal distributions were more similar one to the other.

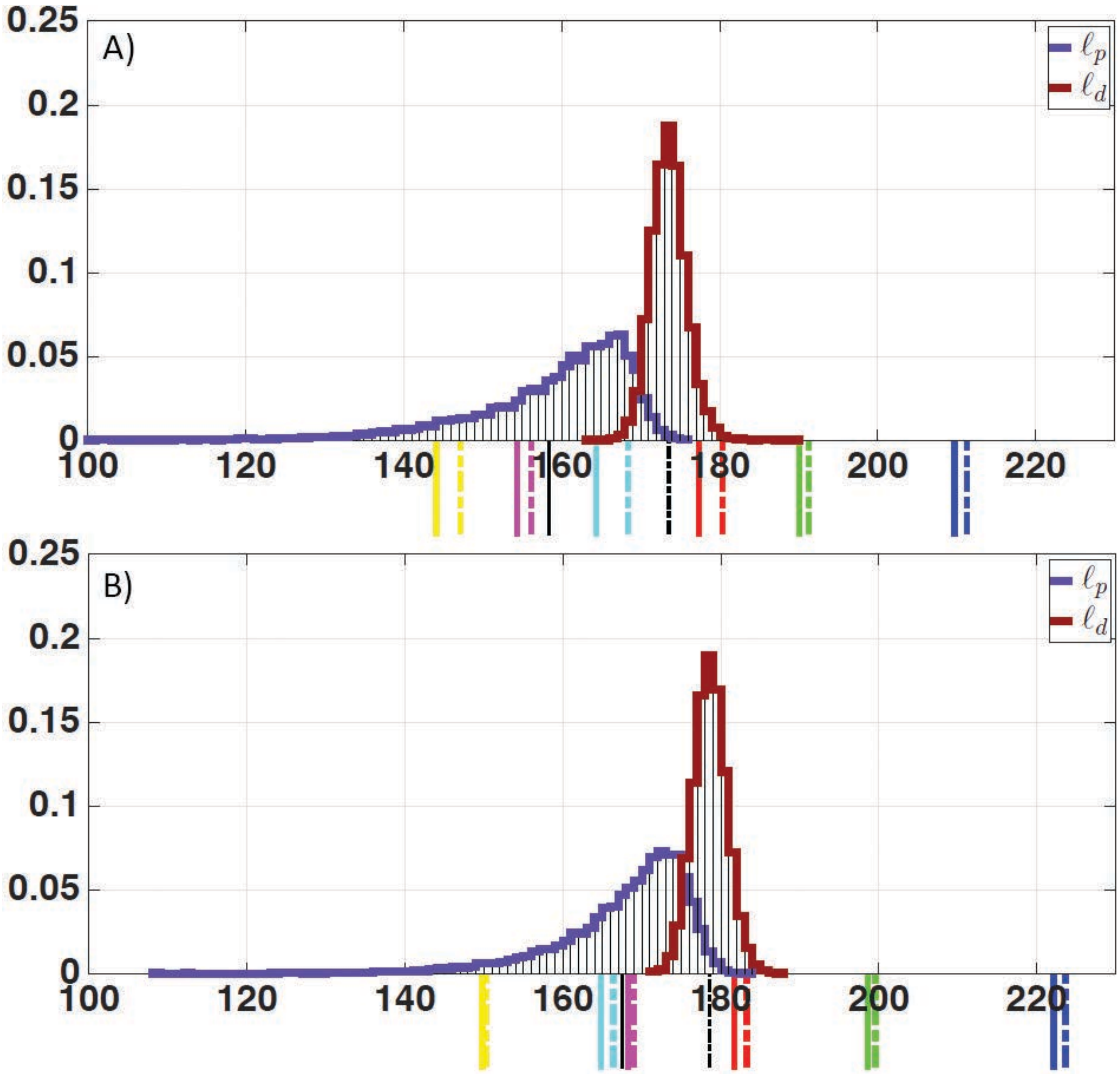


Figure S2. Spectra of ℓ_p (dark blue) and ℓ_d (dark red) persistence lengths computed over an ensemble of 10K sequences for A) PARMBSC0, and B) PARMBSC1 parameter sets, with mean for ℓ_p (black solid line) and mean for ℓ_d (black dashed line). The ℓ_p (coloured solid line) and ℓ_d (dashed solid line) values for the 6 distinct dinucleotide tandem repeats are also indicated in each case.

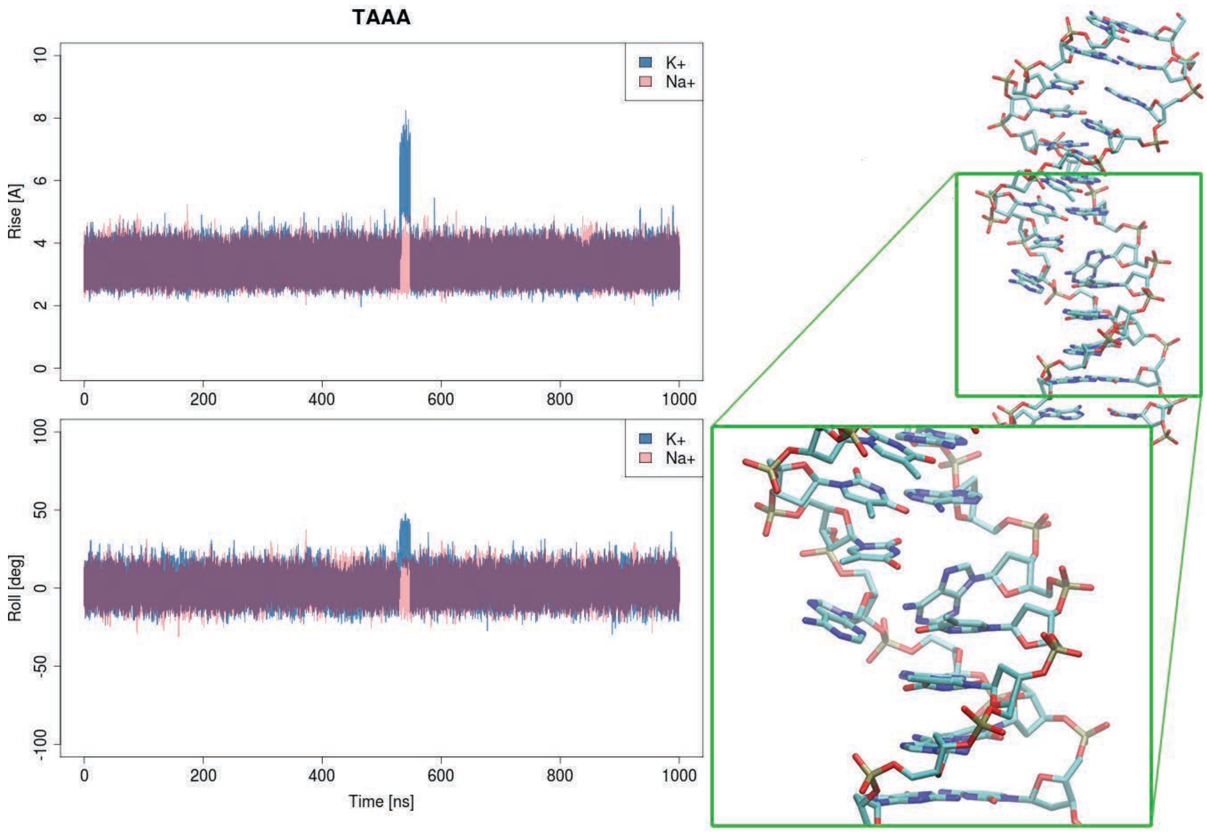


Figure S3. Time evolution of rise and roll for the TAAA tetranucleotide. The trajectory performed in K+ (blue) shows the formation of a reversible kink near 550 ns, not present using Na+ (pink). During the formation of the kink, up to two consecutive adenines lose their Watson-Crick H-bonds and are partially un-stacked. Note that this local distortion does not affect the main double helical structure of the oligomer.

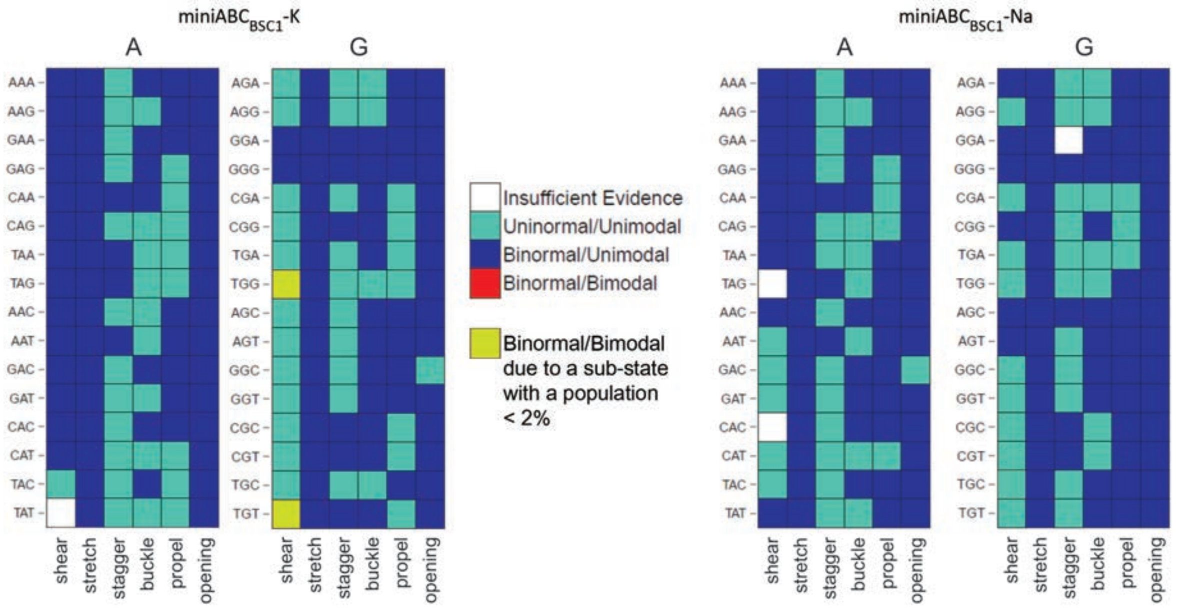


Figure S4. Structural polymorphisms (normality and modality) in base pair helical conformations for all distinct trinucleotides. Results obtained from miniABC$_{BSC1}$-K and miniABC$_{BSC1}$-Na.

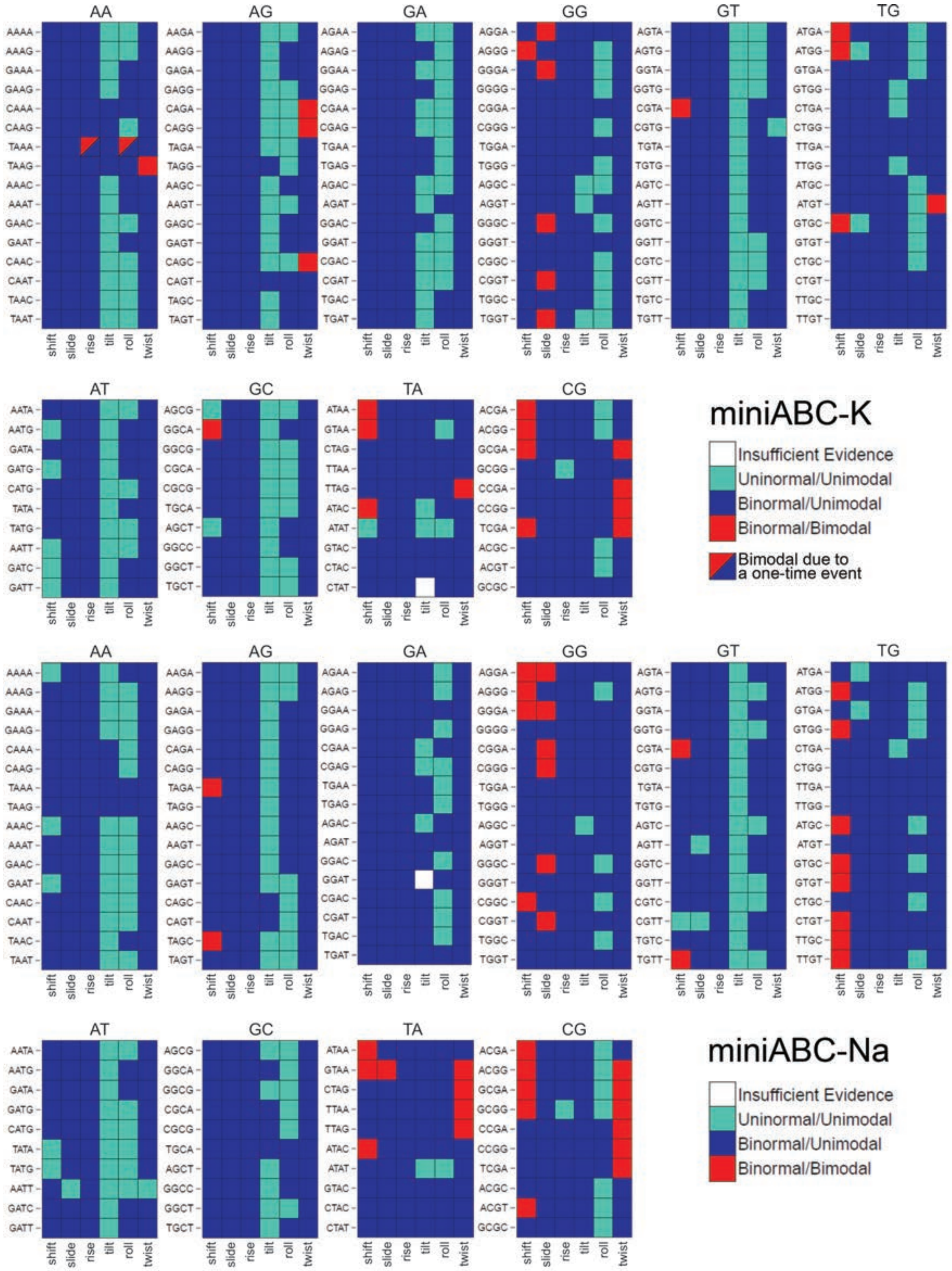


Figure S5. Structural polymorphisms (normality and modality) in base-pair step helical conformations for all the 136 distinct tetranucleotides. Results obtained from miniABC$_{BSC1}$-K (top) and miniABC$_{BSC1}$-Na (bottom). Tetranucleotides classified as binormal/bimodal (red) are considered as polymorphic (exist in two clear conformational sub-states).

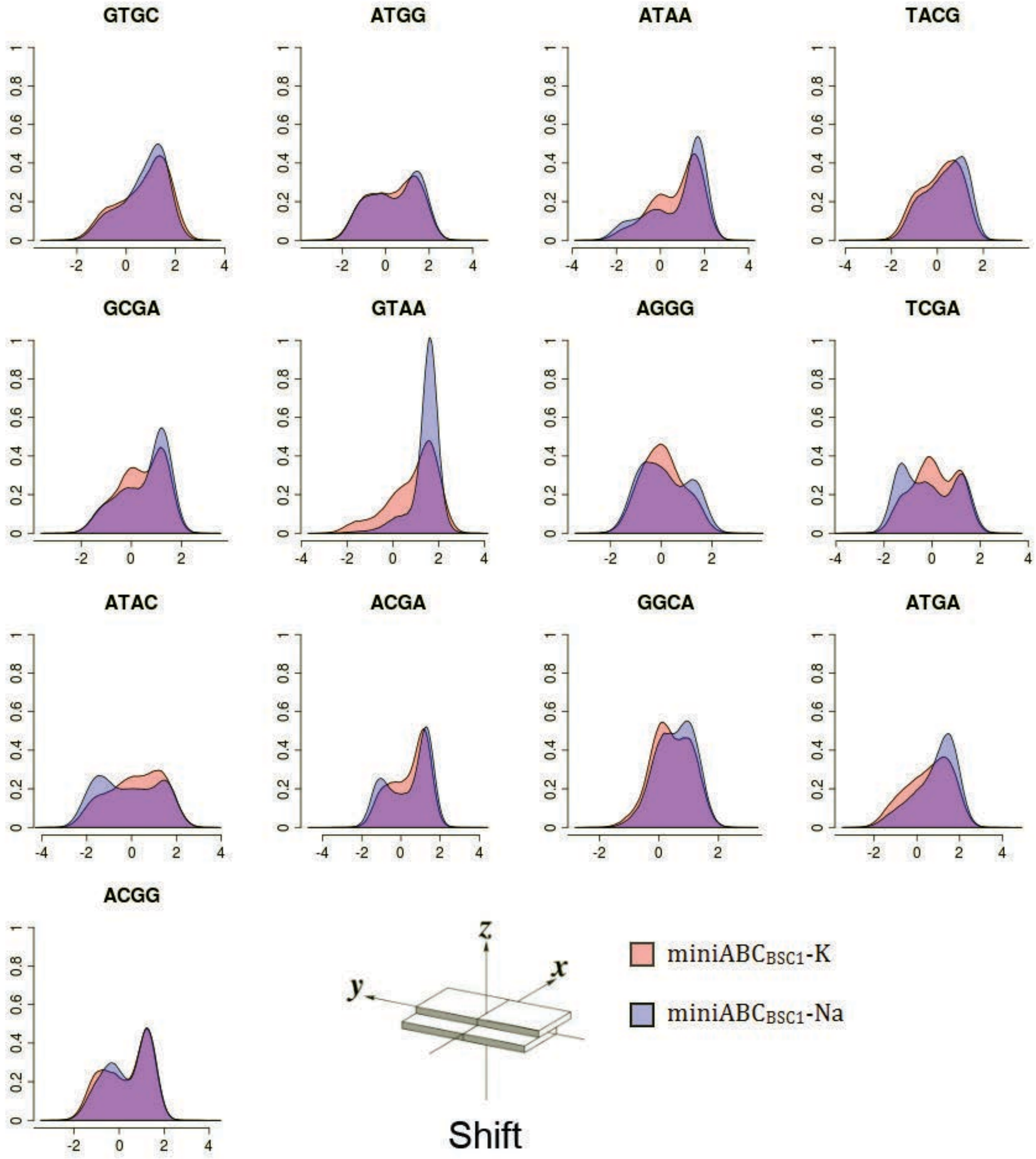


Figure S6. Normalized shift distributions for all the bimodal cases found in the miniABC$_{BSC1}$-K dataset, overlapped with their counterpart computed using Na+. X-axes represent the shift helical parameter in Å.

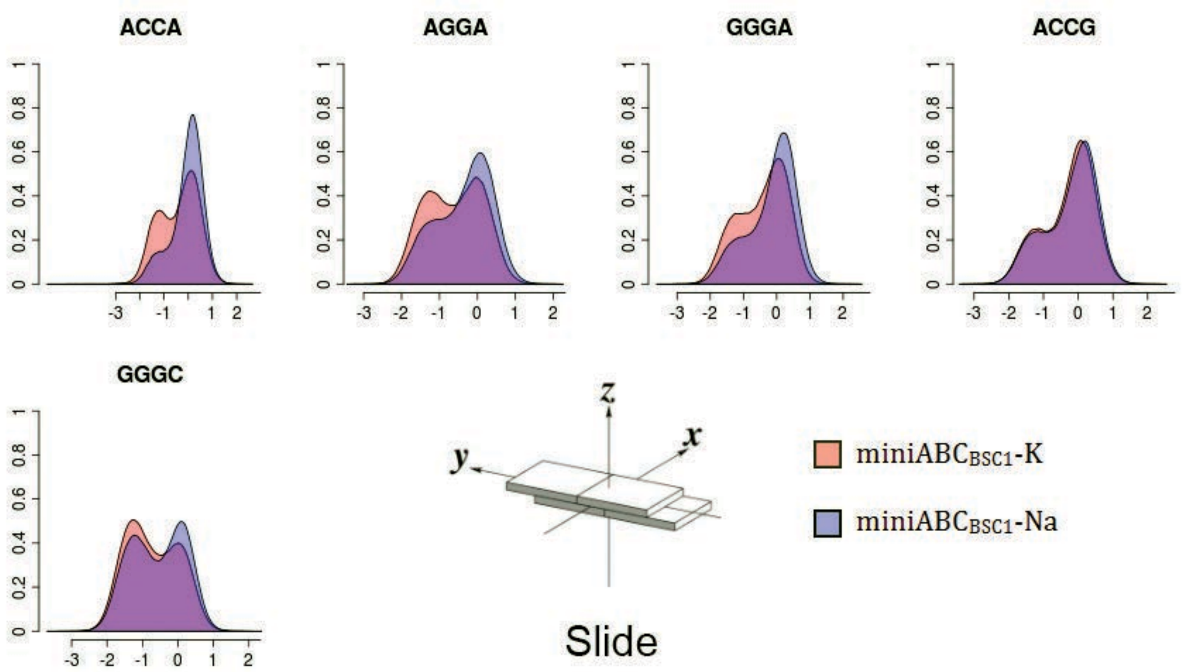


Figure S7. Normalized slide distributions for all the bimodal cases found in the miniABC$_{BSC1}$-K dataset, overlapped with their counterpart computed using Na+. X-axes represent the slide helical parameter in Å.

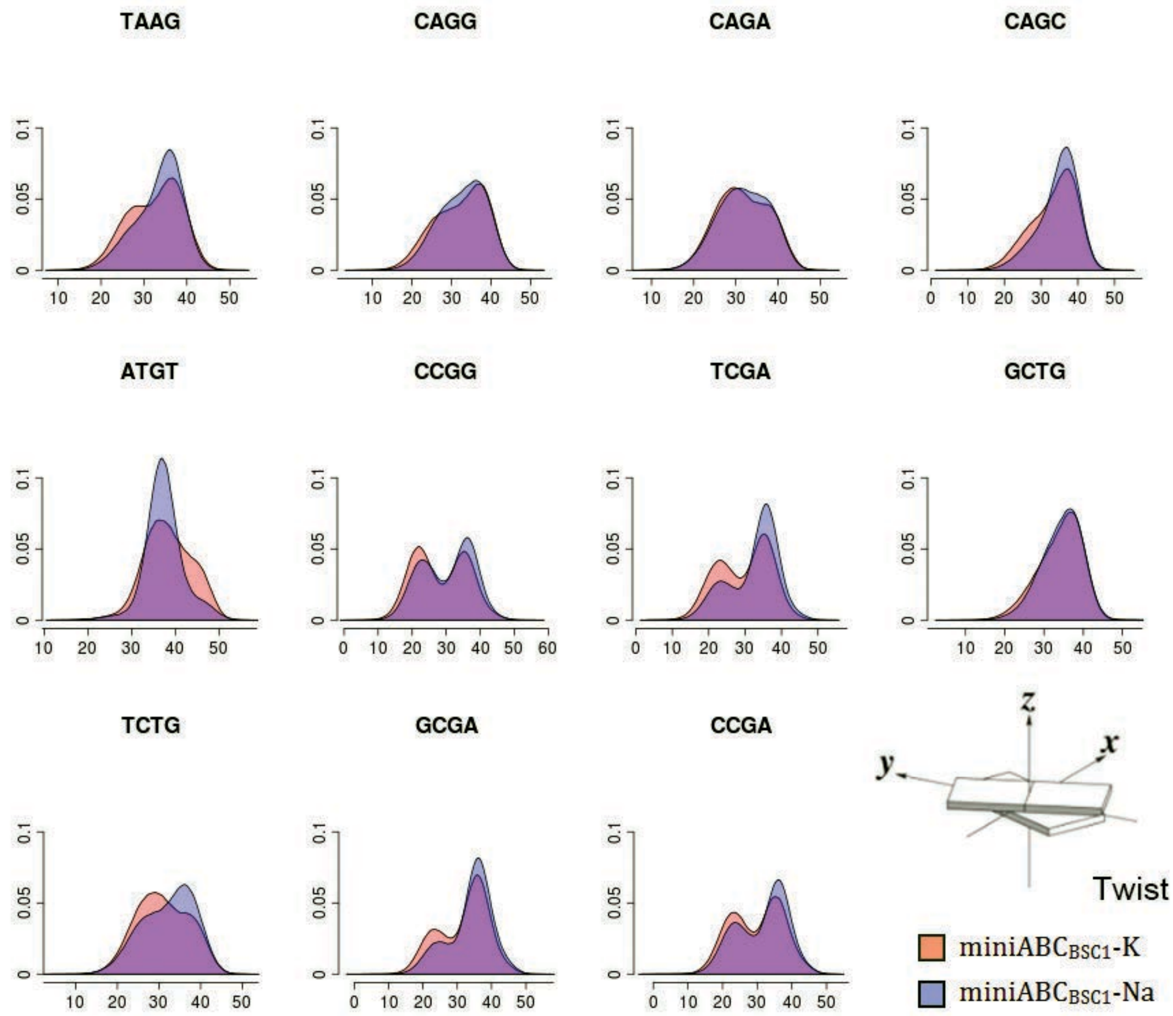


Figure S8. Normalized twist distributions for all the bimodal cases found in the miniABC$_{BSC1}$-K dataset, overlapped with their counterpart computed using Na+. X-axes represent the twist helical parameter in degrees.

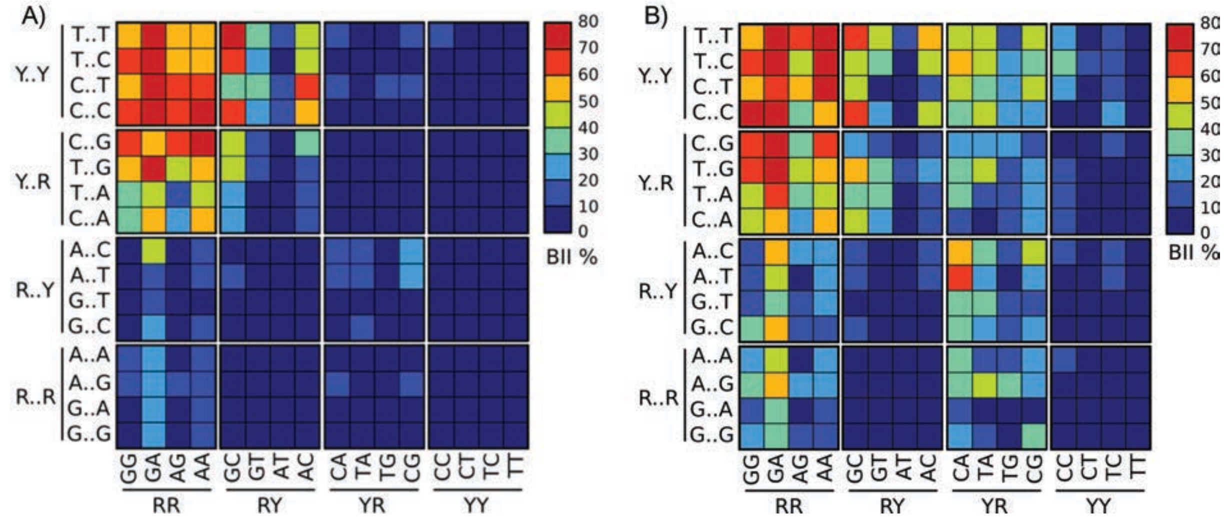


Figure S9. Sequence dependence of BII backbone conformations. The percentage occurrence of BII backbone states for the phosphodiester junction at the central base step of each of the 256 possible tetranucleotide sequences is shown (BII%), using the color code defined on the right (0% is dark blue; 80% is dark red). The sequences are arranged so that each column represents one of 16 dinucleotide steps, and each row corresponds to one of the 16 possible flanking sequences; columns and rows are further grouped on the basis of base type (R = purine and Y = pyrimidine). A) μABC$_{BSC0}$-K BII percentages(24); B) miniABC$_{BSC1}$-K BII percentages.

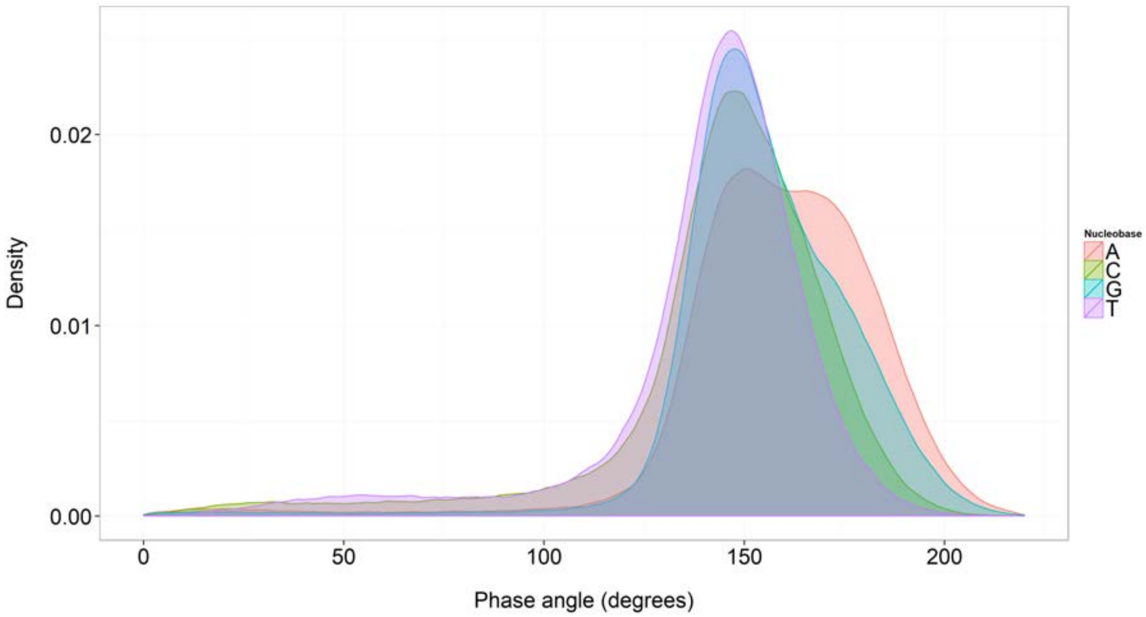


Figure S10. Normalized distribution of the P angle for A, C, G and T bases (in degrees), obtained from miniABC$_{BSC1}$-K dataset.

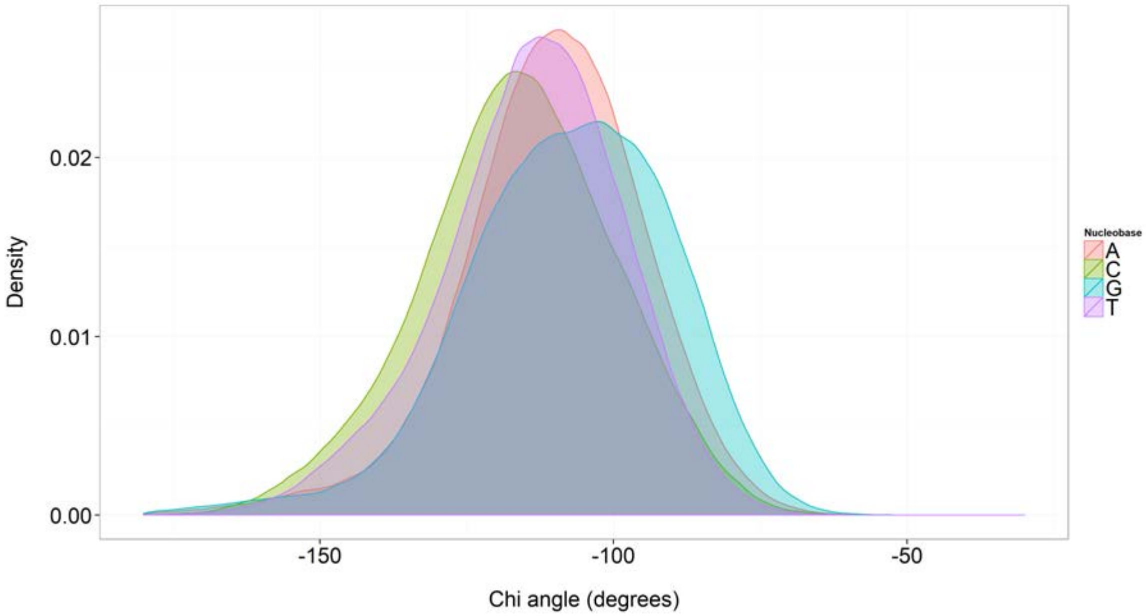


Figure S11. Normalized distribution of the χ angle for A, C, G and T bases (in degrees), obtained from miniABC$_{BSC1}$-K dataset.

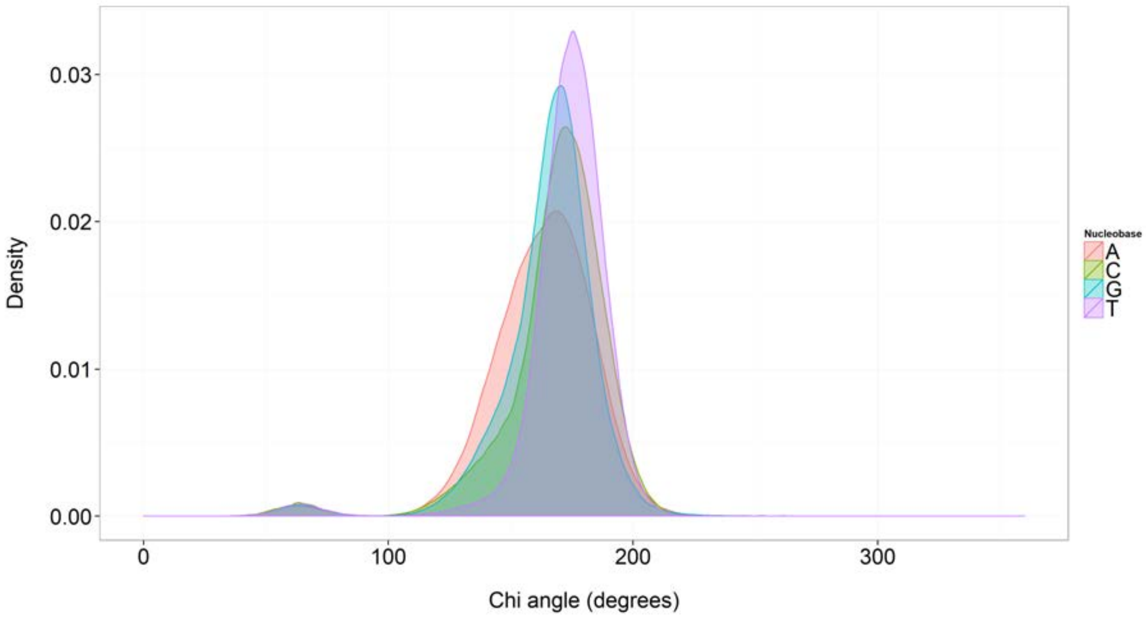


Figure S12. Normalized distribution of the β angle for A, C, G and T bases (in degrees), obtained from miniABC$_{BSC1}$-K dataset.

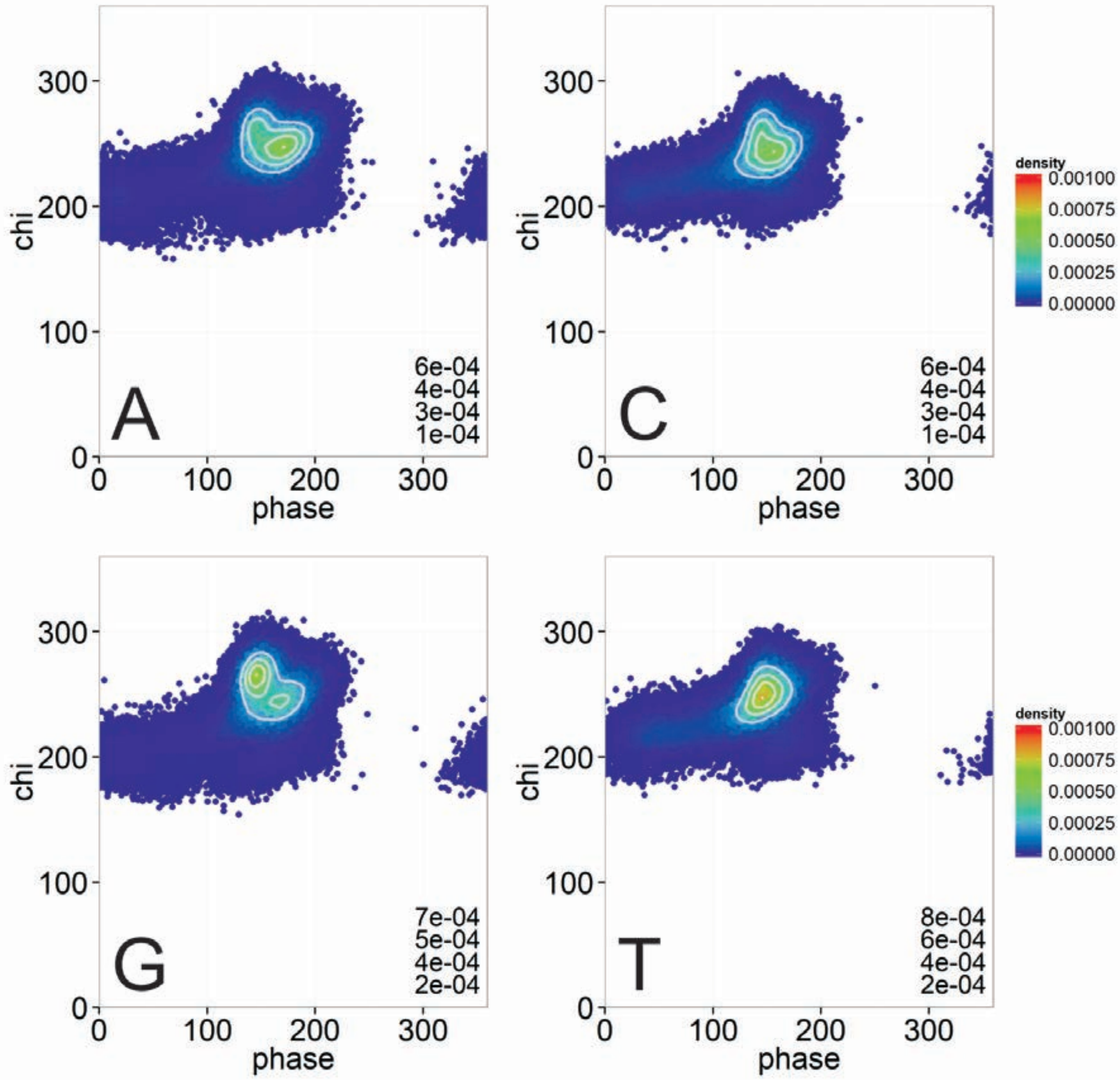


Figure S13. Phase vs χ distribution plot (in degrees) obtained from miniABC$_{BSC1}$-K dataset for A, C, G and T bases.

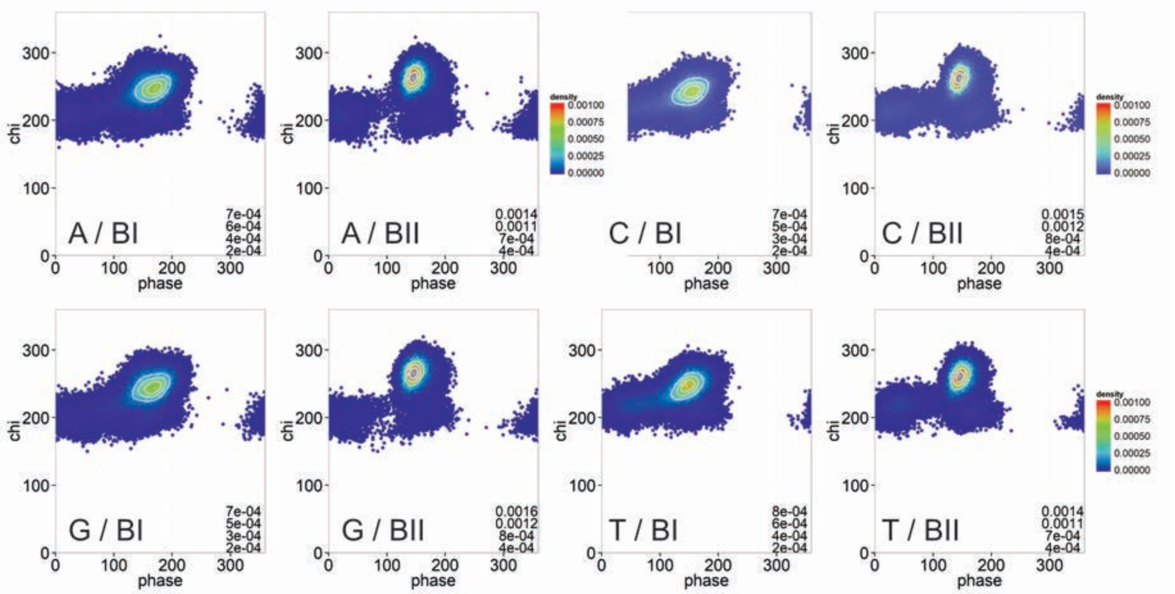


Figure S14. Phase vs χ distribution plot (in degrees) obtained from miniABC$_{BSC1}$-K dataset and filtered according to BI/BII for A, C, G and T bases.

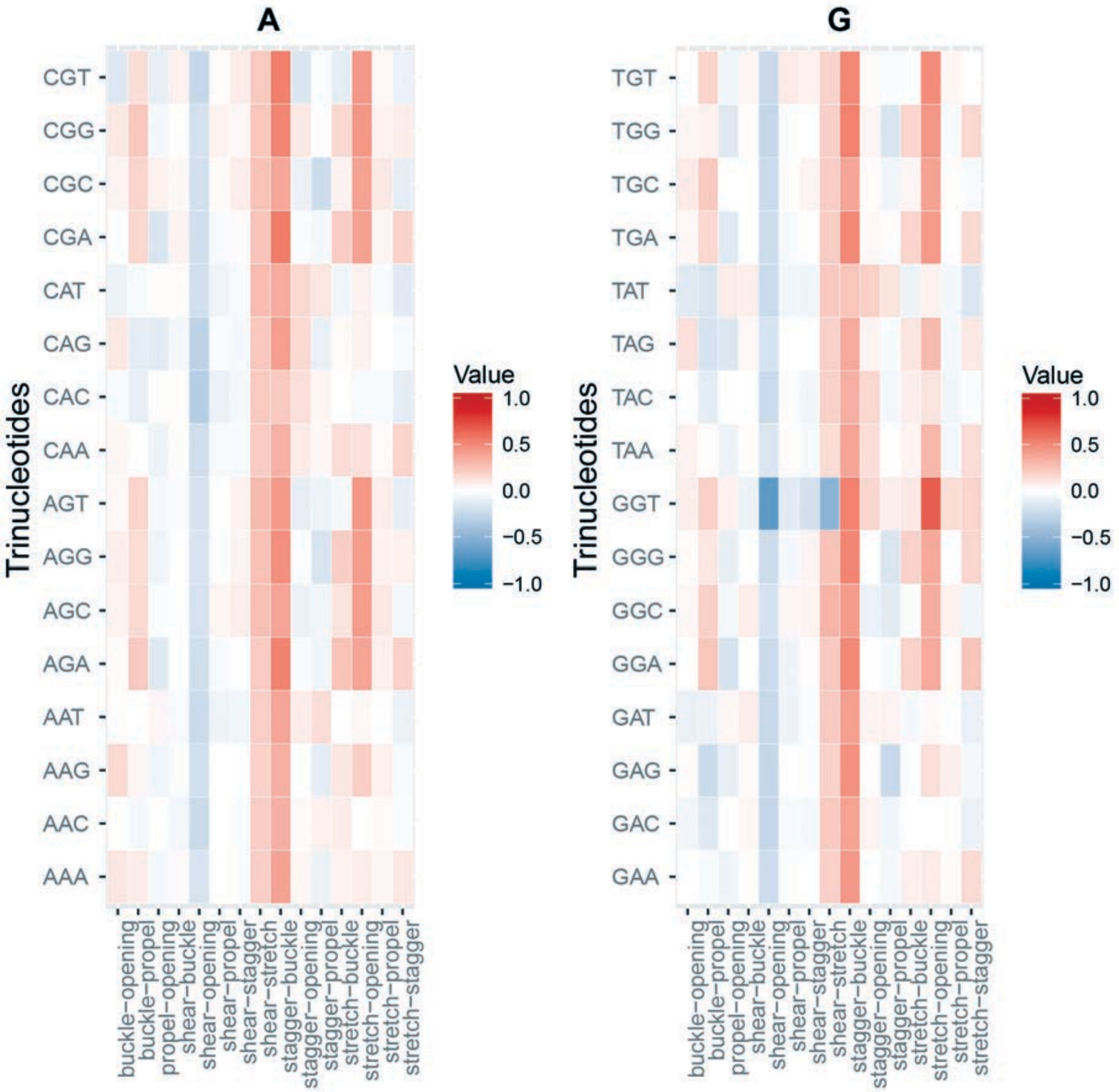


Figure S15. Correlation coefficients between intra-helical parameters (shear, stretch, stagger, propeller, buckle and opening) belonging to the same base-pair in the Watson strand. Results obtained from miniABC$_{\mathrm{BSC1}}$-K dataset for all bps.

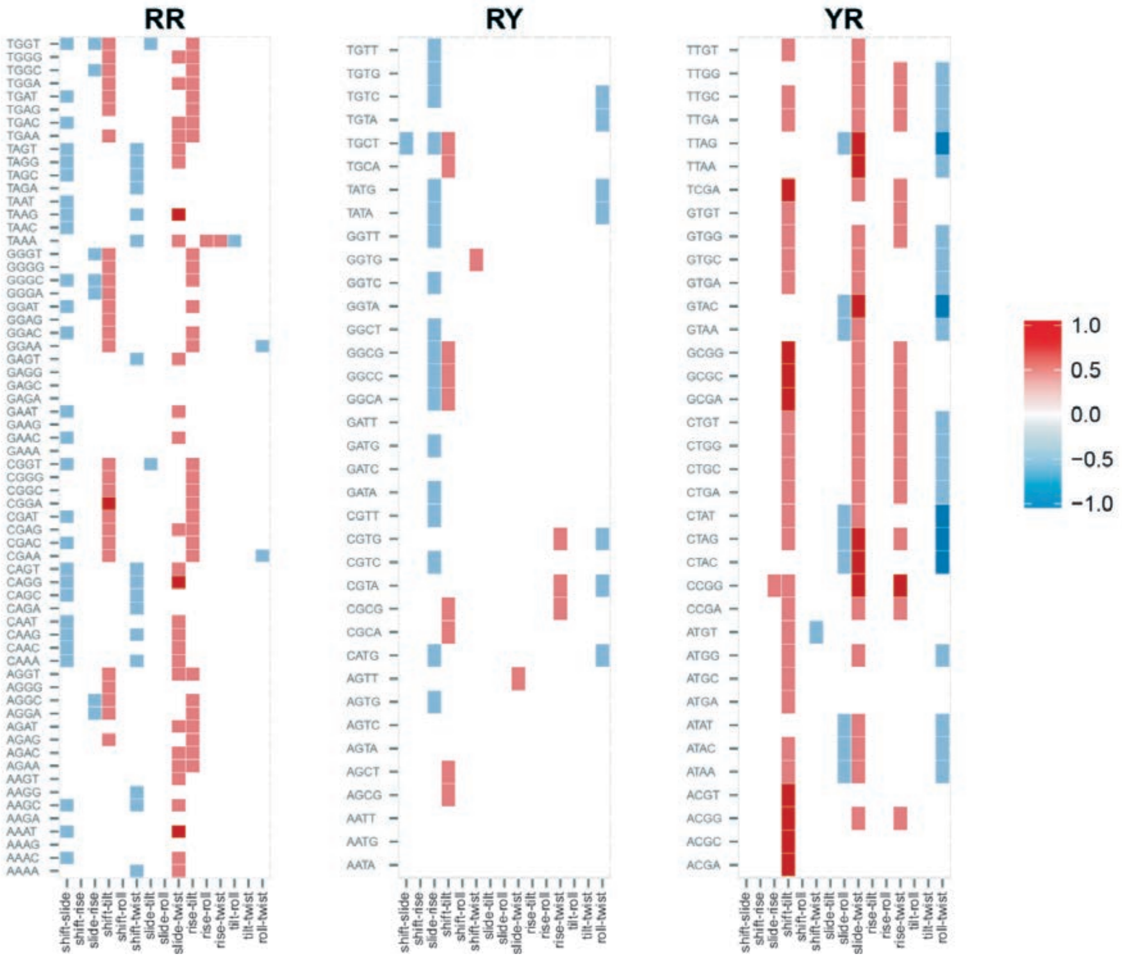


Figure S16. Correlation coefficients between inter-helical parameters (shift, slide, rise, tilt, roll, and twist) belonging to the same step in the Watson strand. Results obtained from miniABC$_{BSC1}$-K dataset for all RR, RY and YR bps.

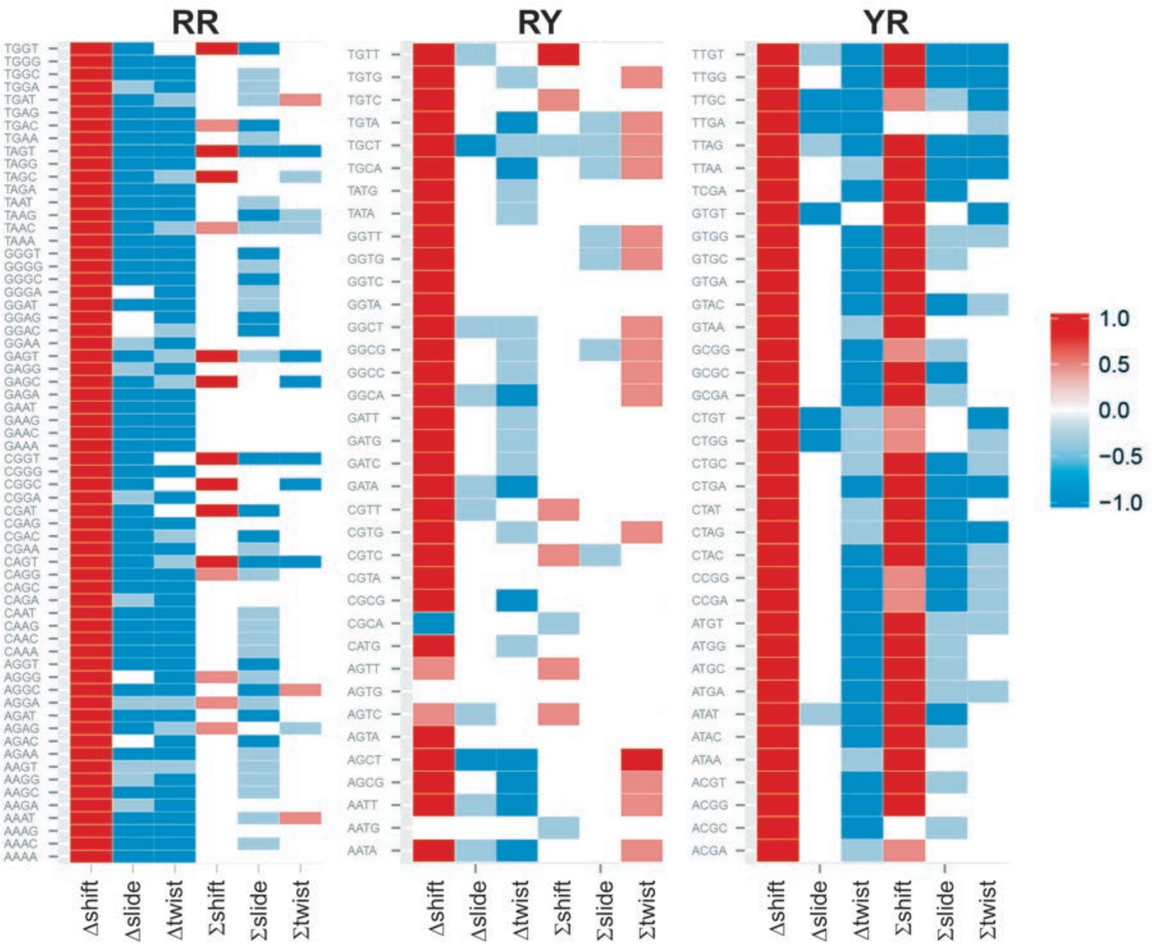


Figure S17. Correlation coefficients between differences (Δ) and sums (Σ) of base-pair steps parameters and the BII state in the central junction. Results obtained from miniABC$_{\mathrm{BSC1}}$-K dataset for all bps grouped by RR, RY and YR according to the central bps.

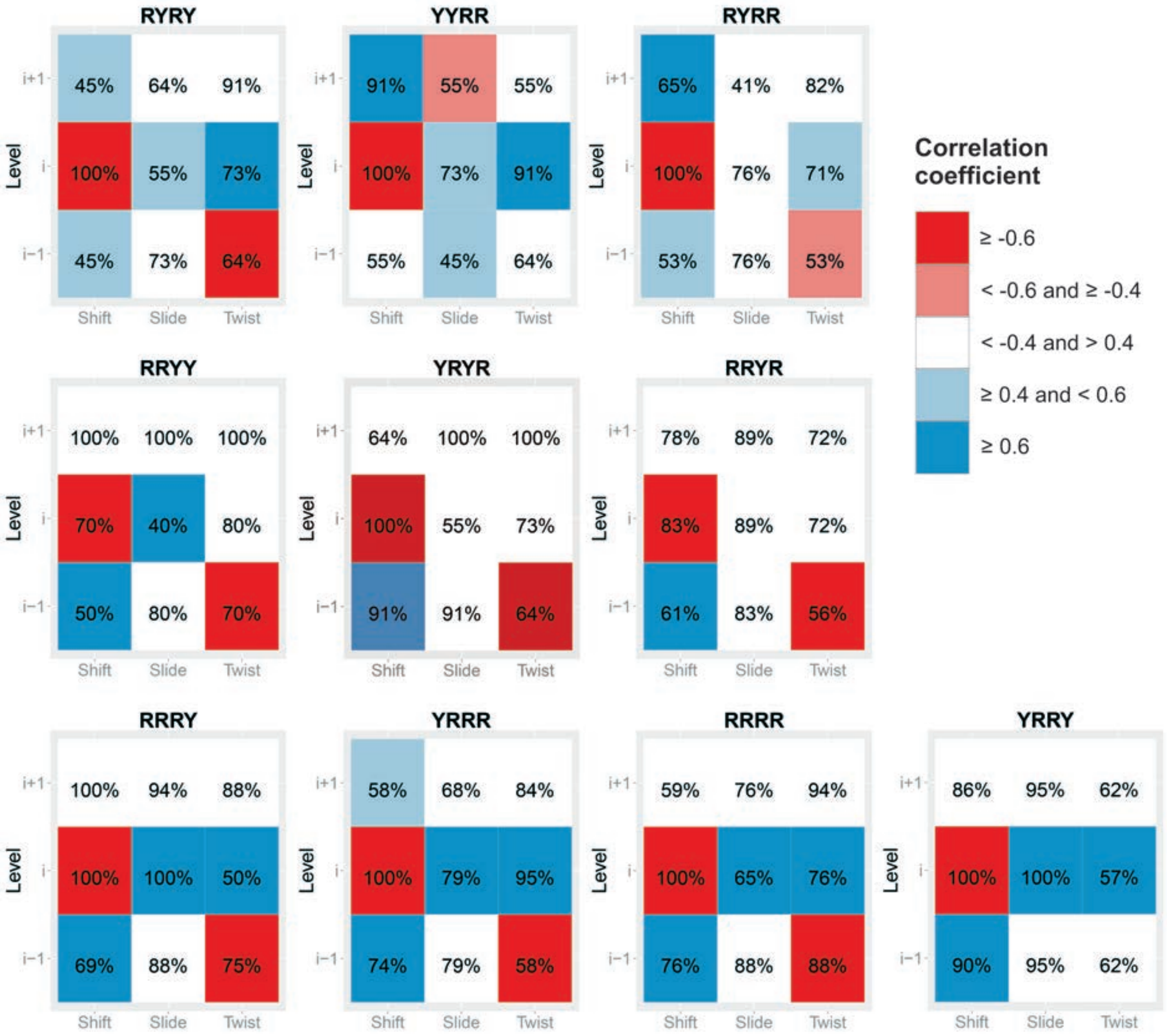


Figure S18. Correlation coefficients between shift, slide, or twist at the positions i-1 (5'-side), i, and i+1 (3'-side), and the backbone sub-state at the junction of base-pair step i in the Watson strand. Results obtained from miniABC$_{\mathrm{BSC1}}$-K dataset. The numbers inside each cell represent the % of specific tetranucleotides within a given family that give rise to the correlation.

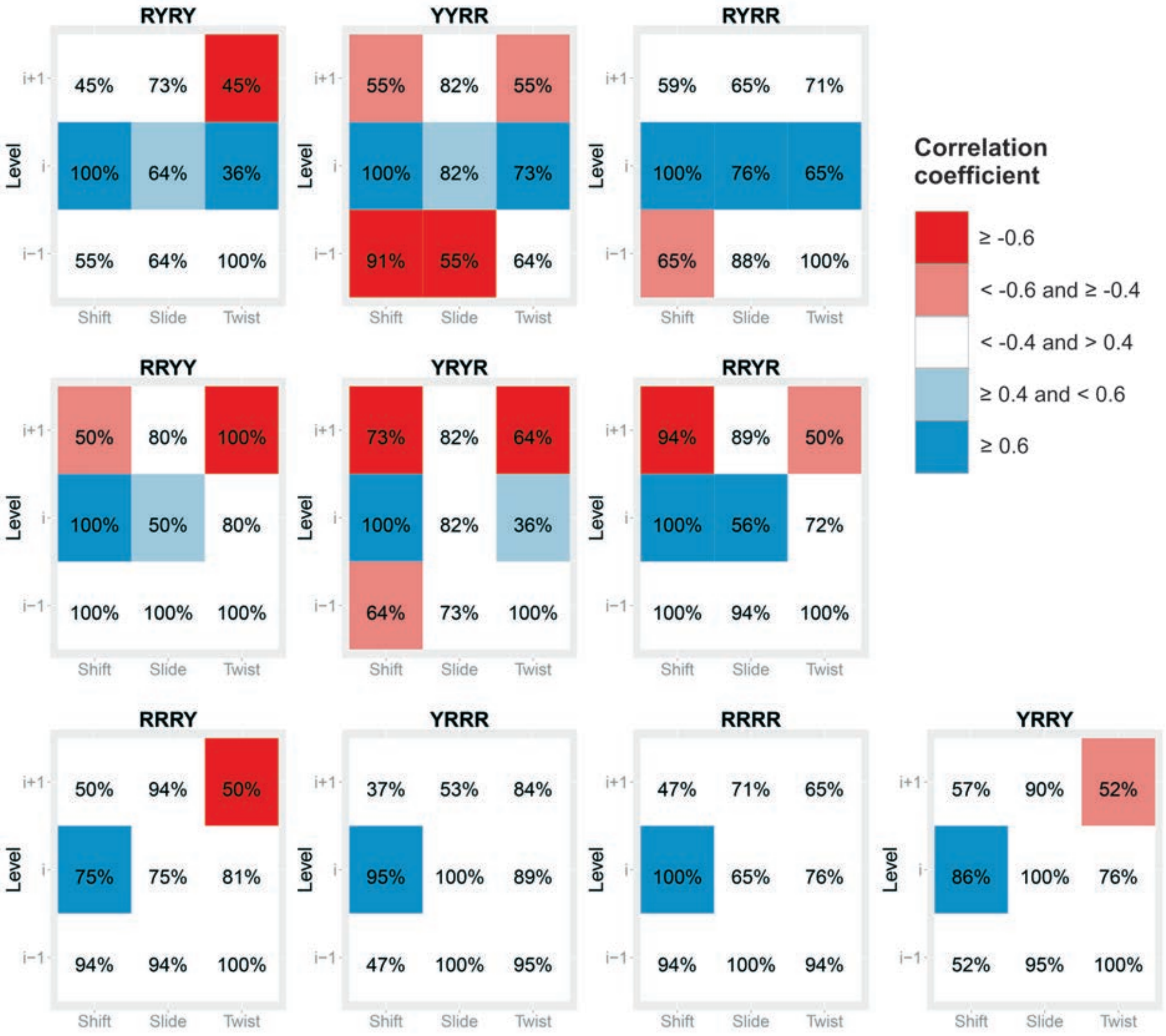


Figure S19. Correlation coefficients between shift, slide, or twist at the positions i-1 (5'-side), i, and i+1 (3'-side), and the backbone sub-state at the junction of base-pair step i in the Crick strand. Note that we refer everything to the Watson strands (see Methods), so in this plot, RRRR means YYYY since we are analyzing the correlation with the Crick strand. Results obtained from miniABC$_{BSC1}$-K dataset.

REFERENCES

1. Arnott S, Hukins DWL (1972) Optimised parameters for A-DNA and B-DNA. *Biochem Biophys Res Commun* 47(6):1504–1509.
2. Berendsen HJC, Grigera JR, Straatsma TP (1987) The missing term in effective pair potentials. *J Phys Chem* 91(24):6269–6271.
3. Pérez A, et al. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J* 92(11):3817–29.
4. Ivani I, et al. (2015) Parmbsc1: A refined force field for DNA simulations. *Nat Methods* 13(1):55–58.
5. Dang LX (1995) Mechanism and Thermodynamics of Ion Selectivity in Aqueous Solutions of 18-Crown-6 Ether: A Molecular Dynamics Study. *J Am Chem Soc* 117(26):6954–6960.
6. Pasi M, et al. (2014) µABC: A systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res* 42(19):12272–12283.
7. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An *N* ·log(*N*) method for Ewald sums in large systems. *J Chem Phys* 98(12):10089–10092.
8. Ryckaert J-P, Ciccotti G, Berendsen HJ. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* 23(3):327–341.
9. Schwarz G (1978) Estimating the Dimension of a Model. *Ann Stat* 6(2):461–464.
10. Kass RE, Raftery AE (1995) Bayes Factors. *J Am Stat Assoc* 90(430):773–795.
11. Dans PD, Pérez A, Faustino I, Lavery R, Orozco M (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res* 40(21):10668–10678.
12. de Helguero F (1904) Sui Massimi Delle Curve Dimorfiche. *Biometrika* 3(1):84.
13. Schilling MF, Watkins AE, Watkins W (2002) Is Human Height Bimodal? *Am Stat* 56(3):223–229.
14. Glass G V, Hopkins KD (2008) *Statistical methods in education and psychology* (Boston : Allyn & Bacon). 3rd ed.
15. Pearson K (1895) Note on Regression and Inheritance in the Case of Two Parents. *Proc R Soc London* 58(1):240–242.
16. Calladine CR (1982) Mechanics of sequence-dependent stacking of bases in B-DNA. *J Mol Biol* 161(2):343–52.
17. Calladine CR (2004) *Understanding DNA : the molecule &amp; how it works* (Elsevier Academic Press).
18. Lindley D V. (1959) Information Theory and Statistics. *Solomon Kullback* . New York: John Wiley and Sons, Inc.; London: Chapman and Hall, Ltd.; 1959. Pp. xvii, 395. $12.50. *J Am Stat Assoc* 54(288):825–827.
19. Mitchell JS, Glowacki J, Grandchamp AE, Manning RS, Maddocks JH (2017) Sequence-Dependent Persistence Lengths of DNA. *J Chem Theory Comput* 13(4):1539–1555.
20. R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing.

21. Oliphant TE (2007) Python for Scientific Computing. *Comput Sci Eng* 9(3):10–20.
22. Hunter JD (2007) Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* 9(3):90–95.
23. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14(1):33–8, 27–8.
24. Balaceanu A, et al. (2017) The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J Phys Chem Lett* 8(1):21–28.

4 Long-Range Effects Modulate Helical Properties of some DNA Dinucleotide Pairs

SUPPORTING INFORMATION

LONG RANGE EFFECTS MODULATE HELICAL PROPERTIES OF SOME DNA DINUCLEOTIDE PAIRS

Alexandra Balaceanu[1], Diana Buitrago[1], Jürgen Walther[1], Adam Hospital[1],
Pablo D. Dans[1] and Modesto Orozco[1,2,*]

[1] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain.
[2] Department of Biochemistry and Biomedicine, University of Barcelona, Barcelona, Spain.

* To whom correspondence should be addressed: Prof. Modesto Orozco, Tel: +34 93 403 7155, Fax: +34 93 403 7157, Email: modesto.orozco@irbbarcelona.org.

SUPPORTING METHODS

The choice of sequences. We built a library of 40 different 16 bp oligomer sequences with a middle d(CpTpApG)$_2$ that cover the entire hexanucleotide space featuring a XpCpTpApGpX sequence pattern (X stands for any nucleotide) as well as all possible pyrimidine(Y)/purine(R) combinations at the octamer level in several (>3) repeats.

System preparation and MD simulations. All the sequences were prepared with the leap program of AMBERTOOLS 16 [1] and simulated using pmemd.cuda code [2]. Following the ABC protocol [3], canonical duplexes were generated using Arnott B-DNA fiber parameters [4], and solvated by a truncated octahedral box with a minimum distance of 10 Å between DNA and the closest face of the box.

Simulations were run using parmbsc1 force filed, SPC/E water model [5] and 150 mM concentration of K^+Cl^- salt using Smith/Dang parameters [6–8]. Systems were optimized and equilibrated as described in our previous works, and simulated for at least 500 ns and up to 10 μs in the NPT ensemble, using Particle-Mesh Ewald corrections [2,9] and periodic boundary conditions. SHAKE was used to constrain bonds involving hydrogen [10], allowing 2 fs integration step. All the trajectories and the associated analysis are accessible in the MuG BigNAsim portal: http://www.multiscalegenomics.eu/MuGVRE/modules/BigNASimMuG/.

Analysis of Molecular Dynamics trajectories. All the trajectories were processed with the *cpptraj* module of the AMBERTOOLS 16 package [1], and the NAFlex server [11] for standard analysis. DNA helical parameters and backbone torsion angles were measured and analyzed with the CURVES+ and CANAL programs [12], following the standard ABC conventions [3]. Duplexes were named following the Watson strand (e.g. CTAG stands for (CTAG)·(CTAG)). The letters R, Y and X stand for a purine a pyrimidine or any base respectively, while X·X and XX represent a base pair and base-pair step respectively. Base pairs flanking the CTAG were denoted using two dots to represent the central tetrad (e.g. R··Y).

The Essential Modes of generic TpA in helical space. We performed Principal Component Analysis (PCA) of the 18 intra- and inter- base-pair parameters that define all degrees of freedom of the central TpA step in a rigid-base model. Before calculating the covariance matrix in helical space, its entries had to be made dimensionally uniform, so all rotational degrees of freedom were scaled by a factor of 10.6 [13]. The covariance was calculated from the joint equilibrated trajectories of all 40 sequences taken at every 100 ps. The first 3 Principal Components, which explain ~60% of the total variance, have their largest projections on a subset of 8 of the original 18 helical parameters. These 3 PCs were used to perform multidimensional clustering in the essential helical space using the mclust package of R. The clustering is performed using the optimal model according to Bayesian Information Criterion (BIC) for an expectation-minimization (EM) algorithm initialized by hierarchical clustering for parameterized Gaussian mixture models.

Distributions of helical parameters that guide specific sequence dependence. The helical parameters that showed the highest variability across trajectories of different sequences were identified using Principal Component Analysis (PCA) of the 18 intra- and inter base pair parameters that define all degrees of freedom of the central TpA step in a rigid-base model. The first 3 Principal Components, which explain ~60% of the total variance have their largest projections on a subset of 8 of the original 18 helical parameters. The Bayesian Information Criterion (BIC) [14,15] was used, limiting the analysis to either two or three components to determine the number of normal functions needed to meaningfully represent the appearance of possible substates in the *shift*, *slide*, *roll* and *twist* 1D distributions of the joint trajectory of all sequences. The normal distributions obtained from the BIC decomposition were compared to the distributions of the same parameters obtained after the multivariate clustering (into 3 clusters) of the first 3 PCs.

From the 8 parameters identified from the PCA as accounting for the most variance, 6 have positive coefficients in the essential helical space, namely the shift, slide and twist of TpA bps, the buckle and propeller twist of dT and the buckle of dA. The distributions of the subset of these 6 parameters were used to evaluate the similarity between the different oligos at the TpA step using the Kullback-Leibler (KL) divergence theorem. For each pair of oligomers we calculated the symmetrized values of the KL divergence and then applied hierarchical cluster analysis using Ward's clustering criterion [16], where the dissimilarities are squared before cluster updating [17] in order to identify specific sequence effects on TpA helical space flexibility.

The 4-state model of TpA dynamics. The 3D and 2D distributions of these three parameters and their paired combinations, respectively, in the meta-trajectory have also been calculated and they show a clear preference of the TpA to occupy one of four states in the Shift-Slide-Twist space. In fact, the states of the 3 helical parameters that display polymorphisms are highly inter-dependent, as shown in the 2- and 3- dimensional distribution plots. The 3 most populated states in the twist-slide-shift space, when considering the entire meta-trajectory of all oligos, are: High Twist/Positive Slide/Negative Shift (HPN), High Twist/Positive Slide/Positive Shift (HPP), and Low Twist/Negative Slide/Zero Shift (LNZ). In order to capture and better understand these effects, we filtered the meta-trajectory into 3 sub-trajectories corresponding to the 3 states, removing all frames that did not belong to any of these. We compared the distribution of helical parameters up to the octamer level in both directions ("-" sign for moving towards the 5' direction on the Watson strand and "+" sign for the 3' direction) between the 3 substate-trajectories and found significant effects in the neighboring shift, slide and twist. We also compared octamer-level backbone torsion (zeta), sugar pucker and glycosidic torsion.

Breaking down the twist, slide and shift contributions to the distal sequence effects, we calculate the Pearson's correlations of these parameters at TpA to the helical parameters at one and two levels away from TpA in each direction and the point biserial correlations to the backbone torsion (zeta – categorized in trans and gauche-), sugar pucker (categorized into South and North) and glycosidic torsion (categorized into Anti and High Anti).

Equilibrium distributions of inter base pair helical parameters at the TpA step vary with higher-than-tetramer sequence. BIC (Bayesian Information Criterion) was used to distinguish between the normal (one Gaussian) or multinormal (a mixture of two or more Gaussians) nature of the distributions of TpA helical parameters [14,15].

Since for each individual trajectory, the BIC decomposition assign the same number of Gaussians (1, 2 and 3) in the respective helical parameters (roll, twist/slide and shift, respectively) and the peaks of the distributions are consistent thought the set of oligomers, we compare the propensities of each Gaussian of the individual trajectories with the total average propensity per peak, assigning them to one of three ranges: mean – sd, mean + sd and within this interval, in order to identify large deviations in population imposed by sequence.

Correlation between twist and zeta states. As previously analyzed in depth for the CpG case, we found strong correlations between the twist state and the BI/BII backbone state at the 3' side of the TpA step on both Watson and Crick strands. The backbone state was defined by discretizing the zeta torsion sub-states into trans (180 ± 40 degrees – associated with a backbone in BII), gauche positive (60 ± 40 degrees – extremely infrequent) and gauche negative (300 ± 40 degrees – associated with a backbone in BI). Just like in the CpG case, a low twist state was found to usually be coupled with BII transitions at both 3' junctions.

Correlation between twist and O3'..C-H hydrogen bond. Relying on strong evidence from previous studies [18,19] of almost perfect correlation between backbone state and the formation of base to backbone hydrogen bonds, we looked at the correlation between twist state at the TpA step and hydrogen bond formation up to the octamer level. We found, as expected, a dependency of 3' side adjacent bond formation to twist state that perfectly mirrors that of the backbone state. But we also discovered an insightful sequential anti-correlation of bond formation from one step to the next that is also highly dependent on sequence, which favors the formation of one or the other.

Analysis of the dynamics of transitions in the Shift-Slide-Twist space. The propensity of each individual state in the 3D distribution in Shift-Slide-Twist varies heavily with sequence. When analyzing transitions between the 3 states defined above for the meta-trajectory, we found that often they use one or both of two extra intermediary states, defined by High Twist/Positive Slide/Zero Shift and High Twist/Negative Slide/Zero Shift. The populations of these states are significantly lower than the main states, but they play an instrumental role in the transitions between them. We analyzed for each sequence trajectory the transition patterns when considering these 5 states in the Twist/Slide/Shift space. The width of the arrows is proportional to the number of transitions and the area of the circles is proportional to the maximum residence time in that particular state.

Classical molecular interaction potentials. Our classical molecular interaction potential (cMIP [20]) was used to analyze the ability of DNA to recognize sodium. The electrostatic interaction term was determined by solving the linear Poisson–Boltzmann equation [21], while the van der Waals contribution was determined using standard AMBER Lennard–Jones parameters [22]. The ionic strength and the reaction-field dielectric constant were set to 0.15 and 78.4 M, respectively, while the dielectric constant for DNA was set to 8 [23]. The calculations were performed using the average structure of each sequence over the last 200 ns of the trajectory.

Stacking and Base-pairing strength. In order to estimate the strength of stacking at the TpA step we used two separate methods. First, we calculated a Stacking Factor based on the distance between the centers of mass of T and A, and the angle between the two planes of the bases. With the stacking coordinate defined as [24]:

$$\xi = \frac{r_M}{S(\alpha)}$$

$$S(\alpha) = e^{-\alpha} + e^{-(\alpha-\pi)^4} + 0.1e^{-(\alpha-0.5\pi)^4}$$

We also show that the Stacking Factor is highly correlated with shift and slightly less correlated with slide.

Secondly, we calculate the Stacking energy as the sum of electrostatic and Van der Waals terms between the bases when removing the negative charge on the phosphate group. Here we have separately contributions from stacked bases on the same strand, from cross terms and the base-pairing hydrogen bonds energies. We calculated the stacking energies separately for the 5 states defined in the transition analysis above to determine the stabilizing factors of the highly preferred states. The TpA or CTAG stacking and hydrogen bond energies don't vary significantly depending on the sequence, but they do depend on the substate in helical space.

Database Analysis of structural features. We retrieved high resolution (< 3Å) X-ray and NMR structures of double stranded DNA containing the CTAG tetrad and distinguished between the protein-bound and free DNA structures. We found 106 instances of CTAG in 29 DNA structures without protein and 160 instances of CTAG in 76 protein-DNA complex structures. We compared helical parameter distributions between the database structures and out results.

Database Analysis of genomic properties. Prevalence of CTAG in the genomes of H. sapiens (hg19), E. coli (NC_000913.3) and S. cerevisiae (sacCer3) was computed, finding low occurrence compared to other tetramers (less than 0.5% in the three species). Occurrences of this tetramer were then mapped, using Homer software

[25], to the annotated regions of each organism obtained from UCSC and compared to the overall frequency of each annotation type. CTAG is enriched at intergenic regions in H. sapiens and E. coli, but not in S. cerevisiae probably due to the low number of intergenic regions in this organism (less than 2.5% compared to more than 20% in the other two). To evaluate resilience to mutation, the frequency of mutations for each tetramer along the genome in 30 different cancer types [26] was computed.

SUPPORTING TABLES

Table S1. Sequence library used to study CTAG polymorphisms, number of replicas and simulation time.

Num.	**Sequence**	**Simulation time**	**Num.**	**Sequence**	**Simulation time**
1	CGTCGGCTAGCCGAGC	500 ns	21	CGGAGACTAGACTCGC	500 ns
2	CGTCTCCTAGGAGAGC	500 ns	22	CGGAGACTAGCCTCGC	500 ns
3	CGAAAACTAGAAAAGC	500 ns	23	CGGAGACTAGGCTCGC	500 ns
4	CGAAAACTAGTTTTGC	500 ns	24	CGGAGACTAGTCTCGC	2 μs
5	CGATATCTAGATATGC	500 ns	25	CGGAGCCTAGACTCGC	500 ns
6	CGTATACTAGTATAGC	2 x 500 ns	26	CGGAGCCTAGCCTCGC	2 x 500 ns
7	CGGGGGCTAGGGGGGC	500 ns	27	CGGAGCCTAGGCTCGC	500 ns
8	CGGGGGCTAGCCCCGC	500 ns	28	CGGAGGCTAGACTCGC	500 ns
9	CGGCGCCTAGGCGCGC	500 ns	29	CGGAGGCTAGCCTCGC	500 ns
10	CGCGCGCTAGCGCGGC	500 ns	30	CGGAGTCTAGACTCGC	2 x 500 ns
11	CGTCTACTAGAGAGGC	500 ns	31	CGCTAGCTAGCTAGGC	4 x 500 ns
12	CGTCTACTAGCGAGGC	2 x 500 ns	32	CGATATCTAGAAATGC	2 μs
13	CGTCTACTAGGGAGGC	2 x 500 ns	33	CGGAGCCTAGAATCGC	2 μs
14	CGTCTACTAGTGAGGC	2 x 500 ns	34	CGGCGCCTAGGGGCGC	2 μs
15	CGTCTCCTAGAGAGGC	2 x 500 ns	35	CGGAGGCTAGCATCGC	2 μs
16	CGTCTCCTAGCGAGGC	500 ns	36	CGAAAACTAGTATAGC	2 μs
17	CGTCTCCTAGGGAGGC	500 ns	37	CGCTAGCTAGCGAGGC	2 μs
18	CGTCTGCTAGAGAGGC	2 μs	38	CGTCTGCTAGACAGGC	2 μs
19	CGTCTGCTAGCGAGGC	6 x 500 ns	39	CGAATCCTAGATAAGC	2 μs
20	CGTCTTCTAGAGAGGC	500 ns	40	CGGACACTAGCGTCGC	2 μs

Table S2. Pearsons correlation coefficient of Shift, Slide and Twist at TpA with flanking bps parameters and selected backbone torsions up to hexamer.

		Shift at TA	Slide at TA	Twist at TA		Shift at TA	Slide at TA	Twist at TA
-2	Shift	0.06	0.002	0.025	zetaW	-0.067	-0.063	-0.123
	Slide	0.157	0.149	0.206	zetaC	-0.471	-0.286	-0.421
	Rise	-0.052	-0.022	-0.086	phaseW	-0.130	-0.023	-0.073
	Tilt	0.086	0.031	0.051	phaseC	-0.061	-0.079	-0.110
	Roll	0.001	0.043	0.038	chiW	0.018	0.002	0.025
	Twist	0.089	0.051	0.021	chiC	-0.074	-0.042	-0.057
-1	Shift	-0.607	-0.149	-0.257	zetaW	-0.454	-0.098	-0.217
	Slide	-0.298	0.089	-0.094	zetaC	0.753	0.295	0.536
	Rise	0.028	-0.089	-0.109	phaseW	-0.425	0.006	-0.105
	Tilt	-0.12	0.057	-0.11	phaseC	0.111	0.102	0.090
	Roll	0.002	0.178	0.157	chiW	-0.140	-0.027	-0.058
	Twist	-0.223	-0.263	-0.453	chiC	0.107	0.173	0.153
Central TpA step								
+1	Shift	-0.607	0.192	0.306	zetaW	-0.736	0.340	0.589
	Slide	0.201	0.098	-0.078	zetaC	0.456	-0.166	-0.260
	Rise	0.017	-0.08	-0.114	phaseW	-0.157	0.130	0.103
	Tilt	-0.104	-0.047	0.12	phaseC	0.431	-0.045	-0.144
	Roll	-0.045	0.176	0.173	chiW	-0.206	0.186	0.170
	Twist	0.232	-0.25	-0.455	chiC	0.166	-0.022	-0.053
+2	Shift	0.185	-0.084	-0.148	zetaW	0.547	-0.332	-0.487
	Slide	-0.251	0.195	0.271	zetaC	0.023	-0.023	-0.061
	Rise	0.09	-0.04	-0.103	phaseW	0.020	-0.072	-0.076
	Tilt	0.156	-0.091	-0.125	PhaseC	0.085	-0.004	-0.054
	Roll	0.012	0.044	0.039	chiW	0.019	-0.012	-0.018
	twist	-0.095	0.079	0.067	chiC	-0.067	0.006	0.024

SUPPORTING FIGURES

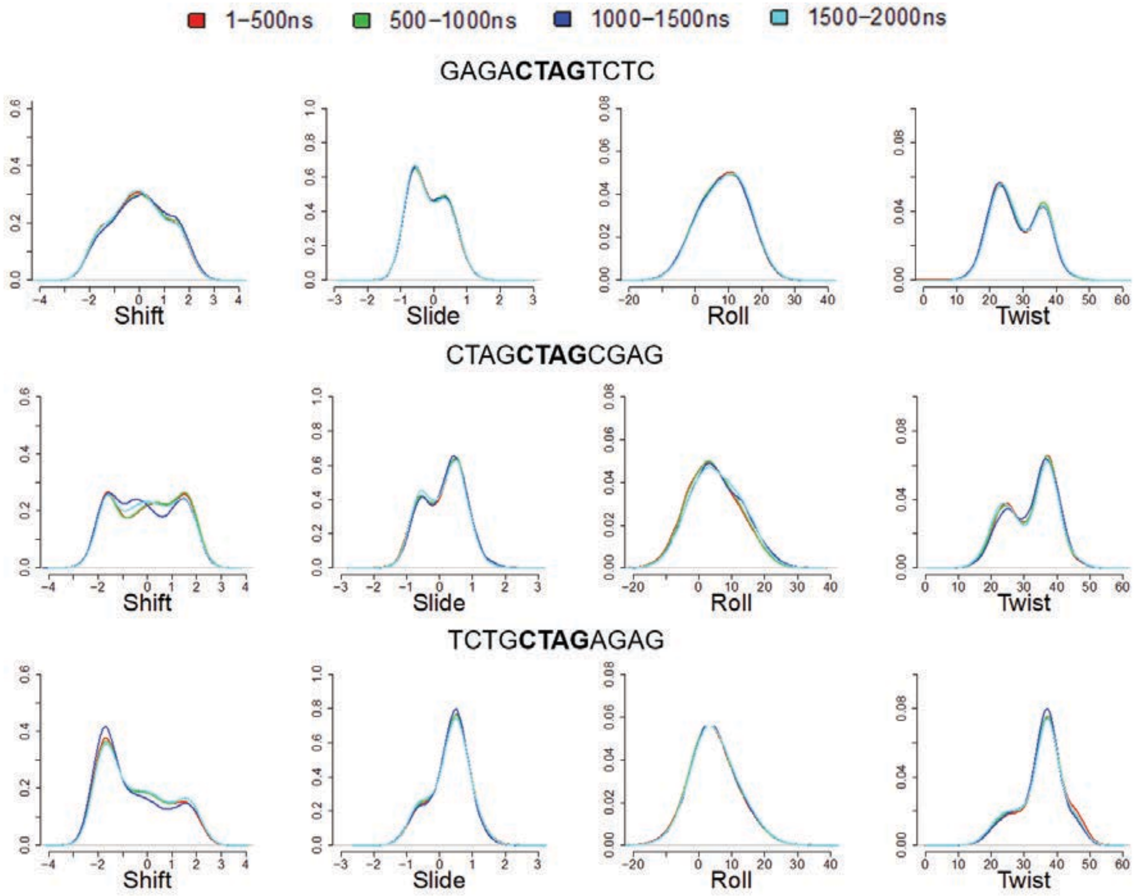


Figure S1. Normalized frequencies of the shift, slide, roll and twist helical parameters for 3 selected sequences, whose trajectories were extended to 2 μs to check for convergence. Four distributions were computed for each helical parameter using segments of 500 ns.

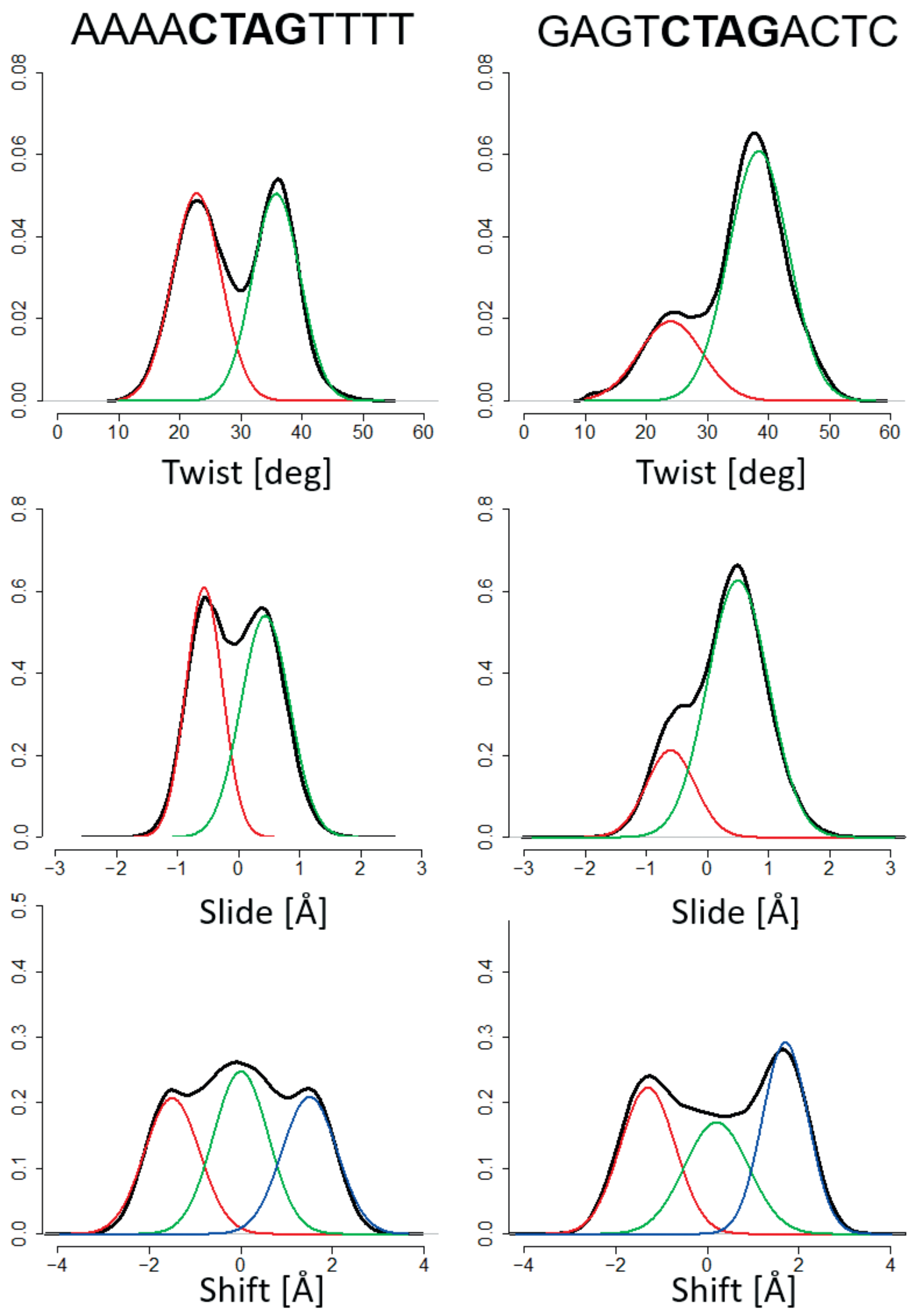


Figure S2. Normalized frequencies of the shift, slide, and twist helical parameters for 2 selected sequences showing clear hexamer effects, which could be appreciated from the change in the relative populations of the bi- and tri-normal distributions.

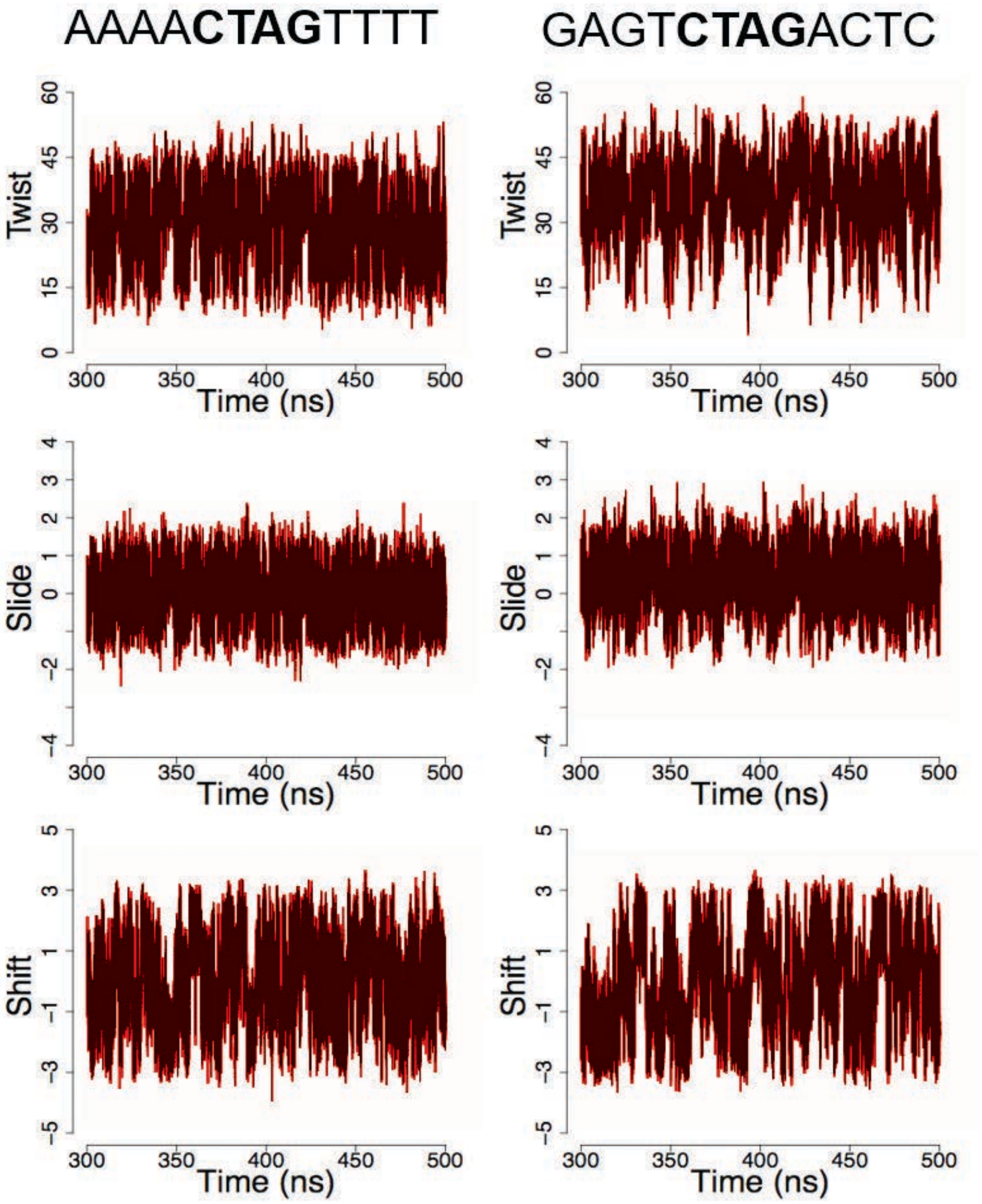


Figure S3. Time evolution (500 ns) of shift, slide and twist for two selected sequences, showing the fast and reversible inter-conversion between high and low substates.

5.

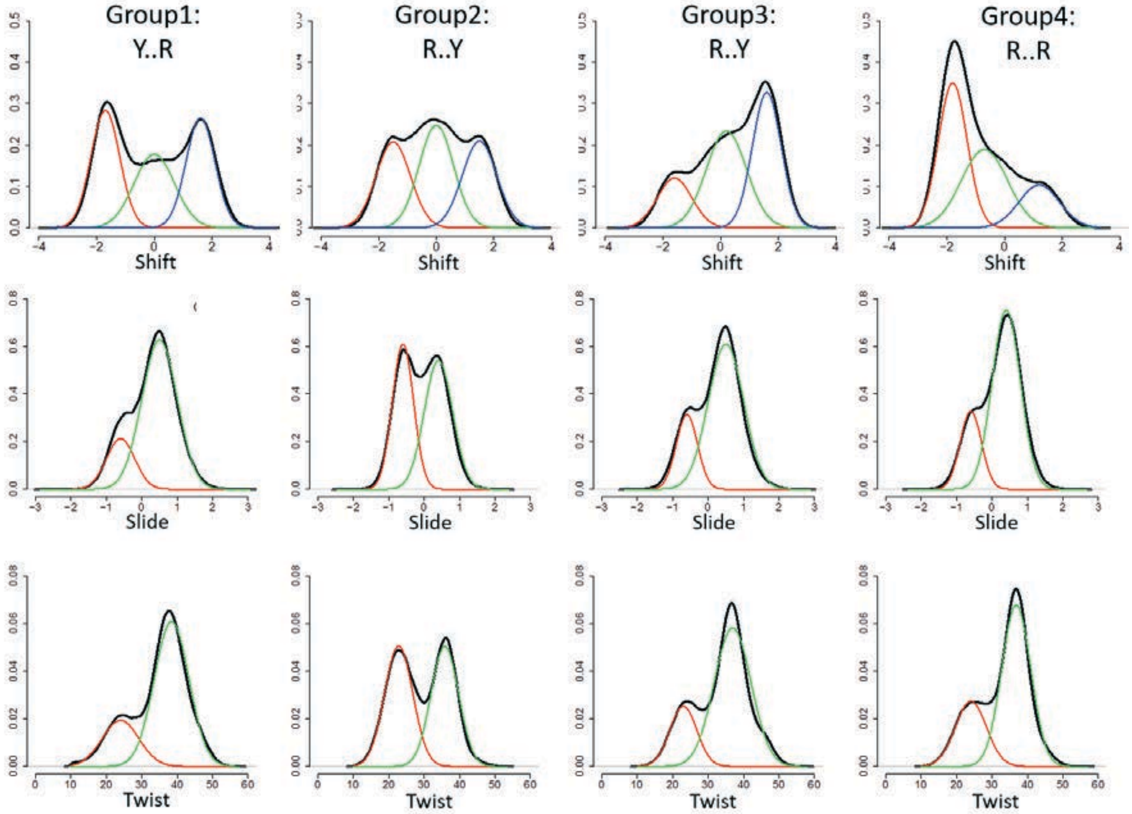


Figure S4. Normalized frequencies for shift, slide and twist (black line), and the BIC decomposition in Gaussians (red, green, and blue lines), showing the behavior of the clusters obtained in the dendogram of Figure 5.

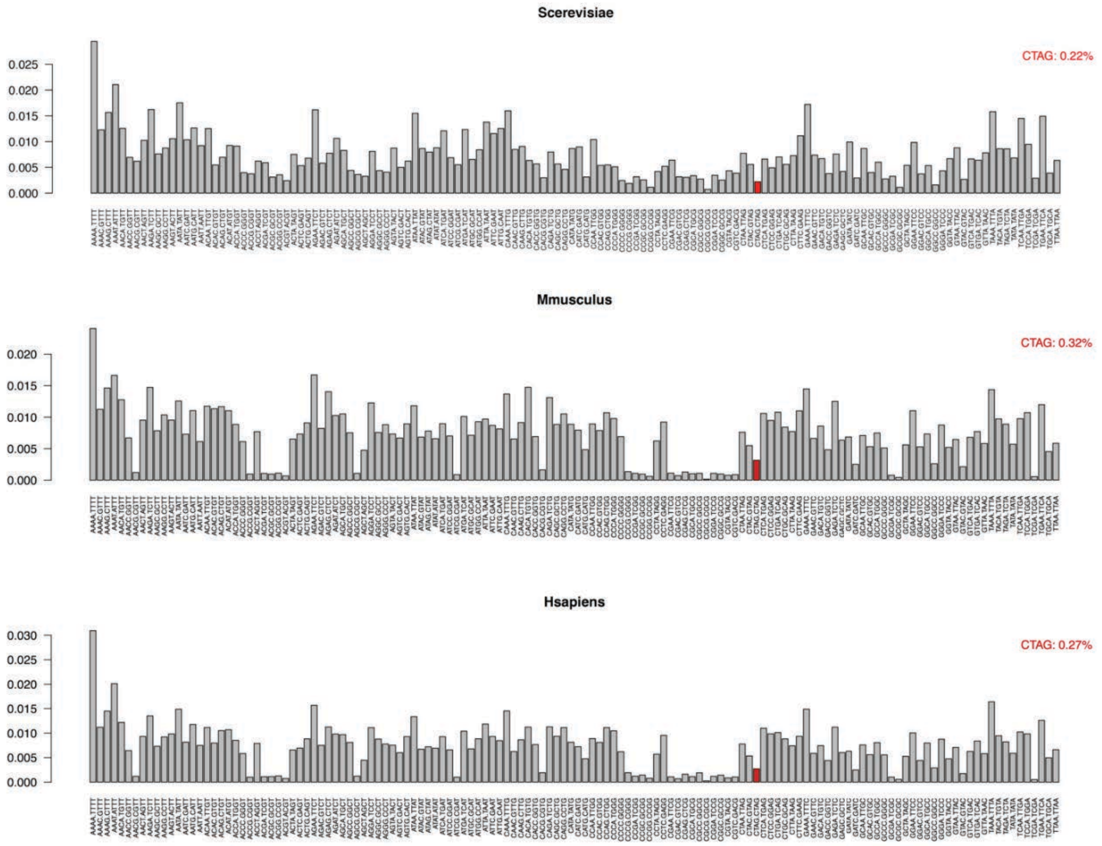


Figure S5. Occurrence of each possible tetranucleotide in 3 different genomes. CTAG is marked in red.

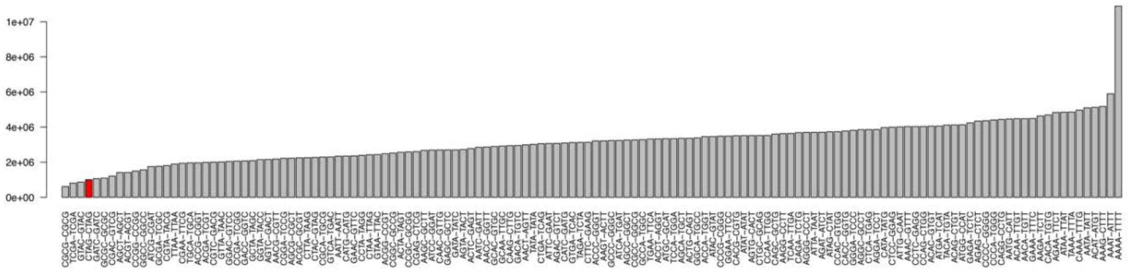


Figure S6. Occurrence of SNPs mapped to each unique tetranucleotide in Human genome. CTAG is marked in red.

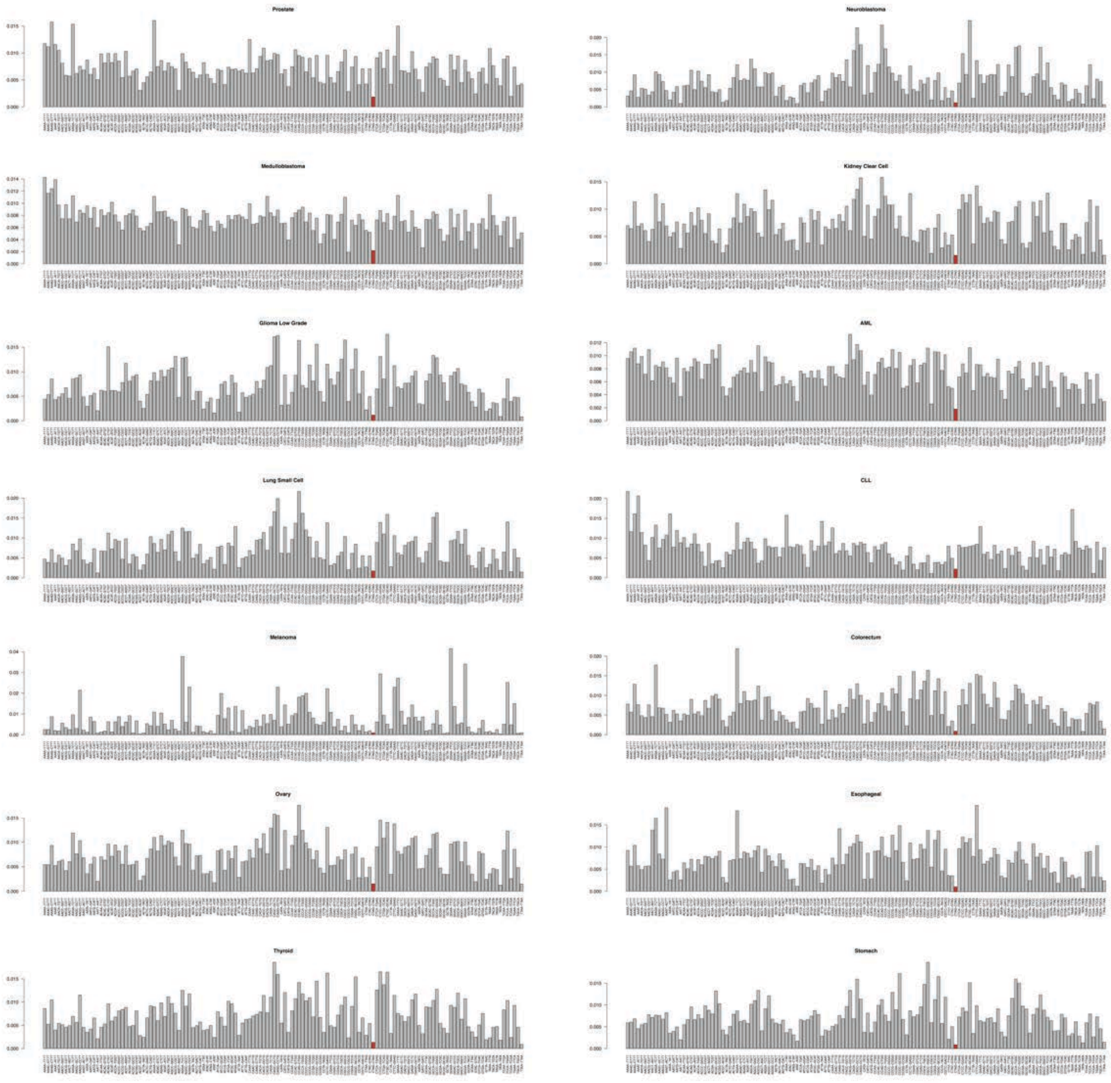


Figure S7. Frequency of mutations for each tetramer along the genome for several cancer types. CTAG is marked in red.

SUPPORTING REFERENCES

1. D.A. Case, R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, LX and PAK. AMBER 2016. 2016.

2. Le Grand S, Götz AW, Walker RC. SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput Phys Commun* 2013;**184**:374–80.

3. Pasi M, Maddocks JH, Beveridge D *et al.* μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res* 2014;**42**:12272–83.

4. Arnott S, Hukins DWL. Refinement of the structure of B-DNA and implications for the analysis of X-ray diffraction data from fibers of biopolymers. *J Mol Biol* 1973;**81**:93–105.

5. Berendsen HJC, Grigera JR, Straatsma TP *et al.* The missing term in effective pair potentials. *J Phys Chem* 1987;**91**:6269–71.

6. Smith DE, Dang LX. Computer simulations of NaCl association in polarizable water. *J Chem Phys* 1994;**100**:3757–66.

7. Dang LX. Mechanism and Thermodynamics of Ion Selectivity in Aqueous Solutions of 18-Crown-6 Ether: A Molecular Dynamics Study. *J Am Chem Soc* 1995;**117**:6954–60.

8. Dang LX, Kollman PA. Free Energy of Association of the K+:18-Crown-6 Complex in Water: A New Molecular Dynamics Study. *J Phys Chem* 1995;**99**:55–8.

9. Darden T, York D, Pedersen L. Particle mesh Ewald: An *N* ·log(*N*) method for Ewald sums in large systems. *J Chem Phys* 1993;**98**:10089–92.

10. Ryckaert J-P, Ciccotti G, Berendsen HJ. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* 1977;**23**:327–41.

11. Ryckaert J-P, Ciccotti G, Berendsen HJ. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* 1977;**23**:327–41.

12. Lavery R, Moakher M, Maddocks JH *et al.* Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res* 2009;**37**:5917–29.

13. Dršata T, Lankaš F. Theoretical models of DNA flexibility. *Wiley Interdiscip Rev Comput Mol Sci* 2013;**3**:355–63.

14. Schwarz G. Estimating the Dimension of a Model. *Ann Stat* 1978;**6**:461–4.

15. Kass RE, Raftery AE. Bayes Factors. *J Am Stat Assoc* 1995;**90**:773–95.

16. Ward JH. Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc* 1963;**58**:236–44.

17. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J Classif* 2014;**31**:274–95.

18. Dans PD, Faustino I, Battistini F *et al.* Unraveling the sequence-dependent

polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res* 2014;**42**:11304–20.

19. Balaceanu A, Pasi M, Dans PD *et al.* The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J Phys Chem Lett* 2017;**8**, DOI: 10.1021/acs.jpclett.6b02451.

20. Gelpí JL, Kalko SG, Barril X *et al.* Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins. *Proteins* 2001;**45**:428–37.

21. Fogolari F, Brigo A, Molinari H. The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J Mol Recognit* 2002;**15**:377–92.

22. Wang J, Wolf RM, Caldwell JW *et al.* Development and testing of a general amber force field. *J Comput Chem* 2004;**25**:1157–74.

23. Cuervo A, Dans PD, Carrascosa JL *et al.* Direct measurement of the dielectric polarization properties of DNA. *Proc Natl Acad Sci* 2014;**111**:E3624–30.

24. Jafilan S, Klein L, Hyun C *et al.* Intramolecular Base Stacking of Dinucleoside Monophosphate Anions in Aqueous Solution. *J Phys Chem B* 2012;**116**:3613–8.

25. Heinz S, Benner C, Spann N *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* 2010;**38**:576–89.

26. Alexandrov LB, Nik-Zainal S, Wedge DC *et al.* Signatures of mutational processes in human cancer. *Nature* 2013;**500**:415–21.

SUPPLEMENTARY DATA

Allosterism and Signal Transfer in DNA

Alexandra Balaceanu[1], Alberto Pérez[2], Pablo D. Dans[1] and Modesto Orozco[1,3,*]

[1] Joint IRB-BSC Program on Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain.

[2] Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, United States.

[3] Department of Biochemistry and Biomedicine, University of Barcelona, Barcelona, Spain.

SUPPLEMENTARY METHODS

Correlations between geometrical variables were evaluated taking into account the nature of the coordinates involved, either linear or circular. Correlation between two directional variables was assessed with the use of Jammalamadaka formula (4). Suppose a sample of n pairs of angles (a_{11} ,a_{21}),(a_{12} ,a_{22}),...,(a_{1n} ,a_{2n}) is available. The circular correlation coefficient is calculated as:

$$r_c = \frac{\sum_{k=1}^{n} \sin\left(a_{1k} - T_{1,1}\right) \sin\left(a_{2k} - T_{2,1}\right)}{\sqrt{\sum_{k=1}^{n} sin^2 \left(a_{1k} - T_{1,1}\right)} \sum_{k=1}^{n} sin^2 \left(a_{2k} - T_{2,1}\right)}$$

where $T_{1,1}$ is the mean direction of the first circular variable and $T_{2,1}$ is the mean direction of second. The correlation between a directional variable and a linear variable was assessed from the correlation of the linear variable with the cosine and sine of the circular variable individually. Thus, the circular-linear correlation coefficient was calculated as:

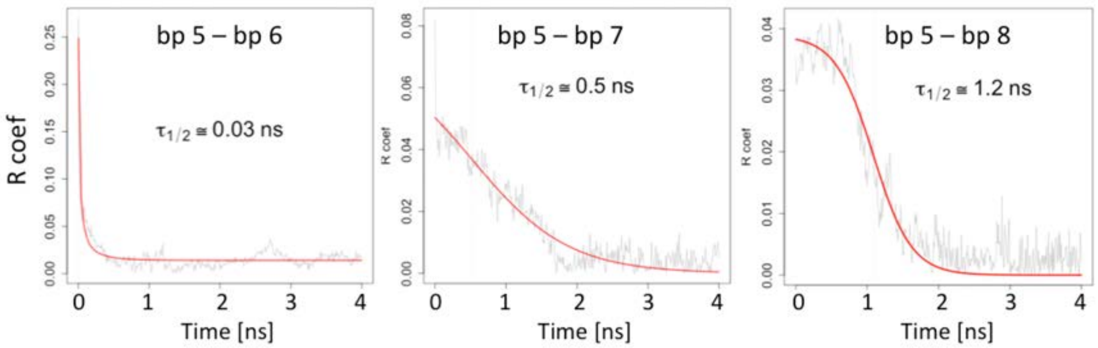
$$\boldsymbol{\rho_{cl}} = \sqrt{\frac{\boldsymbol{r_{cx}^2 + r_{sx}^2 - 2r_{cx}r_{sx}r_{cs}}}{\boldsymbol{1 - r_{cs}^2}}}$$

Where r_{sx}=c(sin α, x), r_{cx}=c(cosα, x) and r_{cs} = c(sin α, cosα), with c(x, y) being the Pearson correlation coefficient.

Delayed Correlations can be described for two time series r_i and r_j considering that $\Delta r_j(t+\tau)$ may be affected by the earlier fluctuations of $\Delta r_i(t)$. The extent of this effect may be quantified by the time delayed correlation function

$$C_{ij}(\tau) = \frac{\sum \Delta r_i(t)\Delta r_j(t+\tau)/N}{\sqrt{< \Delta r_i^2 >< \Delta r_j^2 >}}$$

This leads to directionality in the structure, and because of the asymmetry $C_{ij} \neq C_{ji}$, the delayed correlations can capture causality between the evolutions of two time series. It is however difficult to assess how fast the signal travels along the DNA and therefore which time delay to use at each distance. Using simply the maximum value of the correlation over a certain allowed interval does not yield an incremental sequence of time values because of large noise interference at far away distances. For this reason, we relied on the assumption of a constant speed traveling wave and determined the values of the specific time delays at each position along the DNA from the first 3 strongest signals. We took as example the base pair 5 as the source of perturbation because of its strong contact with protein residues. We fitted the cross-correlation between the major groove width at bp 5 and that at the next three positions as a function of time displacement(allowing for up to 4ns delay)to a logistic decay and calculated for each the half-live. From these 3 values we calculated the slope and considered it to be the speed of the traveling wave to determine the most likely specific time delay at each position.



Correlations network analysis of backbone torsions was performed by computing all circular-circular correlation between backbone angles with absolute values above a threshold of 0.4. Circular-circular correlations as described above were used to build interaction networks of coupled backbone motions in the double helix. A trajectory was build where each base pair was represented by a set of Watson and Crick dihedral angles (alpha, beta, epsilon, gamma and zeta – a set of 10 nodes per base pair) and edges were defined from pairwise circular-circular correlations above the selected threshold. The backbone torsions of both strands were taken into consideration excluding 4 base pairs at each end. Only non-redundant correlations were retained, by removing dependencies between torsions belonging to the same base pair and those between the same dihedral at adjacent bases on the same strand. The data collected was structured and represented as a descriptive network graph using the R package igraph (5). The structure of the network was fixed so that it followed the unraveled DNA helix from 5' to 3'. This method is extremely useful for the visualization of the new communication pathways that appear upon protein binding.

Classical molecular interaction potentials were computed (represented at the -4 kcal/mol contour level) to determine the changes in recognition properties induced by BAMHI binding on the region of DNA that binds GRDBD. Calculations were performed using MD-averaged structures obtained using the last 100 ns of production simulations. Following the defaults in our CMIP code (6), the electrostatic interaction term was determined by solving the linear Poisson–Boltzmann equation, while the van der Waals contribution was determined using standard AMBER Lennard–Jones parameters. The ionic strength and the reaction-field dielectric constant were set to 0.15 and 78.4 M, respectively, while the dielectric constant for DNA was set to 8 following our previous experimental and theoretical estimates of the dielectric response of DNA (7). A protonated methylamine was used as probe particle moving in a 0.5 Å spaced grid. CMIP was also used to calculate the protein-protein interaction in the BAMHI-DNA-GRDBD trimer from the last 10^4 frames of the trajectory.

Cation analysis was performed by determining the cation distribution in curvilinear cylindrical coordinates. The distribution of sodium cations around the DNA was determined from the last 200 ns of each MD trajectory, and analyzed using the latest version of Curves+ (2) that has the ability to calculate the position of ions along a trajectory using a curvilinear cylindrical (CHC) system with respect to the instantaneous helical axis of each snapshot. This data was then analyzed with the Canion utility program (8) and accumulated ion positions in terms of its longitudinal coordinates which describe the position of the ion along the DNA molecule were obtained, accounting for the coupling between ion motions and the global deformations of the DNA molecule. The ion distributions are given in units of molarity. The limits of the grooves are defined according to the default values used in Canion (8,9) and we report on the ion distribution within the internal region of the minor groove as defined there.

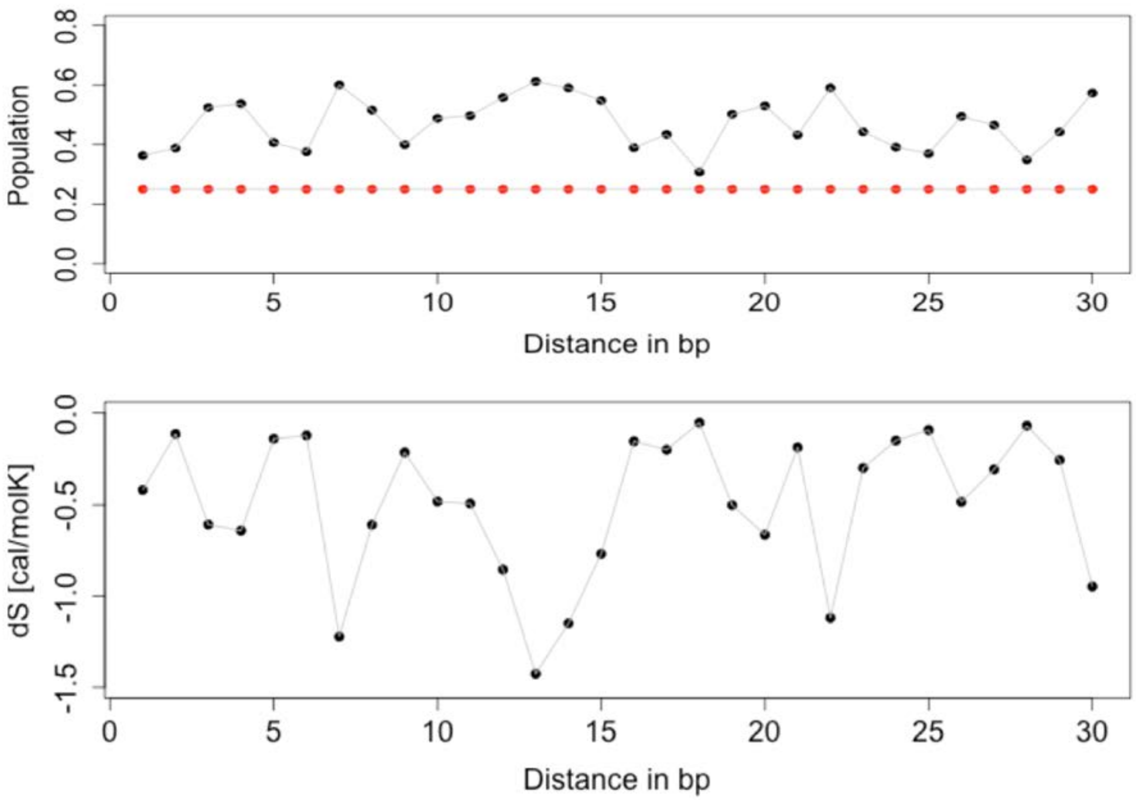
Protein contact analysis "Protein sensing" contacts are defined as pairs of amino-acid/nucleotide that when coming in close proximity to each other (distances between centers of mass below 7 Å and at least one atom pair distance below 3 Å) produce the most significant perturbations in the DNA consistently for complexes of all linker sizes. Selecting for the presence of the combination of these contacts yielded meta-trajectories consisting of at least 10,000 disperse collected frames. These are not always the native contacts, but follow a touch&release pattern all throughout the well equilibrated trajectories (beyond 200 ns).

Entropy calculations were performed using both the Schlitter and Andricioaei/Karplus methods, which give entropy estimates from the diagonalization of the mass-weighted Cartesian covariance matrix (taken from the unbiased simulations) and the application of a modified quantum statistical harmonic oscillator model to the associated quasiharmonic frequencies (10,11). We calculated absolute entropies of the DNA heavy atoms in the secondary binding region, following the two different transformations from Cartesian space atomic oscillations. Since the calculated total entropies are sensitive to the length of the simulation and, to a degree, to the fitting method as well, we checked the robustness of our results by using 2 different methods to align the duplexes (rms-fitting the entire DNA and only the GRDBD binding region) and extending the analyzed time windows of the trajectory. In any case, individual estimates of entropy differences remain very similar after 500 ns as can be seen in Supp. Fig. 7, suggesting the lack of severe convergence problems in results shown in Figure 5 of the main text where the reported values are obtained using Harris' extrapolation scheme (12) to infinite simulation time taking individual estimates in the 50ns-1μs range.

We additionally assess the **dihedral entropy** at the interface between linker and secondary binding regions with a Kullback–Leibler divergence (13) entropy measure. We follow the method described by Cukier (14) to quantify how far from independent the dihedral states are across the analyzed region. If X is the set of all interface (10 base pairs – last 5 bp of each linker region plus the first 5 bp of the GRDBD binding site) backbone conformations as combinations of states for the interface dihedrals, we calculate the Kullback-Leibler divergence for both naked and bound DNA for each linker size as:

$$D(p||p^{ind}) = \sum_{x \in X} p(x) log \frac{p(x)}{p^{ind}(x)}$$

Then we compute the difference $\Delta S_{dihedral} = k_B T[D(1) - D(0)]$ where 1 stands for the bound complex and 0 is the unbound DNA as a measure of dihedral entropy change upon effector protein binding. We acknowledge that the KLD does provide a quantitative measure of the difference between two probability distributions, in this case the observed probability versus the assumed independent probability, but it is not entirely obvious how to interpret the magnitude of this measure. Because $D(p \| p_{ind}) = -\Delta H$ and $\Delta S = R\Delta H$ (where R is the gas constant), we feel the need to attempt at providing a rough reference numerical values of the dihedral entropies calculated with this method in a simplified model that is still similar to our complicated multidimensional calculation. We therefore calculate the associated difference to the independent entropy of pairs of zeta backbone dihedrals along one strand of DNA with increasing distance. The difference in entropy between pairs of dihedrals that show one highly populated state and those for which all states are close to the independent population is of about 1.4 cal/molK in systems with 4 possible states (2 per dihedral). The figure below shows these results: top – population of highest populated state for each dihedral pair and bottom, the entropy change of the corresponding pair (considering all state populations).



<u>Entropy Transfer</u> as described in Schreiber's work (15), we considered the transfer entropy (TE) $T_{i\text{->}j}(\tau)$ from time series *i* to *j* at time τ as:

$$T_{i\rightarrow j}(\tau) = \; S\left(\Delta r_j(t+\tau)\middle|\Delta r_j(t)\right) - S(\Delta r_j(t+\tau)|\Delta r_i(t), \Delta r_j(t))$$

where the conditional entropy of two events separated in time by τ is defined by:

$$S\left(\Delta r_j(t+\tau)\middle|\Delta r_i(t)\right) = -\sum p\left(\Delta r_i(t)\middle|\Delta r_j(t+\tau)\right)\ln p\left(\Delta r_j(t+\tau)\middle|\Delta r_i(t)\right)$$

To estimate the quantities in the equation above, we express conditional probabilities p(j|i) by the Kolmogorov definition, as the quotient of the probability of the joint of events i and j, and the probability of i. In this sense, the TE estimation can be treated as a problem of density estimation from data. One of the most straightforward approaches to density estimation is based on histogram models, which is the method we employ here. One-, two- and three-dimensional histogram functions were used to cluster data into bins with varying widths and partitioning of data is adaptive according to the maximum and minimum of data. Number of bins was selected according to the Sturges' rule. The optimum number of bins is calculated from the Sturges' rule according to

$$\mathrm{n_{opt}} = \text{mean fluctuation} \cdot (1 + \log_2 \mathrm{N})$$

The mean fluctuations refer to the average fluctuations of major groove width divided by their maximum value in the base pairs considered. We take Δri(t) as the magnitude of the major groove width fluctuations at time t. After the estimation of the optimal number of bins and assuming that the trajectory of the system can be approximated by a stationary Markov process, the TE can be expressed as a summation of Shannon entropy terms:

$$\mathrm{T}_{\mathrm{i}\to\mathrm{j}}(\tau) = \mathrm{H}\left(\Delta \mathrm{r_j}(0), \Delta \mathrm{r_i}(0)\right) - \mathrm{H}\left(\Delta \mathrm{r_j}(\tau), \Delta \mathrm{r_j}(0), \Delta \mathrm{r_i}(0)\right) + \mathrm{H}\left(\Delta \mathrm{r_j}(\tau), \Delta \mathrm{r_j}(0)\right) - \mathrm{H}\left(\Delta \mathrm{r_j}(0)\right)$$

In general, $\mathrm{T}_{\mathrm{i}\to\mathrm{j}}(\tau) \neq \mathrm{T}_{\mathrm{j}\to\mathrm{i}}(\tau)$and this will determine the net transfer of entropy from one event to another separated in time by τ. In this study, we tookτ = 2 ns as the representative correlation time of cross correlations based on the characteristic decays depicted in Figure 3. The values of $\mathrm{T}_{\mathrm{i}\to\mathrm{j}}(\tau)$ can vary from zero when the fluctuations of time series i and j are independent, and a maximum of the entropy rate:

$$\mathrm{h}\left(\Delta \mathrm{r_j}(\tau) \middle| \Delta \mathrm{r_j}(0)\right) = -\sum \mathrm{p}(\Delta \mathrm{r_j}(\tau), \Delta \mathrm{r_j}(0) \log \mathrm{p}\left(\Delta \mathrm{r_j}(\tau) \middle| \Delta \mathrm{r_j}(0)\right))$$

when the fluctuations of i and j are completely coupled; this maximum is reached when i = j, for example. High values of $\mathrm{T}_{\mathrm{i}\to\mathrm{j}}(\tau)$ indicate that the fluctuations of base pair i are strongly driving the fluctuations of base pair j, whereas low values indicate a smaller or null dependence. To quantify whether, on average, the motion of base pair j drives the motion of base pair i, or whether the motion of base pair i drives the motion of base pair j, we use the net entropy transferred from base pair i to base pair j, normalized by the entropy rate, with values between -1 and 1:

$$T_{i\to j}^{Net} = \frac{T_{i\to j}(\tau)}{\mathrm{h}\left(\Delta \mathrm{r_j}(\tau) \middle| \Delta \mathrm{r_j}(0)\right)} - \frac{T_{j\to i}(\tau)}{\mathrm{h}(\Delta \mathrm{r_i}(\tau) | \Delta \mathrm{r_i}(0))}$$

This is the quantity represented in the color maps of Figure 7 in the main text. We also calculate the net entropy transferred from base pair i to all other base pair, which is obtained by summing over all *j* as:

$$T_i^{Net} = \sum_{j=1}^{N} (T_{i\to j}(\tau) - T_{j\to i}(\tau))$$

Knowing the major groove width of base pair i and j at time t, we employ the model to evaluate to what extent the uncertainty of the future major groove widths values of atom j was reduced. From this point of view it is the correlation in the fluctuations of the two atoms that causes the decrease in entropy. When major groove width fluctuations are correlated, transfer entropy between residues i and j does not equal the transfer entropy between residues j and i. This explicit asymmetry in transfer entropy can be used to distinguish between the residues that drive the correlated motion (entropy sources) and the residues that respond (entropy sinks); *i.e*. between the cause and the effect of correlated motions. If $\mathrm{T}_{\mathrm{i}\rightarrow\mathrm{j}}^{\mathrm{Net}}$ is greater than zero, we say that the fluctuations of atom i drive those of atom j, or i is the source of correlations. Since $\mathrm{T}_{\mathrm{i}\rightarrow\mathrm{j}}^{\mathrm{Net}}$ is based on mutual information, the measure includes all linear and nonlinear correlations between the two time series.

Free energy of binding calculation. To estimate the free energy of binding GRDBD on the DNA we follow the Confine-Convert-Release (CCR) method described by Roy *et al.* (16), which stems out from previous confinement methods of Tyka *et al.* and Cecchini *et al.* (17,18). We follow the steps shown in the thermodynamic cycle of Figure S8 to compute separately the free energy of binding GRDBD on the naked DNA and on the BAMHI-bound DNA. In order to obtain these estimations, we first calculate the energy of confining each structure to its energy minimum by thermodynamic integration with increasing restraints solute atoms from 5x10^{-5} to 81.92. The negative of this energy is the release term. In the convert step that completes the thermodynamic cycle, we calculate the energy difference of the DNA atoms between the two highly restrained complexes. Finally, the total binding energy is calculated from the sum of these individual contributions and ΔΔG responsible for cooperative effects is estimated. CCR methods have been successfully used for calculating free energies related to conformational changes in proteins. However, in our particular case the free energy change calculated with CCR is highly overestimated because we apply the method to compute binding free energies where the number of atoms changes during the thermodynamic cycle. Even though we use only the DNA atoms to compute energy differences between the two confined states (with and without the GRDBD), the presence of the protein makes some differences and therefore we consider this agreement with experimental data to be only qualitative and we treat its importance and relevance in the text accordingly. It is beyond the scope of this paper to validate the CCR method for binding free energies.

SUPPLEMENTARY TABLES

Table S1. Interaction free energy of a protonated methylamine probe (used to simulate the presence of a charged amino acid sidechain) at the secondary binding site.

	Interaction free energy (Electrostatics + vdW) in kcal/mol		
	Free	Bound	Bound Conf
7bp Linker	-7.1 ± 0.3	-7.1 ± 0.4	-7.1 ± 0.4
11bp Linker	-7.4 ± 0.5	-7.5 ± 0.3	-7.5 ± 0.3
15bp Linker	-7.1 ± 0.5	-6.9 ± 0.5	-6.9 ± 0.5

SUPPLEMENTARY FIGURES

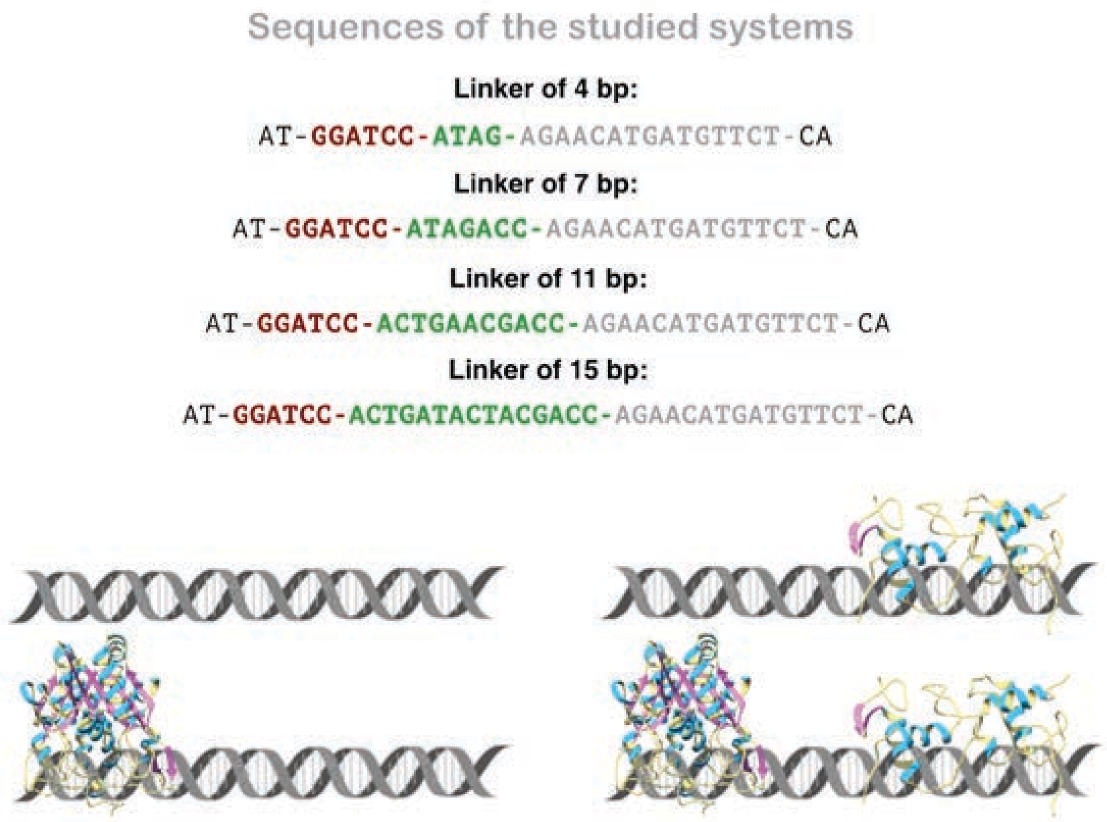

Sequences of the studied systems
Linker of 4 bp:
AT-GGATCC-ATAG-AGAACATGATGTTCT-CA
Linker of 7 bp:
AT-GGATCC-ATAGACC-AGAACATGATGTTCT-CA
Linker of 11 bp:
AT-GGATCC-ACTGAACGACC-AGAACATGATGTTCT-CA
Linker of 15 bp:
AT-GGATCC-ACTGATACTACGACC-AGAACATGATGTTCT-CA

Figure S1. Watson strand sequence of the simulated systems, from 5'to 3'end of duplex DNA containing the canonical BAMHI binding site (d(GGATCC)) and the canonical GRDBD binding site (d(AGAACATGATGTTCT), labeled as the "reverse" sequence in ref. 15 of the main text) separated by linkers of increasing length (4, 7, 11 and 15). In all cases four systems were simulated: the naked DNA, the BAMHI-DNA complex, the GRDBD-DNA complex and the BAMHI-DNA- GRDBD trimer. The 4x4 systems were created using Nucleic Acid Builder and standard B-DNA geometrical parameters, except for the binding region where the geometries were transferred from the respective crystal structures (PDB codes 2BAM and 1R4R). Left top: naked DNA; right top: GRDBD-DNA; left bottom BAMHI-DNA; right bottom: BAMHI-DNA-GRDBD.

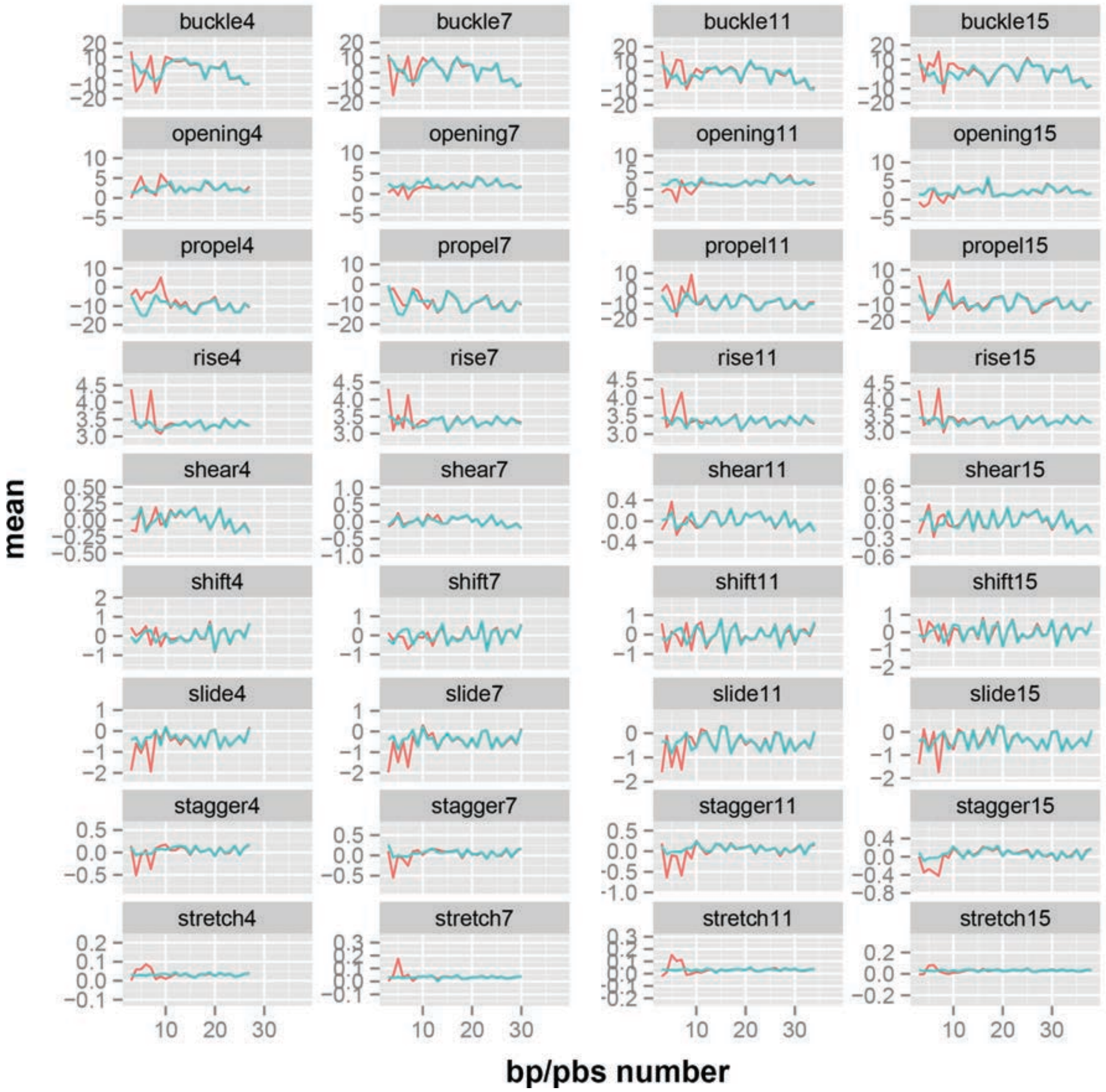


Figure S2. Average (0.2-1 μs) inter and intra base-pair parameters (rotational in degrees, translational in Å) for the DNA duplexes with linkers of different lengths. Values for free DNA (cyan) versus those for BAMHI-bound DNA (red).

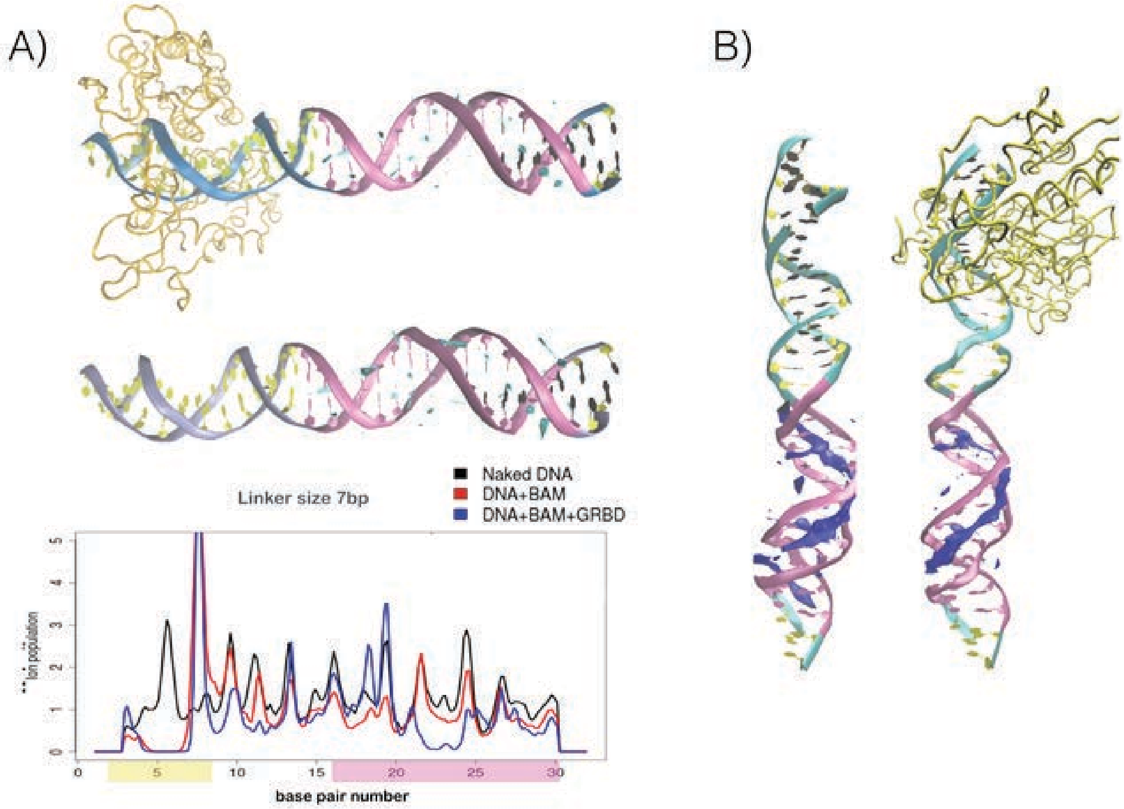


Figure S3. A) Top: Average water density in the GRDBD binding region of a 2 Å spaced grid over the last 30 ns of simulation, where both solute and ions/water are fully equilibrated. Isosurface of water densities of 1.5 times the bulk water density (1 g/ml). The naked and the bound DNA structures were rms-fit to the average structure of naked DNA in the GRDBD binding region. The recognition site of the second protein is shown in magenta. Bottom: Cation distribution along the helical axis in curvilinear cylindrical coordinates; profiles shown for naked DNA (black), BAMHI-bound DNA (red) and the system with both proteins bound to the DNA (blue).B) CMIP calculation of the naked and BAMHI-bound DNA interaction at the secondary binding site with a protonated methylamine. Calculation performed using average structures obtained from the last 100 ns of the production simulations on a 0.5 Å spaced grid.For the sake of comparison, the two averaged structures were aligned and the same isosurface of -4 kcal/mol was computed (see Suppl Table S1 for quantitative results).

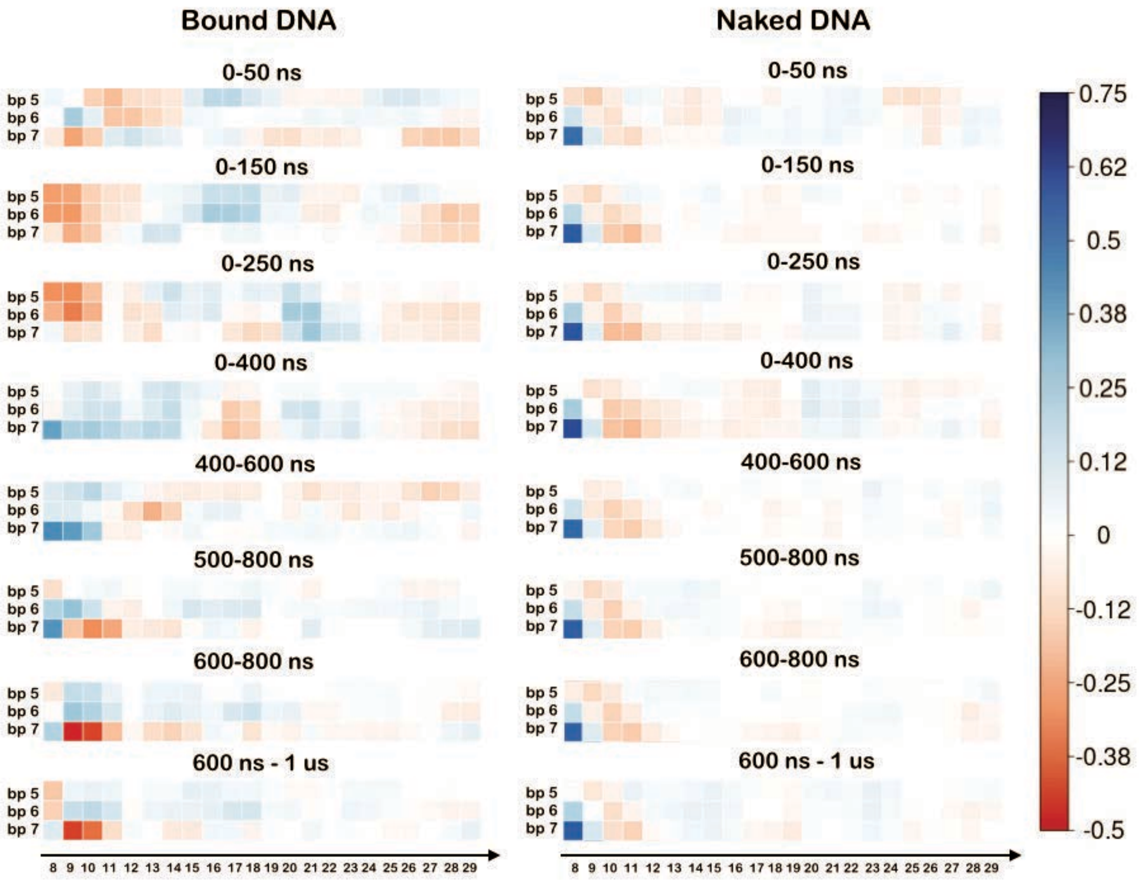


Figure S4. Major groove width correlation from the BAMHI binding site centre along the sequence for several time windows over the trajectory.Comparison between BAMHI-bound DNA and naked DNA. Increasing windows of time are chosen in order to capture the evolution over time of these correlations, since short simulations tend to overestimate correlation strengths due to insufficient sampling and equilibration artifacts. However, even when allowing for long equilibration times, the communication between the two binding sites is noticeable compared to the unbound DNA.

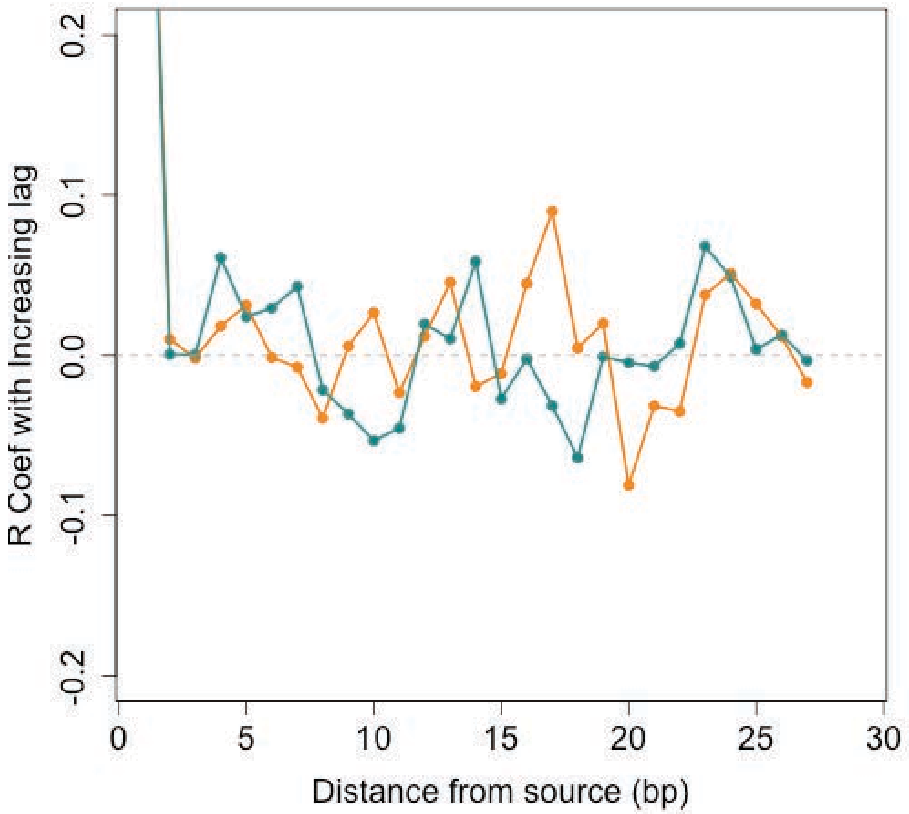


Figure S5. Time-delayed cross correlation of major groove widths between base pair 5 belonging to the BAMHI binding regions and all subsequent bases in all linker size system. The major groove width of the 15 bp linker naked DNA at base pair 5 has been gradually pulled open by harmonic restraints. The abscisa denotes the distance in base pairs from the perturbation source (bp 5). “Forward” correlations (5’ to 3’ on Watson strand, corresponding to cross-correlations from bp 5 to all other bases) are depicted in orange, whilst “reverse” correlations (3’ to 5’ direction, referring to cross-correlations from each bp to bp 5) appear in blue.

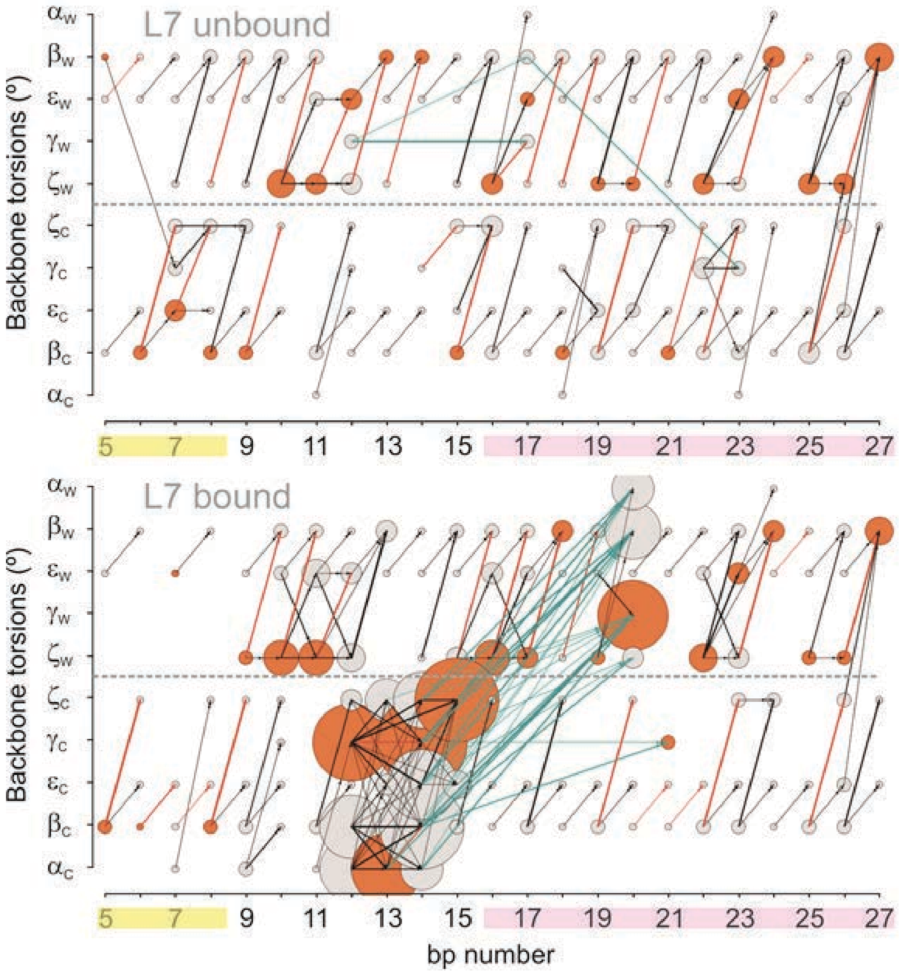


Figure S6. Correlations network analysis showing through space propagation of correlated motions in the DNA backbone above an R coefficient absolute value threshold of 0.4 for naked and BAMHI-bound DNA (case of 7bp linker) from the center of the BAMHI binding site along the sequence. To reduce the noise correlations between torsions at the same base pair have been removed. In the graph, each level represents a base pair, and the vertexes are ordered on the y-axis according to the backbone torsion they represent (both the Watson (W) and Crick (C) strand torsions are shown belonging to each base pair). Vertex size is proportional to their degree (the number of connections) and for each level, the vertex with the highest degree is colored in orange, all others in grey. Correlations between levels more than 4 base pairs apart are depicted as cyan arrows. From each level, the strongest correlation with a starting point on that level is shown as an orange arrow, unless it is already colored in blue. Protein binding sites are highlighted in yellow and magenta.

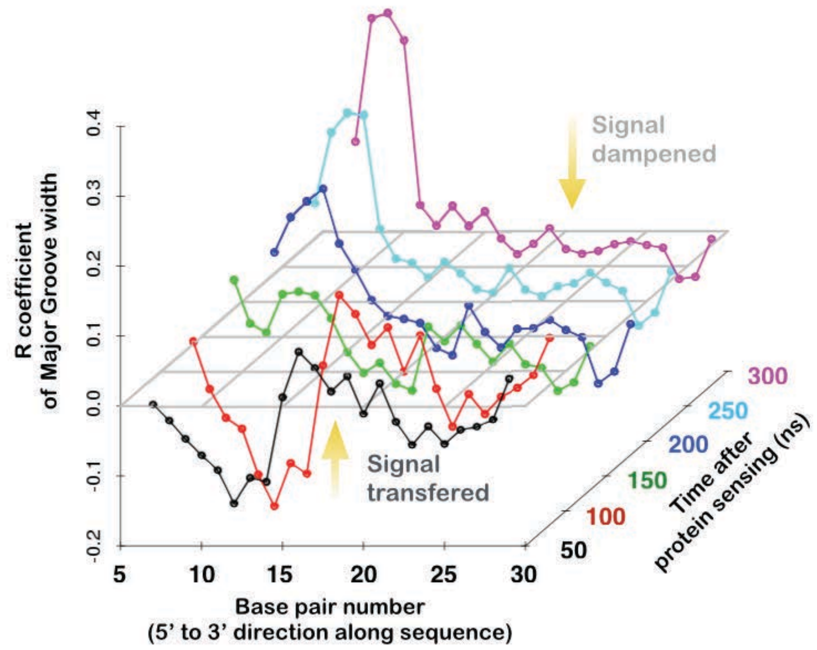


Figure S7. Time evolution of major groove width correlation averages from the first binding site along the sequence after a protein clenching event. The first two curves represent correlations when the contact is active and propagate a long distance, up to the secondary binding site, while after "release" the BAMHI influence is contained within a few bases of its binding site.

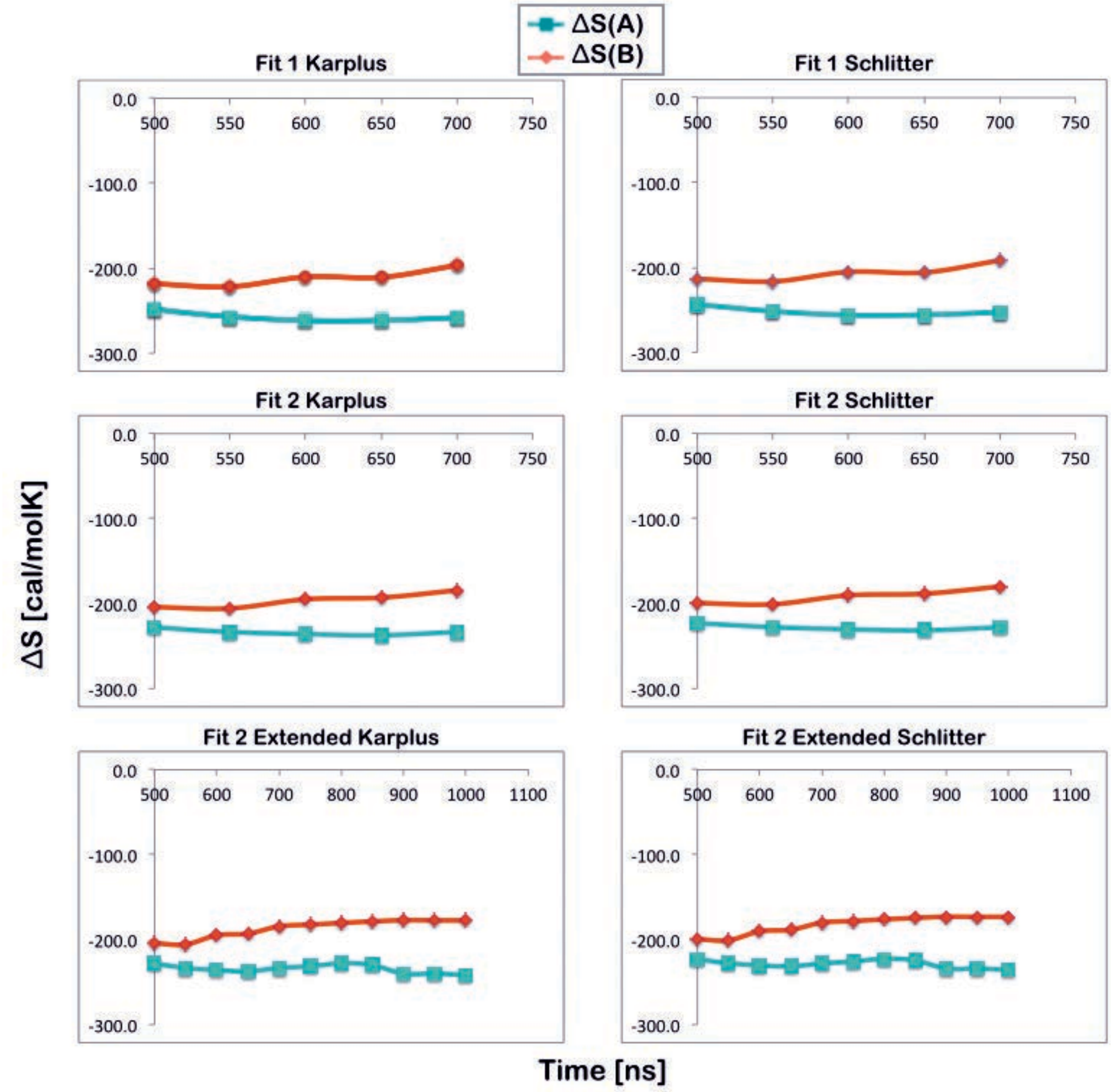


Figure S8. Entropy change upon GRDBD binding on the Naked ΔS(A), and on the bound ΔS(B) DNA. Two alignment methods were used to define the average structure (Fit 1 and Fit 2), and estimates were obtained for different time-windows, showing a stable value of the entropy differences after 500ns irrespective of the procedure followed to transform oscillations into entropy measures (Schlitter or Andricioaiei/Karplus), or even of the extension of the trajectory.

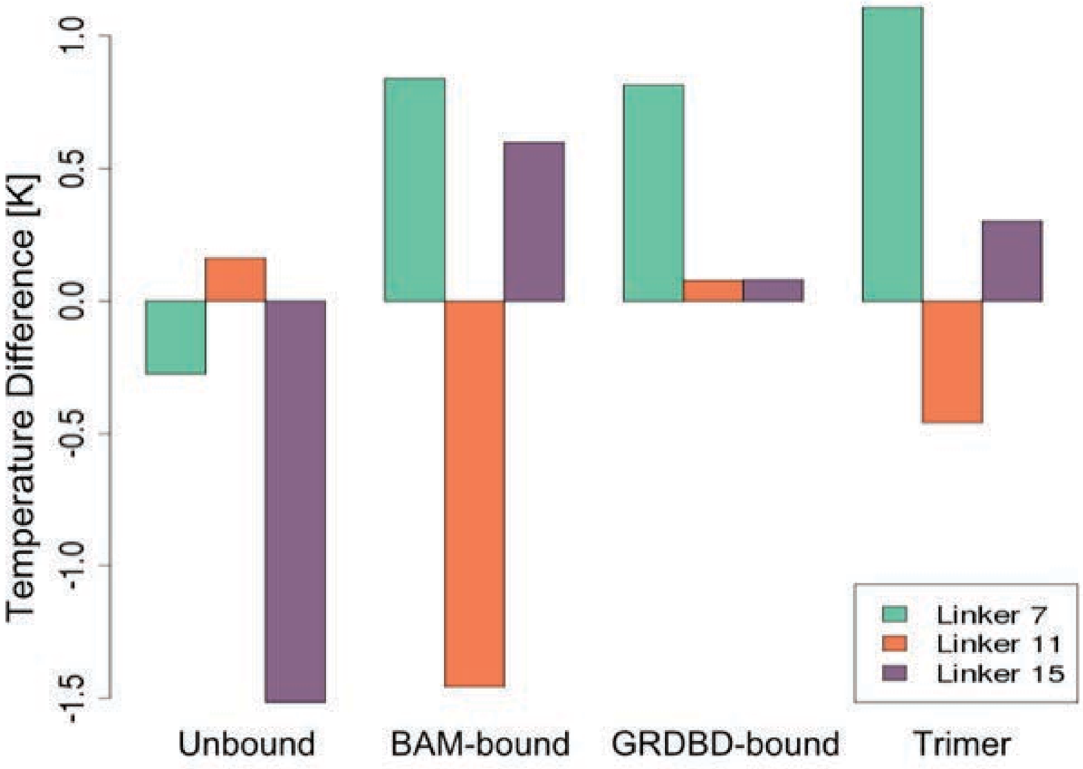


Figure S9. Difference in effective temperature (computed from atomic oscillations) between full DNA duplex and GRDBD region for all 4 different stages of binding: naked DNA, BAMHI-bound, GRDBD-bound and both proteins bound to form the trimer complex. Results support the entropic mechanism behind the cooperative effect. The temperature decrease in the GRDBD region compared to the entire DNA (positive difference) is observed when the first protein binds to the double helix, but only in the cases of the two favorable (cooperative) linker sizes (7 and 15 bp) and not for the 11 bp linker (anticooperative).

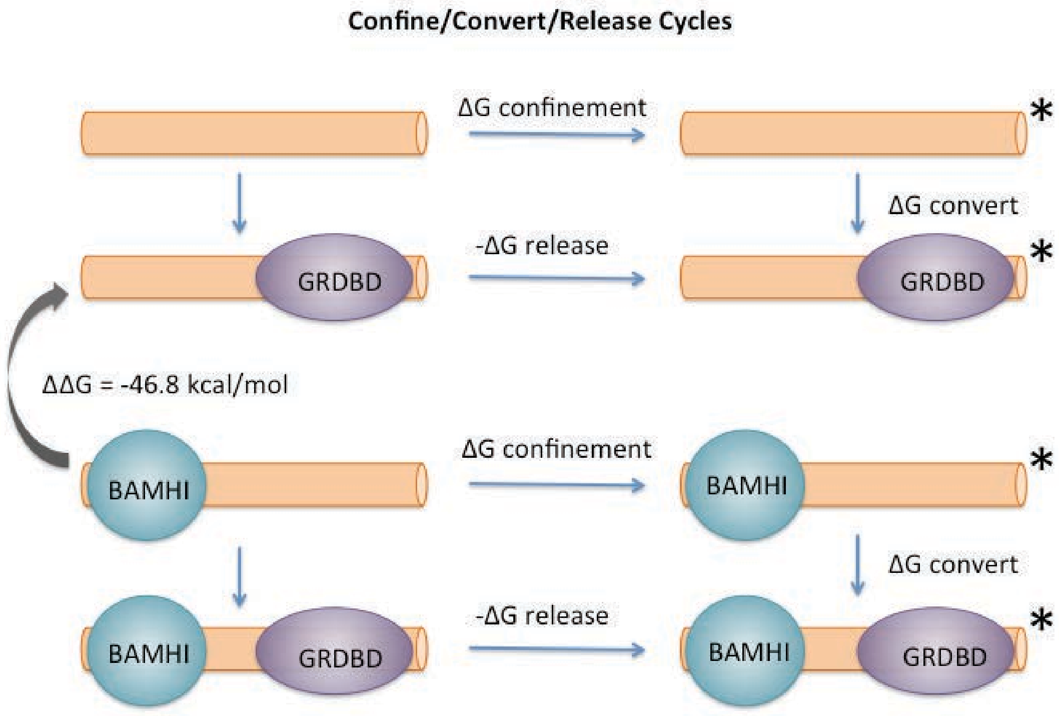


Figure S10. Thermodynamic cycles used for calculating of binding free energies using the Confine-Convert-Release method. The cooperative effect is estimated to increase stability for the binding of GRDBD in the presence of BAMHI by46.8 kcal/mol (ΔΔG = -46.8)for the linker of 7 bp, respect to the binding of GRDBD to naked DNA.Note that this result should be taken as qualitative. See further details and explanations in the Supplementary methods.

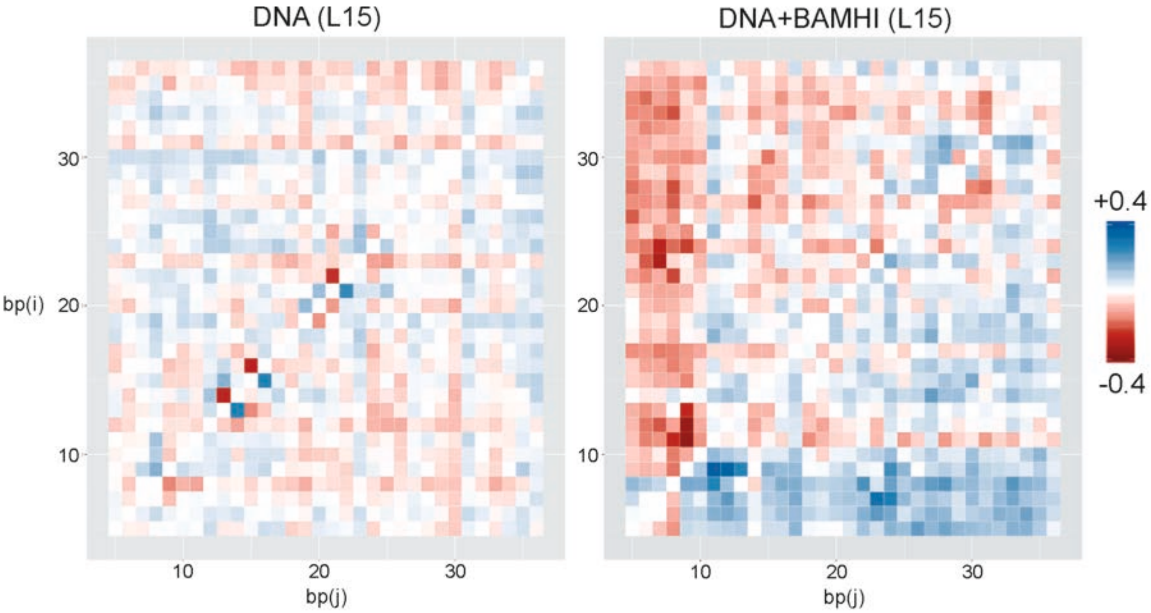


Figure S11. Map of entropy transfer from base pair *i* to base pair *j* in the naked (left) and BAMHI-DNA (right) systems. The normalized net entropy ($T^{NET}{}_{i\to j}$) transferred from residue *i* to residue *j* is obtained as described in the Methods Section of Supplementary Data. Positive values of$\mathrm{T}_{\mathrm{i}\to\mathrm{j}}^{\mathrm{Net}}$ in blue indicate that the major groove fluctuations of bp *i* strongly drive the major groove fluctuations at bp*j*(in the 5'-3' direction), whereas low values (white) indicate a smaller or null dependence, and negative values (red) a correlation from *j* to *i*.

SUPPLEMENTARY REFERENCES

[1] Case, D.A.;Babin, V.; Berryman, J.T.; Betz, R.M.; Cai, Q.;Cerutti, D.S.; Cheatham, T.E.; Darden, T.A.; Duke, R.E.; Gohlke, H. AMBER 14. University of California: San Francisco, 2014

[2] Lavery, R.;Moakher, M.; Maddocks, J.H.;Petkeviciute, D.; Zakrzewska, K. Nucleic Acids Res. 2009, 37, 5917– 5929.

[3] Hospital, A.; Faustino, I.;Collepardo-Guevara, R.; González, C.;Gelpí,J.L;Orozco, M. NucleicAcids Res. 2013, 41 (W1) W47– W55.

[4] Jammalamadaka SR, SenGupta A. Topics in Circular Statistics, Section 8. World Sci. Press: Singapore, 2001.

[5] Csárdi, G.; Nepusz, T. Inter. J. Comp.Syst.2006, 1695, 1-9.

[6] Gelpi, J.L.;Kalko,S.G.;Barril, X.; Cirera, J.; de la Cruz,X.;Luque, F.J.; Orozco, M. ProteinsStruct. Funct. Genet.2001, 45, 428-437.

[7] Cuervo, A.; Dans, P.D.; Carrascosa, J.L.; Orozco, M.; Gomila, G.; Fumagalli, L. Proc. Natl. Acad. Sci. U.S.A. 2014, 111, E3624– E3630.

[8] Lavery, R.;Maddocks, J.H.;Pasi, M.; Zakrzewska, K. Nucleic Acids Res. 2014, 42, 8138-8149.

[9] Pasi, M.; Maddocks, J.H.; Lavery, R. Nucleic Acids Res. 2015; 43, 2412-23.

[10] Schlitter, J. Chem. Phys. Lett. 1993, 215, 617.

[11] Andricioaei, I. and Karplus, M. J Chem Phys. 2001, 115, 6289–6292.

[12] Harris, S.A.; Gavathiotis, E.; Searle, M.S.; Orozco, M.; Laughton, C.A. J Am Chem Soc. 2001, 123, 12658-63.

[13] Cover, T. M.; Thomas, J. A. Elements of Information Theory; John Wiley & Sons: New York, 1991.

[14] Cukier, R. I. J. Phys. Chem. B 2015, 119, 3621–3634.

[15] Schreiber T. Measuring Information Transfer. Phys Rev Lett 2000, 85, 461–464.

[16] Roy, A., Perez, A., Dill, K., MacCallum, J. Structure 2013, 22, 168–175.

[17] Tyka, M.D., Clarke, A.R., and Sessions, R.B. J. Phys. Chem. B 2006, 110, 17212–17220.

[18] Cecchini, M., Krivov, S.V., Spichty, M., and Karplus, M. J. Phys. Chem. B 2009, 113, 9728–9740.