PARMBSC1: A REFINED FORCE-FIELD FOR DNA SIMULATIONS

Ivan Ivani[1,2], Pablo D. Dans[1,2], Agnes Noy[3], Alberto Pérez[4], Ignacio Faustino[1,2], Adam Hospital[1,2], Jürgen Walther[1,2], Pau Andrio[2,5], Ramon Goñi[2,5], Alexandra Balaceanu[1,2], Guillem Portella[1,2,6], Federica Battistini[1,2], Josep Lluis Gelpí[2,7], Carlos González[8], Michele Vendruscolo[6], Charles A. Laughton[9], Sarah A. Harris[3], David A. Case[10],and Modesto Orozco[1,2,7]

[1]Institute for Research in Biomedicine (IRB Barcelona), the Barcelona Institute of Science and Technology, Barcelona, Spain.

[2]Joint BSC-IRB Research Program in Computational Biology, Barcelona, Spain.

[3]School of Physics and Astronomy, University of Leeds, Leeds, UK.

[4]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, USA.

[5]Barcelona Supercomputing Center, Barcelona, Spain.

[6]Department of Chemistry, University of Cambridge, Cambridge, UK.

[7]Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona, Spain.

[8]Instituto de Química Física "Rocasolano", CSIC, Madrid, Spain.

[9]School of Pharmacy and Centre for Biomolecular Sciences, University of Nottingham, Nottingham, UK.

[10]Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, USA.

Corresponding author: Modesto Orozco, modesto.orozco@irbbarcelona.org

Editorial summary:

Parmbsc1, a force-field for DNA simulations, is presented. It has been broadly tested on nearly 100 DNA systems and overcomes simulation artifacts that affect previous force-fields.

We present parmbsc1, a force-field for DNA atomistic simulation, which has been parameterized from high-level quantum mechanical data and tested for nearly 100 systems (representing a total simulation time of ~140 μs) covering most of DNA structural space. Parmbsc1 provides high quality results in diverse systems. Parameters and trajectories are available at http://mmb.irbbarcelona.org/ParmBSC1/.

The Force-field, the energy functional used to describe the dependence between system conformation and energy, is the core of any classical simulation including molecular dynamics (MD). Its development is tightly connected to the extension of simulation time scales. As MD trajectories are extended to longer timescales, errors previously undetected in short simulations emerge, creating the need to improve the force-fields[1]. For example, AMBER (Assisted Model Building with Energy Refinement) parm94-99 was the most used force-field in DNA simulations until multi-nanosecond simulations revealed severe artifacts[2,3], thus fueling the development of parmbsc0[4], which, in turn, started to show deviations from experimental data in the μsec regime (for example an underestimation of the twist, deviations in sugar puckering, biases in ε and ζ torsions, excessive terminal fraying[2,5], and severe problems in representing certain non-canonical DNAs[1,6]). Various force-field modifications have been proposed to address these problems, such as the Olomouc (OL)-ones[5,6] designed to reproduce specific forms of DNA. While these and other tailor-made modifications are useful, there is an urgent need for a new general-purpose AMBER force-field for DNA simulations to complement recent advances in the CHARMM (Chemistry at HARvard Macromolecular Mechanics) family of force-fields (Online Methods). We designed theparmbsc1 force-field presented here to solve these needs, with the aim of creating a general-purpose force-field for DNA simulations. We demonstrate its performance by testing its ability to simulate a wide variety of DNA systems (**Supplementary Table 1**).

Parmbsc1 shows good ability to fit quantum mechanical (QM) data (QM data fitting section in **Supplementary Discussion**), improving on previous force-field results (Online

Methods, **Supplementary Table 2**). We first tested QM-derived parameters on the Drew-Dickerson dodecamer (DDD), a well-studied DNA structure[2,7], typically used as benchmark in force-field developments.Parmbsc1 trajectories sampled a stable B-type duplex that remained close to the experimental structures (**Fig. 1** and **Supplementary Table 2**), preserving hydrogen bonds and helical characteristics, even at the terminal base pairs, where fraying artifacts are common using other force-fields[2,8] (see Online Methods and **Supplementary Discussion**). The average sequence-dependent helical parameters (**Fig. 1** and **Supplementary Figs. 1** and **2**), and BI/BII conformational preferences (**Supplementary Table 2** and **Supplementary Fig. 3**) matched experimental values (for the comparisons with estimates obtained with other force-fields see Online Methods). Furthermore, parmbsc1 reproduced residual dipolar couplings (Q-factor = 0.3) and NOEs (Nuclear Overhauser Effect; only two violations), yielding success metrics similar to those obtained in the NMR (Nuclear Magnetic Resonance)-refined structures (**Supplementary Table 3**).

We next evaluated the ability of parmbsc1 to represent sequence-dependent structural features from simulations on 28 B-DNA duplexes (**Supplementary Table 4**). The agreement between simulation and experiment was excellent (Root Mean Square deviation (RMSd) per base pair of 0.1 or 0.2 Å). Almost no artifacts arising from terminal fraying were present, and the average helical parameters (twist and roll from simulations: 33.9 º and 2.5 º respectively), matched values from the analysis of the PDB (33.6 º and 2.9 º)[9]. Moreover, parmbsc1 was able to reproduce the unique properties of A-tract sequences[10] (**Supplementary Figs. 4–6**), and capture sequence-dependent structural variability (**Supplementary Fig. 7**). We also studied longer duplexes (up to 56 bp) to ensure that a possible accumulation of small errors given by the force-field did not compromise the description of the DNA, finding excellent results (**Supplementary Table 5**). The expected spontaneous curvature was clearly visible in both static and dynamical descriptors, demonstrating that parmbsc1 trajectories were able to capture complex polymeric effects (**Supplementary Table 5**).

We also explored the ability of parmbsc1 to represent unusual DNAs, such as a Holliday junction, a complex duplex-quadruplex structure which was fully preserved in μsec-long trajectories (**Supplementary Figs. 8** and **9**); or the Z-DNA, a *levo* duplex containing nucleotides in *syn*, for which parmbsc1 not only provided stable trajectories (**Fig. 2a**), but also reproduced the experimentally known salt dependence, confirming that the conformation is stable only at high (4 M) salt concentration[11]. For Hoogsteen-DNA (H-DNA), simulations with parmbsc1 showed a stable duplex for more than 150 ns (**Fig. 2b**), and severe distortions in longer simulation periods (**Supplementary Fig. 10**), as expected from its metastable nature[12]. We obtained equivalent results for another metastable structure, the parallel poly-d(AT) DNA (**Supplementary Fig. 11**)[13]. Parmbsc1 simulations not only reproduced the known structure of parallel d(T-A·T) and d(G-G·C) triplexes (**Figs. 2c,d**), but also showed correctly that the equivalent antiparallel structures are unstable in normal conditions (**Fig. 2e**)[14]. Finally, parmbsc1 was able to reproduce experimental structures of both parallel and antiparallel DNA quadruplexes with RMSd < 2 Å (**Figs. 2f,g**).

We explored also the ability of parmbsc1 to reproduce the complex conformation of hairpins and loops, exceptionally challenging structures for force-fields[15]. We performed μs simulations of the d(GCGAAGC) hairpin (PDB: 1PQT), the 4T-tetraloop in *Oxytricha nova* quadruplex $d(G_4T_4G_4)_2$ (OxyQ; PDB: 1JRN), and the junction loops in the human telomeric quadruplex (HTQ; PDB: 1KF1). Parmbsc1 provided excellent representations (RMSd around 1 Å) of the d(GCGAAGC) hairpin (**Fig. 2h**), and of the OxyQ quadruplex (**Fig. 2i**). For the very challenging HTQ structure, parmbsc1 maintained the stem structure 20 times longer than in previous simulations[15], and recognized the large flexibility of the loops in the absence of the lattice-contacts (**Supplementary Fig. 12**), showing that, as predicted[16], not only the crystal, but also other loop conformations were sampled (**Fig. 2j**).

As an additional critical test of the new force-field we predicted NMR observables from parmbsc1 trajectories (Online Methods). We obtained equivalent NOE violation statistics to those determined from NMR-derived ensembles (**Supplementary Tables 6** and **7**, and **Supplementary Fig. 13**). This agreement was maintained in *de novo* predictions, *i.e.* in those cases where NMR observables were collected in one of our laboratories after parmbsc1 development (**Supplementary Table 8**). Finally, it is worth noting that parmbsc1 trajectories reproduced the structure of DNA in crystal environments, yielding a RMSd between the simulated and crystal structures of only 0.7 Å, and average twist differences below one degree, improving on previous calculations (Online Methods and **Supplementary Figs. 14** and **15**).

In our final structural test we explored the ability of parmbsc1 to reproduce the conformation of DNA in complex with other molecules. We studied four diverse protein DNA complexes (PDB: 1TRO, 2DGC, 3JXC and 1KX5), and two prototypical drug DNA complexes. In all cases, we found excellent agreement (RMSd for DNA around 2–3 Å in protein-DNA complexes, and 1–2 Å in drug-DNA complexes) with experiments (**Fig. 3** and **Supplementary Figs. 16** and **17**).

A force-field should not only reproduce the structure of DNA, but also its mechanical properties[1]. To evaluate the performance of parmbsc1 we firstly evaluated the μs-scale dynamics of the central 10 base pairs of the DDD. The agreement between parmbsc0 and parmbsc1 normal modes and entropy estimates (Online Methods and **Supplementary Table 9**) demonstrated that parmbsc1 does not "freeze" the DNA structure, a risk for a force-field reproducing well average properties. This was further confirmed by the ability of parmbsc1 to reproduce the DNA dielectric constant (8.0 ± 0.3 for DDD *versus* the experimental estimate of 8.5 ± 1.4; see **Supplementary Fig. 18**), and also the cooperative binding (around 0.7 kcal mol^{-1}) of Hoechst 33258 to DNA. We then computed the helical stiffness matrices for the ten unique base pair steps[17,18]. Parmbsc1 values were intermediate between parmbsc0 and CHARMM27 stiffness parameters[18],

and substantially smaller than those suggested by Olson and coworkers[17] (**Supplementary Table 10** and **Supplementary Fig. 19**); the dependence of the stiffness parameters on sequence were similar for parmbsc1 and parmbsc0[17].

The persistence length, the torsional, and the stretching modules were obtained from simulations of long (up to 56 bp) duplexes (Online Methods). Parmbsc1 predicted persistence lengths in the range of 40–57 nm (**Supplementary Table 11**), close to the generally accepted value of 50 nm. The computed static persistence length, stretch and twist torsion modules were around 500 nm, 1,100–1,500 pN, and 50–100 nm respectively, also in agreement with experimental values (**Supplementary Table 11**). Finally, we explored the ability of parmbsc1 to describe relaxed and stressed DNA minicircles. We performed three 100 ns simulations of a 106-bp minicircle with ten turns (106t10), which should have zero superhelical density (σ =0) and therefore no denatured regions[19,20] (**Supplementary Fig. 20**). A kink was observed only in a single replica for one of the register angles, while in the remaining simulations the DNA remained intact (**Supplementary Fig. 20**). On the contrary, negatively supercoiled 100-bp (100t9; σ = –0.05) and 106-bp (106t9, σ = –0.10) minicircles formed distortions due to the superhelical stress, as previously reported experimentally using enzymes that digest single stranded DNA[19,20].

Having demonstrated the ability of parmbsc1 to describe stable and metastable DNA structures and DNA flexibility, we finally studied conformational transitions. Parmbsc1 reproduced the spontaneous A to B-form DNA transition in water, and the A form was found, as expected, to be stable in 200 ns control simulations in a 85% ethanol and 15% water mixture (**Supplementary Fig. 21**). Parmbsc1 also reproduced the unfolding of DNA $d(GGCGGC)_2$ in a 4 Molar pyridine solution (**Supplementary Fig. 21**), and the effective folding of d(GCGAAGC) in water (**Supplementary Fig. 22**), suggesting the ability to capture long-scale conformational changes in DNA.

Based on the wide series of tests we report, we conclude that parmbsc1 provides good representations of the static and dynamic properties of DNA. We anticipate that parmbsc1 will be a valuable reference force-field for atomistic DNA simulations under a diverse range of conditions.

METHODS

Methods and associated references are available in the online version of the paper.

ACKNOWLEDGEMENTS

MO thanks Spanish Ministry of Science (BIO2012-32868), the Catalan SGR, the Instituto Nacional de Bioinformática, and the European Research Council (ERC SimDNA) for support.MO is an ICREA academia researcher. MO thanks CPU-GPU time on MareNostrum-MinoTauro (BSC).CAL, SAH and AN thanks the UK HECBioSim Consortium for HPC time on ARCHER (Grant EP-L000253-1).AN was supported by the Biotechnology and Biological Sciences Research Council (BBSRC, grant number BB-I019294-1), and thanks ARC Leeds for computational resources. PDD is a PEDECIBA and SNI (ANII, Uruguay) researcher. DAC thanks C. Liu for assistance with the crystal simulation analysis.

AUTHOR CONTRIBUTION

II derived the parmbsc1 force-field parameter set. II, PDD, AN, AP, IF, AH, JW, AB, GP, FB, CAL, and SAH performed validation simulations. CG, MV, and GP validate results from NMR. CG did *de novo* NMR measures. DAC performed crystal MD simulations. RM, PA, AH, and JLG created the database infrastructure and web application. All authors contributed to the analysis of data. MO had the idea, directed the project, and wrote the manuscript which was improved by the rest of the authors.

FINANCIAL STATEMENT

The authors declare no competing financial interests.

REFERENCES

1. Pérez, A., Luque, F. J. & Orozco, M. *Acc. Chem. Res.***45,** 196–205 (2011).

2. Pérez, A., Luque, F. J. & Orozco, M. *J. Am. Chem. Soc.***129,** 14739–14745 (2007).

3. Varnai, P.& Zakrzewska, K. *Nucleic Acids Res.***32,** 4269-4280 (2004).

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

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

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

7. Wing, R. *et al. Nature***287,** 755–758 (1980).

8. Lavery, R. *et al. Nucleic Acids Res.***38,** 299–313 (2010).

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

10. Lankaš, F., Špačková, N., Moakher, M., Enkhbayar, P. & Šponer, J. *Nucleic Acids Res.***38,** 3414–3422 (2010).

11. Thamann, T.J., Lord, R.C., Wang, A.H.J. & Rich, A. *Nucleic Acids Res.***9,** 5443–5458 (1981).

12. Abrescia, N.G.A., González, C., Gouyette, C. & Subirana, J.A. *Biochemistry***43,** 4092–4100 (2004).

13. Cubero, E., Luque, F.J. & Orozco, M. *J. Am. Chem. Soc.***123,** 12018–12025 (2001).

14. Soyfer, V.N. & Potaman, V.N. in*Triple-helical nucleic acids* 1st edn. (Springer - Verlag New York, 1996).

15. Fadrná, E. *et al. J. Chem. Theory Comput.***5,** 2514–2530 (2009).

16. Martín-Pintado, N. *et al. J. Am. Chem. Soc.***135,** 5344–5347 (2013).

17. Olson, W.K., Gorin, A.A., Lu, X.-J., Hock, L.M. & Zhurkin, V.B. *Proc. Natl. Acad. Sci.***95,** 11163–11168 (1998).

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

19. Moroz, J.D. & Nelson, P. *Proc. Natl. Acad. Sci.***94,** 14418–14422 (1997).

20. Du, Q., Kotlyar, A. & Vologodskii, A. *Nucleic Acids Res.***36,** 1120–1128 (2008).

FIGURE LEGENDS

Figure1|Analysis of the Drew-Dickerson dodecamer. (**a**) Visual comparison of MD average structure (brown) and NMR structure (PDB id: 1NAJ) (light blue) and X-ray structure (PDB id: 1BNA) (green). (**b**) RMSd of 1.2 μs trajectory of DDD compared with B-DNA (blue) and A-DNA (green) form (coming from the standard geometries derived from fiber diffraction, see Online Methods section Validation of MD simulations). (**c**) RMSd compared to experimental structures (with (dark) and without (light) ending base-pairs): X-ray (green) and NMR (blue). Linear fits of all RMSd curves are plotted on top. (**d**) Evolution of total number of hydrogen bonds formed between base pairs in the whole duplex. (**e**) Helical rotational parameters (twist, roll, and tilt) comparison of average values per base-pair step (standard deviations are shown by error bars) coming from NMR (cyan), X-ray (dark green), 1 μs parmbsc0 trajectory[2] (black) and 1.2 μs parmbsc1 trajectory (violet).

Figure2|Analysis of non-canonical DNA structures. (**a**) Comparison of Z-DNA (PDB id: 1I0T) simulations in neutralized conditions (green) and in 4 M solution of Na^+Cl^- (blue). Structural comparisons at given time points are shown above the RMSd curves. (**b**) Simulation of anti-parallel H-DNA (PDB id: 2AF1) showing deviation of the structure over time (highlighted in red). RMSd of (**c**) parallel d(T-A•T)$_{10}$, (**d**) parallel d(G-G•C)$_{10}$,and (**e**) antiparallel d(G-G•C)$_{10}$ triplexes. (**f**) Parallel (PDB id: 352D) and (**g**) anti-parallel (PDB id: 156D) quadruplex showing stable structures over time. (**h**) Structural stability of d(GCGAAGC) hairpin (PDB id: 1PQT) and (**i**) OxyQ quadruplex (PDB id: 1JRN) with ions, over time. (**j**) Human Telomeric Quadruplex (PDB id: 1KF1) with highlighted loops. RMSd of HTQ backbone, loop 1, loop 2 and loop 3 regions are shown below. In all panels, parmbsc1 (final, averaged or at a given trajectory point) structures (light blue; also

green for Z-DNA) are overlapped over experimental structure (grey) for comparison. See **Supplementary Table 1** for information on the PDB structures.

Figure3|Analysis of DNA-protein complexes. Structural details of microsecond trajectories of four complexes with PDB id: 1TRO (**a**), 2DGC (**b**), 3JXC (**c**) and 1KX5 (**d**) (500 ns trajectory). Each plot shows overlap of the MD starting (red) and final (blue) structures, time dependent mass-weighted root mean square deviation (RMSD in Å) of all DNA (red) and protein (cyan) heavy atoms, and comparison of the values of rotational helical parameter roll (in degrees) at each base pair step calculated from the X-ray crystal structure (cyan) and averaged along the MD simulation (red line with the standard deviation envelope in light red). For clarity, in the 1KX5 plot of the roll value, the base pair steps are defined by the number of the position along the DNA strand and not by the base pair step name.

ONLINE METHODS

General parameterization strategy.

AMBER charges and van der Waals parameters for DNA are able to reproduce high-level QM data[21–23] and hydration free energies[24–26], as well as producing reasonable hydrogen bond stabilities[2, 21–23, 27] and complex properties such as sequence-dependent stabilities of duplex DNA[2, 28, 29]. Thus, we decided to keep the non-bonded parameters unaltered in this force-field revision, and focus our efforts in the parameterization of the backbone degrees of freedom: sugar puckering, glycosidic torsion, and ε and ζ rotations (taking the recently re-parameterized α and γ torsions from parmbsc0[4]). Parameterization of the different torsion angles (see below) was done from high-level QM calculations using the refined gas phase fitted parameters as initial guesses for the refinement of parameters

in solution taken now as reference high level Self-Consistent Reaction Field (SCRF)-QM data. In cases where fitting of one force-field parameter requires the knowledge of another parameter for the optimization, an iterative procedure using parmbsc0 parameters in the first iteration was employed.

Quantum mechanical calculations.

Model compounds, shown in **Supplementary Fig. 23**, were first geometrically optimized at the B3LYP/6-31++G(d,p) level[30] from which single-point energies were calculated at the MP2/aug-cc-pVDZ level[31]. To reduce errors in the fitting, optimizations were done while selected backbone and sugar dihedral angles were constrained to typical values obtained from a survey of DNA crystal structures[9]. We obtained both vacuum and solvent profiles for all structures calculated. 3D profiles of ε and ζ were sampled with 10 º increment in the region of interest (ε = [175 º, 275 º], ζ = [220 º, 330 º]), and with 40 º increment in the rest of the profile. Profiles of χ were sampled with 15 º increment and profiles of sugar pucker by 10 º in the range of phase angles from 0 º to 180 º, and considering the four nucleosides. To increase the accuracy of the profiles, we performed CCSD(T)/complete basis set (CBS) calculations[32, 33] on key point along the Potential Energy Surface (for ε and ζ these points were B_I, B_{TRANS} and B_{II} states; for χ minima of *anti* and *syn* regions, and maximum between them; and minima of *North*, *East* and *South* conformations for the sugar pucker). These calculations were performed first by optimization at the MP2/aug-cc-pVDZ level, followed by single-point calculations at the MP2/aug-cc-pVXZ (X = Triplex and Quadruplex) levels. CBS energies were obtained by extrapolating to infinite basis set, from the scheme of Halkier *et al.*[32], and adding the correction term of the difference from CCSD(T) and MP2 with the 6-31+G(d) basis set. These high level points were introduced with increased weights in the global fitting (see below). All QM calculations were performed with Gaussian09 (http://www.gaussian.com).

Solvation corrections in QM calculations.

The solvent calculations were done at the single-point level using our version of the polarizable continuum model (PCM) from Miertus, Scrocco and Tomasi (MST)[34–40]. For comparison, test calculations were performed using Cramer and Truhlar SMD (Solvent Model based on Density) model[41], and the standard Integral Equation Formalism (IEF)-PCM[36] as implemented in the Gaussian09 package, obtaining very similar results (data not shown). Consequently, only MST values were used in this work.

Molecular mechanics and Potential of Mean Force calculations.

Molecular mechanics (MM) reference calculations of the QM-optimized structures *in vacuo* were obtained from MM single-point energy calculations using the AMBER 11 package (http://ww.ambermd.org). MM profiles in solution were recovered from potential of mean force (PMF) calculations created with umbrella sampling (US)[42] procedures in explicit solvent conditions (no restraints were used on any dihedrals out of the reaction coordinate in these calculations). US calculations were carried out with a weak biasing harmonic potential of 0.018 kcal mol^{-1} deg^{-2}. The resulting populations were integrated using the Weighted Histogram Analysis Method (WHAM, http://membrane.urmc.rochester.edu/content/wham).US calculations typically involve 40–100 windows, each consisting of 2–5 ns of equilibration and sampling times in the order of 1–2 ns. Simulation details in PMF-US calculations were the same as those outlined below in the validation of MD simulations section.

Force-field fitting.

The procedure of force-field fitting was similar to parmbsc0 parameterization process[4]. In order to avoid altering other torsional parameters of the general force-field, we introduced new atom types depending on the parameterization. For ε, ζ, and sugar pucker parameterization we assigned the atom type *CE* to C3' atom. For χ parameterization we assigned *C1* to the C8 atom of adenine and *C2* to the C6 atom of thymine, while keeping unchanged the atom types *CK* for guanine and *CM* for cytosine. Charges for model systems used in the parameterization were calculated from standard

RESP methods mimicking the original amber parameterization. We used the standard torsions definition, *i.e*. ε = C4'-C3'-O3'-P, ζ = C3'-O3'-P-O5', χ = O4'-C1'-N9-C8 (for dA and dG) and χ = O4'-C1'-N1-C6 (for dC and dT). For sugar pucker parameterization we chose ν_1=O4'-C1'-C2'-C3', the δ backbone and the ν_2=C1'-C2'-C3'-C4' dihedrals, since they connect the two corrections: ε/ζ and χ[43–45].

As in the parmbsc0 parameterization, we used a Monte Carlo method for fitting residual energy, or QM-MM difference (Eq. I), to a Fourier series limited to the third order to maintain the AMBER force-field philosophy (Eq. II). The rotational barrier V_n and the phase angle α of each periodicity (n = 1, 2, 3) were fitted to obtain the minimal error in:

$$E_{\mathrm{dih},x} = E_{QM} - E_{ffbsc0(x=0)} \qquad \text{(I)}$$

where *x* stands for a specific torsion or a combination of torsions (in the case of ε and ζ) and *ffbsc0(x=0)* refers to the standard parameters and the specific *x* torsion set to zero (that used in reference MM or US calculations noted above). The dihedral term is defined as:

$$E_{\mathrm{dih}} = \sum_{torsions} \sum_{n}^{3} \frac{V_n}{2} \left[1 + \cos(n\varphi - \alpha)\right] \quad \text{(II)}$$

where *torsions* denotes a torsion, *n* stands for the periodicity of the torsion, V_n is the rotational barrier, φ is the torsion angle, and α is the phase angle.

Our flexible Metropolis Monte Carlo algorithm allows the introduction of different weights in the fitting for each point of the profile, as well as weighting of energy slopes to guarantee smooth transitions, or even mixing information from different profiles obtained in different conditions or with different levels of QM data. Fittings were done taking all the data in consideration, but with increased weighting at the profile minima (typically five times more than others) specially at the key points computed through the

most accurate CCSD(T)/CBS approach (typically weighted nine times more than others). For certain cases like the sugar puckering, detailed attention was needed to properly reproduce the transition region, which was achieved by increasing the importance of the energy maximum and by also introducing weights to the slopes in the calculations. As described before[4], around 5–10 acceptable solutions of the Monte Carlo refinement were tested on short MD simulations (around 50–100 ns) for one small duplex $d(CGATCG)_2$ rejecting those leading to distorted structures. The optimum parameter set (see **Supplementary Discussion** and **Supplementary Table 12**), without additional refinement was extensively tested against experimental data. Note that the way in which the parameters were derived does not guarantee their validity for RNA simulations, for which the use of others already validated RNA force-fields are recommended[45].

Validation of MD simulations.

We performed MD simulations with the PMEMD code from the programs AMBER 11-12 (http://www.ambermd.org), or with GROMACS[46], depending on the given simulation. As shown in **Supplementary Fig. 24**, results are insensitive to the simulation engine or to the use of CPU or GPU-adapted codes[47]. Unless otherwise noted NPT conditions with default temperature and pressure setting, at 300 K and pressure of 1 atm, where used. Calculations employed an integration step of 2 fs in conjunction with SHAKE[48] (or LINCS[49] in the case of GROMACS), to constrain X-H bonds with the default values. The TIP3P[50] or SPCE[51] water models were used, with a minimum buffer of 10 Å solvation layer beyond the solute, and the negatively charged DNA was neutralized with Na^+or K^+ ions[52]. Test simulations with added salt (Na^+Cl^-) showed that DNA helical conformations were not much dependent on the surrounding ionic strength in the 0 to 0.5 M range (**Supplementary Discussion** and **Supplementary Fig. 25**). Long range electrostatic interactions were calculated using the particle mesh Ewald method (PME)[53] with default grid settings and tolerance. All structures were first optimized, thermalized and pre-

equilibrated for 1 ns using our standard protocol[8] and were subsequently equilibrated for an additional 10 ns period. Conformational snapshots were saved every 1, 10, 20, or even 100 ps depending on the system size, the objective of the simulation, and its length. Simulations mimicking crystal environments were carried out as described elsewhere[54] for d(CGATCGATCG)$_2$ (PDB: 1D23) using 2 μsec simulation with 12 unit cells (or 32 duplexes) in the simulation periodic box (**Supplementary Fig. 14**), for a total of 64 μsec of duplex simulation.

For annotation of conformational regions at the nucleotide level we used standard criteria. Sugar puckering (C3'-endo for P between 0 º and 36 º (canonical North) C4'-exo for P between 36 º and 72 º, O4'-endo for P between 72 º and 108 º (canonical East), C1'-exo for P between 108 º and 144 º, C2'-endo for P between 144 º and 180 º (canonical South), C3'-exo for P between 180 º and 216 º, C4'-endo for P between 216 º and 252 º, O4'-exo for P between 252 º and 288 º (canonical West), C1'-endo for P between 288 º and 324 º, and C2'-exo for P between 324 º and 360 º), glycosidic torsion (*anti* for 90º to 180 º or -60 º to -180 º, and *syn* for -60 º to 90 º). BI (ε trans, ζ gauche-) and BII (ε gauche-, ζ trans). An H-bond is annotated using standard GROMACS rules and was considered broken when donor-acceptor distance was greater than 3.5 Å for at least ten consecutive picoseconds. Reference A-DNA and B-DNA fiber conformations were taken from Arnott's values[55]. Whenever possible, the simulations were validated against experimental data obtained in solution.

A variety of analyses were performed to characterize the mechanical properties of DNA based on MD simulations. Flexibility analysis was performed using essential dynamics algorithms[56–58], base step stiffness analysis[17, 59, 60], and quasi-harmonic entropies computed by using either Andricioaei-Karplus[61] or Schlitter[62] procedures. Similarities between essential deformation movements were determined using standard Hess's metrics[63] as well as energy-corrected Hess-metrics[59]. The calculation of polymer deformation parameters (persistence length, stretch and twist torsion modules) was

done following different approaches to reduce errors associated to the use of a single method to move from atomistic simulations to macroscopic descriptors: i) extrapolation of base step translations and rotations[17, 59], ii) analysis of the correlations in the conformations and fluctuations of the DNA at different lengths[64], and iii) an implementation of Olson's hybrid approach, which requires additional Monte Carlo simulations using MD-derived stiffness matrices[65]. Dielectric constants of DNA were computed using Pettit's procedure[66, 67].

The trajectories were analyzed using AMBERTOOLS (http://www.ambermd.org), GROMACS[46], MDWeb[68], NAFlex[69], and Curves+[70], as well as with in-house scripts (http://mmb.irbbarcelona.org/www/tools).

NMR analysis.

Analysis of the ability of MD trajectories to reproduce NMR observables (NOE-derived interatomic distances and residual dipolar couplings) was done using the last 950 ns of microsecond trajectories. We used the Single Value Decomposition (SVD) method implemented in the program PALES[71] to obtain the orientation tensor that best fitted the calculated and observed RDC values. Violations of the NOE data were computed using the tool *g_disre*, included in the GROMACS package, using distance restraints derived from the deposited BioMagResBank database[72], or as described below when NOEs were collected *de novo* using full relaxation matrix experiments.

The *novo* NMR experiments.

Samples (3 mM oligonucleotide concentration) were suspended in 500 µL of either D2O or H2O/D2O 9:1 in 25 mM sodium phosphate buffer, 125 mM Na^+Cl^-, pH 7. NMR spectra were acquired in Bruker spectrometers operating at 800 MHz, and processed with Topspin software. DQF-COSY (Double Quantum Filter – Correlation spectroscopy), TOCSY (Total Correlation spectroscopy), and NOESY (Nuclear Overhauser effect

spectroscopy) experiments were recorded in D2O and H2O/D2O 9:1. The NOESY spectra were acquired with mixing times of 75, 100, 200, and 300 ms, and the TOCSY spectra were recorded with standard MLEV 17 spin lock sequence, and 80 ms mixing time. NOESY spectra were recorded at 5 and 25 ºC.

The spectral analysis program Sparky (https://www.cgl.ucsf.edu/home/sparky) was used for semi-automatic assignment of the NOESY cross-peaks and quantitative evaluation of the NOE intensities**.** Quantitative distance constraints were obtained from NOE intensities by using a complete relaxation matrix analysis with the program MARDIGRAS[73]. Error bounds in the inter-protonic distances were estimated by carrying out several MARDIGRAS calculations with different initial models, mixing times and correlation times (2.0, 4.0 and 6.0 ns). Final constraints were obtained by averaging the upper and lower distance bounds in all the MARDIGRAS runs.

Availability of force-field parameters and porting to different MD codes.

The refined parameters are incorporated in amber-format libraries accessible from http://mmb.irbbarcelona.org/ParmBSC1/. Porting to GROMACS format was done from amber topology files using external utilities (amb2gmx[74] and acpype[75] tools accessible at https://simtk.org/home/mmtools and https://github.com/choderalab/mmtools). Porting to NAMD (http://www.ks.uiuc.edu/Research/namd) is not required since direct reading of AMBER topology files is possible.

Data Management.

Trajectories and the analysis performed were placed in a novel dual database framework for nucleic acid simulations using Apache's Cassandra to manage trajectory data, and MongoDB to manage trajectory metadata and analysis. Results are available at http://mmb.irbbarcelona.org/ParmBSC1/. Details on the Barcelona's nucleic acids database will be presented elsewhere.

Online Methods references

21. Šponer, J., Jurecka, P. & Hobza, P. *J. Am. Chem. Soc.***126,** 10142–10151 (2004).

22. Hobza, P., Kabeláč, M., Šponer, J., Mejzlík, P. & Vondrášek, J. *J. Comput. Chem.***18,** 1136–1150 (1997).

23. Šponer, J. *et al.Chem. Eur. J.***12,** 2854–2865 (2006).

24. Orozco, M. & Luque, F.J. *Chem. Phys.***182,** 237–248 (1994).

25. Colominas, C., Luque, F.J. & Orozco, M. *J. Am. Chem. Soc.***118,** 6811–6821 (1996).

26. Orozco, M., Cubero, E., Hernández, B., López, J.M. & Luque, F.J. in *Computational Chemistry. Reviews of Current Trends***4,** 191–225 (World Scientific Publishing, 1999).

27. Pérez, A. *et al.Chem. Eur. J.***11,** 5062–5066 (2005).

28. Beveridge, D.L. *et al.Biophys. J.***87,** 3799–3813 (2004).

29. Portella, G., Germann, M.W., Hud, N.V. & Orozco, M. *J. Am. Chem. Soc.***136,** 3075–3086 (2014).

30. Krishnan, R., Binkley, J.S., Seeger, R. & Pople, J.A. *J. Chem. Phys.***72,** 650–654 (1980).

31. Woon, D.E. & Dunning Jr, T.H. *J. Chem. Phys.***98,** 1358–1371 (1993).

32. Halkier, A. *et al.Chem. Phys. Lett.***286,** 243–252 (1998).

33. Halkier, A., Helgaker, T., Jørgensen, P., Klopper, W. & Olsen, J. *Chem. Phys. Lett.***302,** 437–446 (1999).

34. Miertuš, S., Scrocco, E. & Tomasi, J. *Chem. Phys.***55,** 117–129 (1981).

35. Miertus, S. & Tomasi, J. *Chem. Phys.***65,** 239–245 (1982).

36. Cances, E., Mennucci, B. & Tomasi, J. *J. Chem. Phys.***107,** 3032–3041 (1997).

37. Bachs, M., Luque, F. J. & Orozco, M. *J. Comput. Chem.***15,** 446–454 (1994).

38. Soteras, I., Curutchet, C., Bidon-Chanal, A., Orozco, M. & Luque, F.J. *J. Mol. Struct. THEOCHEM***727,** 29–40 (2005).

39. Soteras, I., Forti, F., Orozco, M. & Luque, F.J. *J. Phys. Chem. B***113,** 9330–9334 (2009).

40. Soteras, I., Orozco, M. & Luque, F.J. *J. Comput. Aided Mol. Des.***24,** 281–291 (2010).

41. Marenich, A.V., Cramer, C.J. & Truhlar, D.G. *J. Phys. Chem. B***113,** 6378–6396 (2009).

42. Torrie, G.M. & Valleau, J.P. *J. Comput. Phys.***23,** 187–199 (1977).

43. Hart, K. *et al.J. Chem. Theory Comput.***8,** 348–362 (2011).

44. Wu, Z., Delaglio, F., Tjandra, N., Zhurkin, V.B. & Bax, A. *J. Biomol. NMR***26,** 297–315 (2003).

45. Zgarbová, M. *et al.J. Chem. Theory Comput.***7,** 2886–2902 (2011).

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

47. Galindo-Murillo, R., Roe, D.R. & Cheatham III, T.E. *Nat. Commun.***5,** (2014).

48. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H.J.C. *J. Comput. Phys.***23,** 327–341 (1977).

49. Hess, B., Bekker, H., Berendsen, H.J.C. &Fraaije, J.G.E.M. *J. Comp. Chem.***18**, 1463–1472(1997).

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

51. Berendsen, H.J.C., Grigera, J.R. &Straatsma, T.P. *J. Phys. Chem.***91**, 6269-6271 (1987).

52. Smith, D.E. & Dang, L. X. *J. Chem. Phys.***100,** 3757–3766 (1994).

53. Darden, T., York, D. & Pedersen, L. *J. Chem. Phys.***98,** 10089–10092 (1993).

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

55. Arnott, S. & Hukins, D.W.L. *Biochem. Biophys. Res. Comm.***47**, 1505-1509 (1972).

56. Orozco, M., Pérez, A., Noy, A. & Luque, F.J. *Chem. Soc. Rev.***32,** 350–364 (2003).

57. Pérez, A. *et al.J. Chem. Theory Comput.***1,** 790–800 (2005).

58. Amadei, A., Linssen, A. & Berendsen, H.J.C. *Proteins Struct. Funct. Bioinforma.***17,** 412–425 (1993).

59. Lankaš, F., Šponer, J., Hobza, P. & Langowski, J. *J. Mol. Biol.***299,** 695–709 (2000).

60. Noy, A., Perez, A., Lankas, F., Luque, F.J. & Orozco, M. *J. Mol. Biol.***343,** 627–638 (2004).

61. Andricioaei, I. & Karplus, M. *J. Chem. Phys.***115,** 6289–6292 (2001).

62. Schlitter, J. *Chem. Phys. Lett.***215,** 617–621 (1993).

63. Hess, B. *Phys. Rev. E***62,** 8438 (2000).

64. Noy, A. & Golestanian, R. *Phys. Rev. Lett.***109,** 228101 (2012).

65. Zheng, G., Czapla, L., Srinivasan, A.R. & Olson, W.K. *Phys. Chem. Chem. Phys.***12,** 1399–1406 (2010).

66. Cuervo, A. *et al. Proc. Natl. Acad. Sci*.**111**, E3624–E3630 (2014).

67. Yang, L., Weerasinghe, S., Smith, P.E. & Pettitt, P.M. *Bioph. J*.**69**, 1519–1527 (1995).

68. Hospital, A. *et al. Bioinformatics***28,** 1278–1279 (2012).

69. Hospital, A. *et al. Nucleic Acids Res.***41,**W47-W55(2013).

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

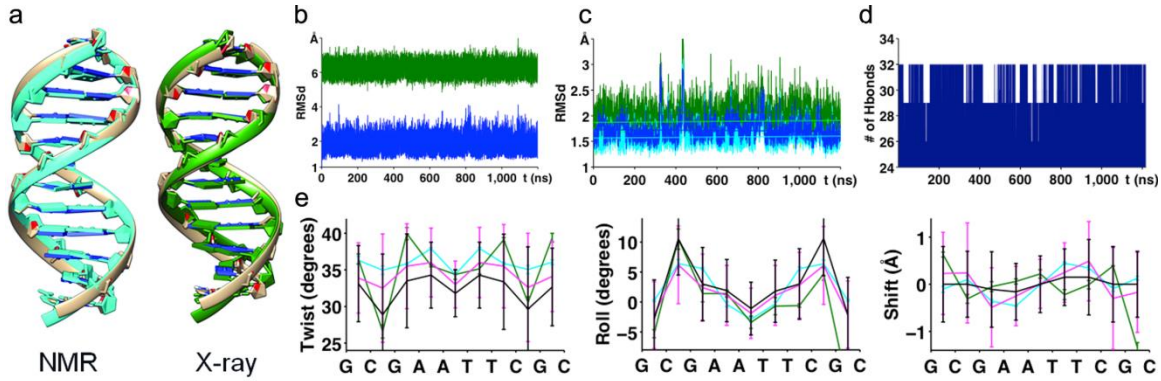
71. Zweckstetter, M. *Nat. Protoc.***3,** 679–690 (2008).

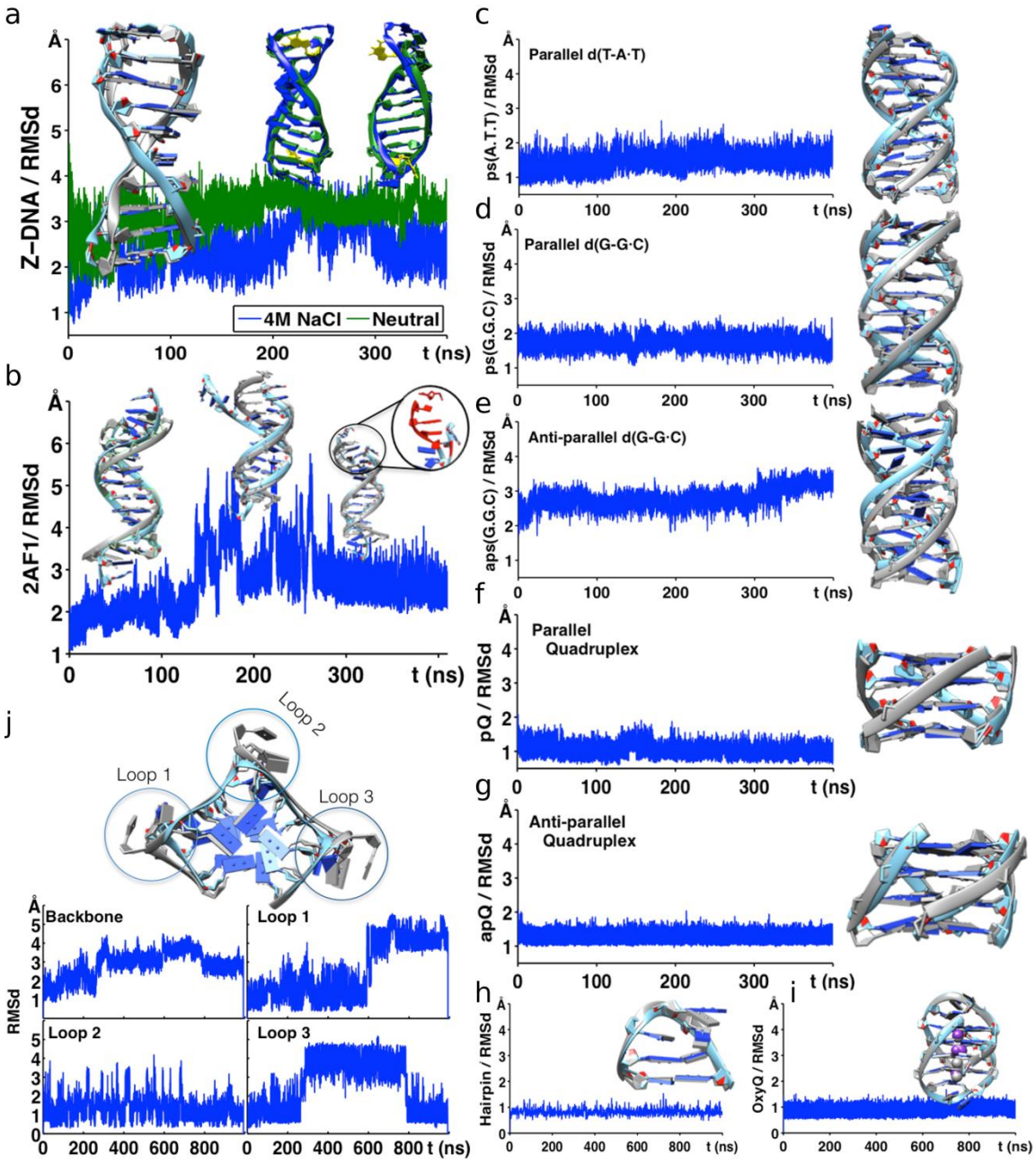
72. Bernstein, F.C. *et al. Eur. J. Biochem.***80,** 319–324 (1977).

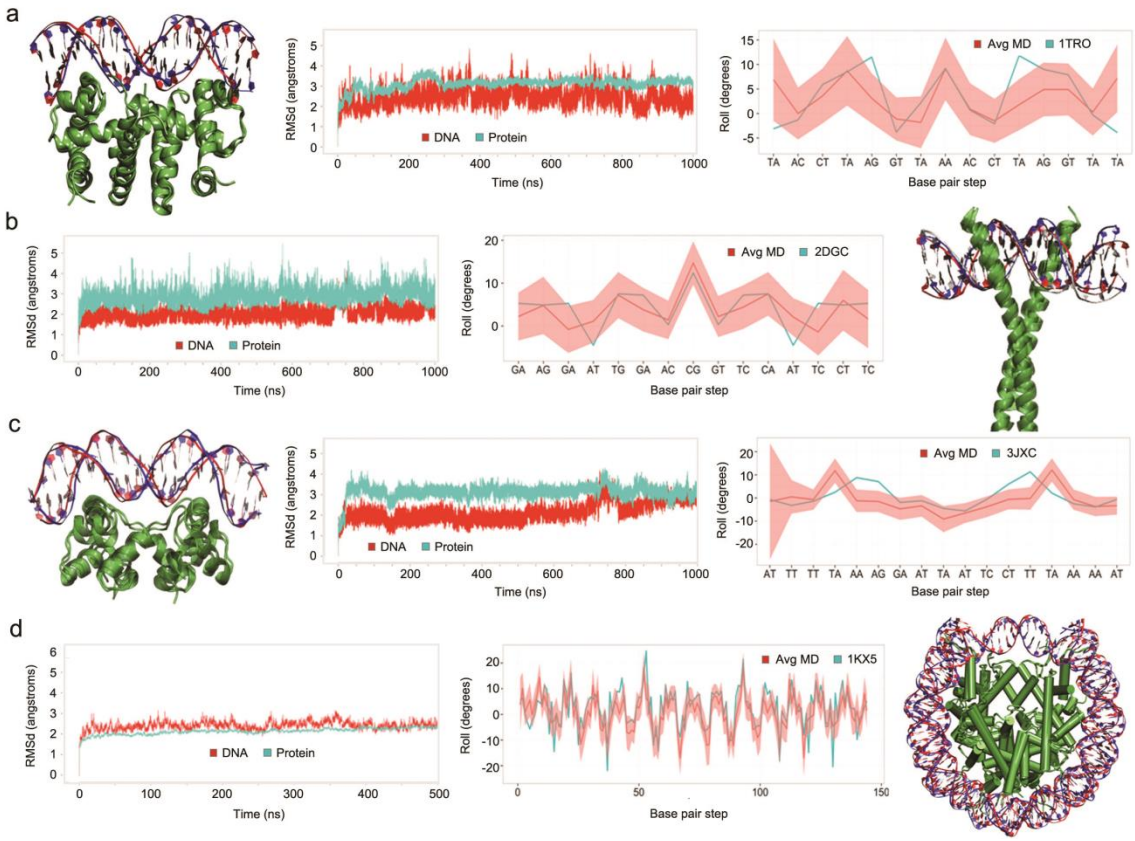
73. Borgias, B.A. & James, T.L. *Journal of Magnetic Resonance (1969)***87,** 475–487 (1990).

74. Mobley, D.L., Chodera, J.D, Dill, K.A., *J.Chem.Phys*.**125,** 084902 (2006).

75. Sousa da Silva, A.W. & Vranken, W.F., *BMC Res Notes***5**, 367 (2012).


a
NMR
X-ray
b
RMSd
Å
t (ns)
c
RMSd
Å
t (ns)
d
# of Hbonds
t (ns)
e
Twist (degrees)
Roll (degrees)
Shift (Å)
G C G A A T T C G C


a
Z–DNA / RMSd
Å
4M NaCl
Neutral
t (ns)
b
2AF1/ RMSd
c
Parallel d(T-A·T)
ps(A.T.T) / RMSd
d
Parallel d(G-G·C)
ps(G.G.C) / RMSd
e
Anti-parallel d(G-G·C)
aps(G.G.C) / RMSd
f
Parallel
Quadruplex
pQ / RMSd
g
Anti-parallel
Quadruplex
apQ / RMSd
h
Hairpin / RMSd
i
OxyQ / RMSd
j
Loop 1
Loop 2
Loop 3
Backbone
RMSd


a
RMSd (angstroms)
DNA
Protein
Time (ns)
Roll (degrees)
Avg MD
1TRO
TA AC CT TA AG GT TA AA AC CT TA AG GT TA TA
Base pair step
b
RMSd (angstroms)
DNA
Protein
Time (ns)
Roll (degrees)
Avg MD
2DGC
GA AG GA AT TG GA AC CG GT TC CA AT TC CT TC
Base pair step
c
RMSd (angstroms)
DNA
Protein
Time (ns)
Roll (degrees)
Avg MD
3JXC
AT TT TT TA AA AG GA AT TA AT TC CT TT TA AA AA AT
Base pair step
d
RMSd (angstroms)
DNA
Protein
Time (ns)
Roll (degrees)
Avg MD
1KX5
Base pair step