11090–11102 *Nucleic Acids Research, 2019, Vol. 47, No. 21* *Published online 18 October 2019*
doi: 10.1093/nar/gkz905

The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules

Pablo D. Dans[1,2,*], Alexandra Balaceanu[1,†], Marco Pasi[3,4,†], Alessandro S. Patelli[5,†], Daiva Petkevičiūtė[5,6,†], Jürgen Walther[1,†], Adam Hospital[1], Genís Bayarri[1], Richard Lavery[4], John H. Maddocks[5,*] and Modesto Orozco[1,7,*]

[1]Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology. Baldiri Reixac 10–12, 08028 Barcelona, Spain, [2]Department of Biological Sciences, University of the Republic (UdelaR), CENUR Gral. Rivera 1350, 50000 Salto, Uruguay, [3]LBPA, École normale supérieure Paris-Saclay, 61 Av. du Pdt Wilson, Cachan 94235, France, [4]Bases Moléculaires et Structurales des Systèmes Infectieux, Univ. Lyon I/CNRS UMR 5086, IBCP, 7 Passage du Vercors, Lyon 69367, France, [5]Institute of Mathematics, Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland, [6]Faculty of Mathematics and Natural Sciences, Kaunas University of Technology, Studentų g. 50, 51368 Kaunas, Lithuania and [7]Department of Biochemistry and Molecular Biology. University of Barcelona, 08028 Barcelona, Spain

Received May 18, 2019; Revised September 25, 2019; Editorial Decision September 27, 2019; Accepted October 06, 2019

ABSTRACT

We present a multi-laboratory effort to describe the structural and dynamical properties of duplex B-DNA under physiological conditions. By processing a large amount of atomistic molecular dynamics simulations, we determine the sequence-dependent structural properties of DNA as expressed in the equilibrium distribution of its stochastic dynamics. Our analysis includes a study of first and second moments of the equilibrium distribution, which can be accurately captured by a harmonic model, but with nonlocal sequence-dependence. We characterize the sequence-dependent choreography of backbone and base movements modulating the non-Gaussian or anharmonic effects manifested in the higher moments of the dynamics of the duplex when sampling the equilibrium distribution. Contrary to prior assumptions, such anharmonic deformations are not rare in DNA and can play a significant role in determining DNA conformation within complexes. Polymorphisms in helical geometries are particularly prevalent for certain tetranucleotide sequence contexts and are always coupled to a complex network of coordinated changes in the backbone. The analysis of our simulations, which contain instances of all tetranucleotide sequences, allow us to extend Calladine–Dickerson rules used for decades to interpret the average geometry of DNA, leading to a set of rules with quantitative predictive power that encompass nonlocal sequence-dependence and anharmonic fluctuations.

INTRODUCTION

DNA is a flexible and structurally polymorphic polymer whose overall equilibrium geometry strongly depends on its sequence, the solvent environment, and the presence of ligands (1,2). Conformational changes in DNA are mediated by a complex choreography of backbone rearrangements such as the BI/BII transition (3,4), the low-twist/high-twist equilibrium (5,6), or concerted α/γ rotations (7–9). Such backbone rearrangements lead to local and global changes in the helix geometry (9,10) impacting on the ability of the DNA to recognize ligands (11), and consequently on its functionality.

Binding-induced conformational changes in DNA are required for function and are expected to follow the sequence-dependent intrinsic deformation modes of DNA, *i.e.* are implicitly coded in the spontaneous deformability of isolated DNA. This suggests that evolution has refined DNA sequence not only to maximize ligand-DNA interactions but also to reduce the energetic cost of moving from a canonical to a bioactive conformation (11,12). This leads the notion of ‘indirect readout’, which suggests that the ability of the DNA to adopt the ‘bioactive’ conformation plays a major role in determining the target sequences of a given DNA ligand. Understanding the sequence-dependent physical prop-

[*]To whom correspondence should be addressed. Tel: +34 934039073; Email: pablo.dans@unorte.edu.uy
Correspondence may also be addressed to John H. Maddocks. Tel: +41 216932762; Email: john.maddocks@epfl.ch
Correspondence may also be addressed to Modesto Orozco. Tel: +34 934037155; Email: modesto.orozco@irbbarcelona.org
[†]The authors wish it to be known that, in their opinion, these co-authors should be regarded as Joint Second Authors.

erties of DNA then becomes crucial to rationalizing how ligands and, most notably, proteins, recognize and modulate DNA activity, i.e. the structural basis of gene regulation.

Understanding the sequence-dependent physical properties of DNA has been traditionally hampered by the lack of experimental data. Using simple steric considerations and geometric constraints Calladine developed in 1982 a set of principles to describe the mechanics of DNA (13), which have been used for decades to gain some qualitative insight into the sequence-dependence of expected, or average, local helical geometry. In their original version, those principles suggested that clashes between bases are avoided by a combination of a concerted change in twist, roll, and slide, as the basepair propeller increases to improve stacking (13). One year later, Dickerson formulated a simple numerical algorithm allowing for a quantification of Calladine's principles, coining the procedure as the 'Calladine's Rules' (14,15), which explains qualitatively the local variation (at the basepair level) in twist, roll, propeller twist and the torsion angle δ for a few B-DNA sequences which were previously determined experimentally (14). Unfortunately, the extension and predictive power of these rules, even in the most recent versions, is limited (1,16). Attempts to gain more quantitative information were based on the analysis of the variability in local helical parameters in structural databases (17,18), but to date (data accessed on the 19 March 2019), isolated B-DNA structures in the Nucleic acid Data Bank (NDB) allowed us to obtain flexibility data for only 5 of the 136 distinct tetranucleotides (only AATT, CGCG, CGAA, GCGA and ATTC are represented more than 500 times from 10 134 tetranucleotide analyzed belonging to 727 PDB structures). Even when the database is extended by including protein–DNA complexes (155 316 tetranucleotides belonging to 3434 PDB structures), the sampling is not dense enough to describe sequence-dependent DNA flexibility at the tetranucleotide level (24 out of the 136 tetranucleotides are still represented less than 500 times). In this context, atomistic molecular dynamics (MD) simulations are the only alternative to obtain robust and transferable parameters (10,19,20).

The first requirement for deriving physical descriptors of DNA from MD simulations is the availability of extended simulations for a library of sequence fragments containing all distinct tetranucleotides. This requires a significant computational effort which has encouraged joint projects such as the Ascona B-DNA Consortium (ABC, https://bisi.ibcp.fr/ABC), which have been instrumental, not only in describing physical properties of DNA but also in refining simulation protocols (10,21–23). The second major requirement is the availability of accurate force fields, such as the recently developed PARMBSC1 (24), which has been shown to represent DNA with a quality indistinguishable from experimental measurements (25). Thanks to the coordinated effort of several ABC groups, a series of microsecond-scale simulations on a library of DNA duplexes covering all of the 136 distinct tetranucleotides have been performed, and with a number of different simulation conditions, e.g. using PARMBSC0 (26) or PARMBSC1, different counter ions, etc. Consequently, there is a minimum of six total simulations of each independent tetranucleotide. The analysis of this large ensemble of data allows us to not only decipher the rules defining the sequence-dependent equilibrium geometry of B-DNA, but also those determining coordinated backbone conformational changes, and the correlations between various helical deformations. A new and comprehensive extension of Calladine–Dickerson rules emerges from the analysis of these simulations, including the first predictions of anharmonicity based on sequence context.

MATERIALS AND METHODS

The choice of sequences

The new ABC sequence library was designed to optimize the number of relatively short oligomers needed to include one copy of each of the distinct 136 tetranucleotides. Applying an adapted version of the Orenstein and Shamir algorithm (27–29), we generated 13 oligomers, each containing 18 bp (including GC terminals on each end), covering the complete tetranucleotide space (see Supplementary Table S1 for a list of the designed sequences), and 117 (of the 2080 possible) distinct hexanucleotide sequences. The smaller number of oligomers with respect to previous training libraries (6,10) made it more practical to obtain multi-microsecond trajectories under several simulation conditions (e.g. using both the PARMBSC1 (24) and PARMBSC0 (30) force fields, labelled miniABC$_{\mathrm{BSC1}}$ and miniABC$_{\mathrm{BSC0}}$ respectively), and by changing the ionic environment (from KCl to NaCl, labelled miniABC$_{\mathrm{BSC1}}$-K and miniABC$_{\mathrm{BSC1}}$-Na respectively). Comparison of results obtained with this library of sequences (miniABC) with respect to the standard ABC-set (μABC (10)) allowed us to check for the robustness of our conclusions as a function of the duplexes from which the tetranucleotide parameters were derived.

System preparation and MD simulations

All oligonucleotides were constructed with the *leap* program of AMBERTOOLS 15 (31) and simulated using the *pmemd.cuda* code (32) from AMBER14 (31), following the standard ABC protocol (10). Canonical duplexes were generated using Arnott B-DNA fiber parameters (33), and solvated by a truncated octahedral box of SPC/E (34) 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 (30) and PARMBSC1 (24) force fields were used to describe DNA, while Dang's parameters were used for ions (35). Systems were optimized and equilibrated as described elsewhere (10), and simulated for 1 μs in the NPT ensemble, using Particle-Mesh Ewald corrections (36) and periodic boundary conditions. SHAKE was used to constrain bonds involving hydrogen (37), allowing 2 fs integration step. Typically, analyses presented here correspond to the second part of the trajectory (last 500 ns). Trajectories are accessible at the BIGNASim server: https://mmb.irbbarcelona.org/BIGNASim/ (38).

Global analysis

Trajectories were processed with the *cpptraj* (39) module of the AMBERTOOLS 15 package (31), and the NAFlex server (40) for standard analysis. DNA helical parameters and backbone torsion angles were measured and analysed with the CURVES+ and CANAL programs (41), following the standard ABC conventions (10). Duplexes were named following the Watson strand. The letters R, Y and X stand for a purine, a pyrimidine, or any base respectively; base pairs flanking a dinucleotide step were denoted using two dots to represent the central step (e.g. R..Y), and one dot when trinucleotides are considered (e.g. R.Y), while X:X and XX represent a basepair and a basepair step, respectively. Bayesian Information Criterion (or BIC) (42,43) was used to quantify the normal or binormal (i.e. a mixture of two normal functions) nature of the distributions of the helical parameters (see Supplementary Methods). An extension of Helguerro's theorem (44,45) was used to distinguish those binormal distributions where the two Gaussians are very close (unimodal distributions) from those where they are significantly separated (bimodal distributions). The similarity between first and second moments (i.e. average and covariance) of the helical parameter distributions for different simulation libraries was evaluated using the Kullback-Leibler (KL) divergence. More specifically sequence-dependent Gaussian coarse grain cgDNA (46–48) model parameters were computed from each of the four MD training libraries used in this work (i.e. μABC$_{BSC0}$-K, miniABC$_{BSC0}$-K, miniABC$_{BSC1}$-K, miniABC$_{BSC1}$-Na) in order to be able to generate associated predictions of first and second moments of the helical parameters for fragments of arbitrary sequence. In particular this allowed us to compare PARMBSC0 simulations of the μABC library with the PARMBSC0 simulations of the miniABC library, even though the two libraries have different sequence fragments. See the Supporting Methods for specific details on BIC, Bayes Factors, Helguerro's theorem, KL divergence between configuration distributions, and DNA Persistence Length (PL) calculations.

Correlations between helical structural substates

For each tetranucleotide we calculated the correlation between the backbone state at the central step (inter-basepair *i*) and the helical parameters at two 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°; *trans* (t) = 180 ± 40° and gauche negative ($g-$) = 300 ± 40°. 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 did not 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 (49) correlation coefficient, mathematically equivalent to the Pearson correlation (50), was used as a measure of the correlation between these discrete substates of the backbone and the continuous values of the inter-basepair helical 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).

Statistics, graphics and molecular plots

The statistical analysis, including the Bayesian Information Criterion (BIC), Bayes Factor analysis, Helguerro's theorem, Kullback–Lieber divergence, PL and correlations measurements, as well as associated graphics, were obtained with R 3.0.1 statistical package (51), the MatLab R2016b package, numpy (52) and matplotlib (53). The molecular plots were generated using VMD 1.9 (54).

RESULTS AND DISCUSSION

Sources of uncertainty: the sequence library and the type of salt

Before going into detail with a conformational analysis, we first considered the robustness of our results to changes in the choice of sequence library, because large differences would challenge the general validity of our conclusions. Fortunately, only one of the 1,632 distributions analysed (namely of 6 intra- plus 6 inter-basepair helical parameters for each of the 136 distinct tetranucleotides), showed significant differences (according to BIC-Helguerro analysis) depending on the choice of library (the previous μABC library, or the current miniABC library; see Supplementary Figure S1). Furthermore, no differences were found depending on the salt (see Supplementary Table S2 and raw data in https://mmb.irbbarcelona.org/miniABC/), which suggests that our results are robust to the choice between K and Na for the counter-ion. To gain additional confidence in the robustness of our results, we used the explicit form of Kullback–Leibler divergence available for Gaussian (i.e. multivariate normal) distributions to quantify three pairwise differences in cgDNA model predictions (see Materials and Methods, and Supplementary Methods) of the mean and covariance for each of the 13 miniABC library sequences for the four different parameter sets extracted from the μABC$_{BSC0}$-K, miniABC$_{BSC0}$-K, miniABC$_{BSC1}$-K and miniABC$_{BSC1}$-Na simulations. As can be seen in Figure 1, no significant difference arises from the change in the sequence library, nor from the difference between K and Na counter ions. However, the results are quite sensitive to the change in force field from PARMBSC0 to PARMBSC1. This is to be expected since the latest PARMBSC1 force field leads to a considerably more realistic representation of twist/roll and BI/BII distributions (see the analysis and discussion published elsewhere (9,25)), and to straighter average configurations of duplexes than those obtained from prior force fields. This can be confirmed by considering the differences between

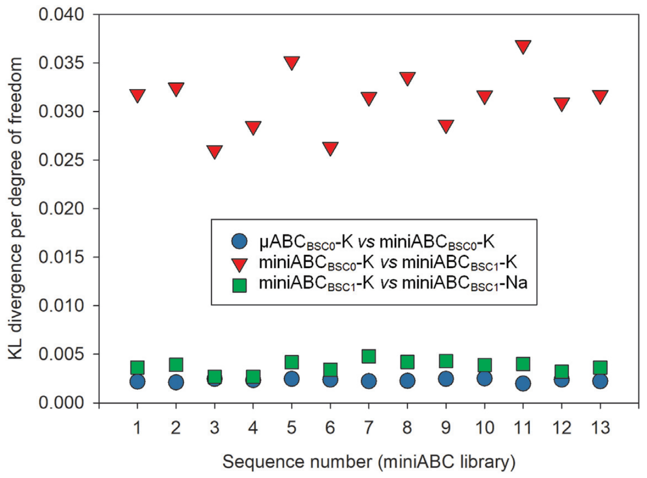

Figure 1. Symmetric Kullback–Leibler divergence per degree of freedom between Gaussian distributions, which is a combined measure of differences in values of first and second moments, for each of the thirteen oligomers in the miniABC training library, but for cgDNA model parameter sets fitted to different MD simulation protocols (see Materials and Methods and Supplementary Methods).

static and dynamic persistence lengths (as introduced elsewhere (55)) over a large ensemble of sequences (see Supplementary Figure S2).

Strong anharmonic distortions do arise

One of the most important extreme deformations of DNA is the disruption of base pairing, which can be analyzed in detail by aggregating data over all instances of G:C and A:T basepairs. This allowed us to accumulate ensembles of independent 3D conformations, which added together reach the millisecond time scale. Terminal basepairs (G:C pairs in all the cases) showed open states (water molecules in between H-bonding Watson–Crick groups) in 1–2% of the total simulation time, with short average open lifetimes (around 3 ns, see Supplementary Table S3) in agreement with time-resolved Stokes shifts spectroscopy (56), but most probably too short to lead to isotope exchange signals in NMR experiments (57). The opening of central basepairs is less likely to occur (between 0.01% in G:C and 0.05% in A:T of the simulation time), but when it happens, the open state can survive considerably longer (up to 50 ns). Whether or not this time is sufficient to allow proton interchange with the solvent is unclear. Another example of a strong anharmonic deformation arising in our simulations is the temporary formation of a sharp kink (Supplementary Figure S3) associated with anomalous rise and roll (58,59) at an AA basepair step within a TAAA tetranucleotide belonging to a relatively long continuous stretch of A:T basepairs (seq. 9, see Supplementary Table S1). Very interestingly, this deformation (not previously seen in simulations) has been characterized before as one of the origins of bubbling and kinking in natural DNA (60,61), and correlates with the extreme flexibility of NTAN (where N = A, C, G, T) sequences described by NMR experiments (62) and MD simulations (63).

Equilibrium distributions of basepair deformations are close to harmonic

A BIC analysis was carried out for the distributions of all six of the intra-basepair helical parameters at the central base pair in all 64 possible trinucleotide contexts. These distributions were all observed to be rather close to Gaussian, cf. Supplementary Figure S4, with the save for exceptionally rare events, as discussed in the last paragraph. Certainly no multi-peaked distribution was ever observed. Nevertheless the average value, or first moment, of each of the six intra-basepair parameters is strongly sequence-dependent to at least the trinucleotide sequence context, see Figure 2. Some qualitative rules on the sequence-dependent variation in the means can be observed. Shear values in G:C base pairs, when G is followed by Y are below average, while the opposite happens for A:T base pairs. Buckle in G:C shows large variations depending on the nature of the 3′-base of G, with an R leading to large positive buckles, and a Y leading to large negative buckles. Propeller also shows clear sequence effects, with A:T base pairs having a sizeable negative value when there is an R 5′ to the A, while propeller is close to zero for G:C basepairs within YGR trinucleotides.

Equilibrium distributions of basepair step deformations are frequently strongly anharmonic

Bi-normality (i.e. deviation from gaussianity) in the equilibrium distributions of the inter-basepair helical coordinates is common, but clear bimodality (i.e, the appearance of distinct multiple peaks) is observed in only 3% (miniABC$_{BSC1}$-K$^+$) to 5% (miniABC$_{BSC1}$-Na$^+$) of the inter-basepair helical parameter distributions (Figure 3 and Supplementary Figure S5). Bimodality appears systematically only for slide (several tetranucleotides containing a central GG step), shift (typically in a few tetranucleotides containing a YR central step) and twist (mainly in tetranucleotides containing central CG or AG steps). These conclusions are completely compatible with our prior analysis of PARMBSC0 simulations (see the μABC work (10), particularly Figure 8). There are few cases where bimodality affects simultaneously two or more helical parameters, for example, AGGA and GGGA are bimodal in shift and slide (in agreement with experimental data (64)) and ACGG, GCGA and GCGG are bimodal in shift and twist in agreement with results derived from the data mining of PDB structures (5). The central step of the GTAA tetranucleotide is the only case displaying bimodality in three helical parameters (shift, slide and twist) simultaneously. In general, shift bimodality is coupled with the appearance of high-shift values (>1 Å). The reverse situation was found for slide, where bimodality displaces the distribution to lower values. Finally, twist bimodality displays more complex behavior, as in some cases the second peak of the distribution occurs at lower than canonical values (<30°), while in others it is at high twist values (>40°). See Figure 3 and Supplementary Figures S6–S8 for a detailed analysis.

While inter-basepair, or junction, helical coordinates are frequently far from having a normal distribution, the first and second moments of their equilibrium distributions are still well defined, and can be approximated by evaluating

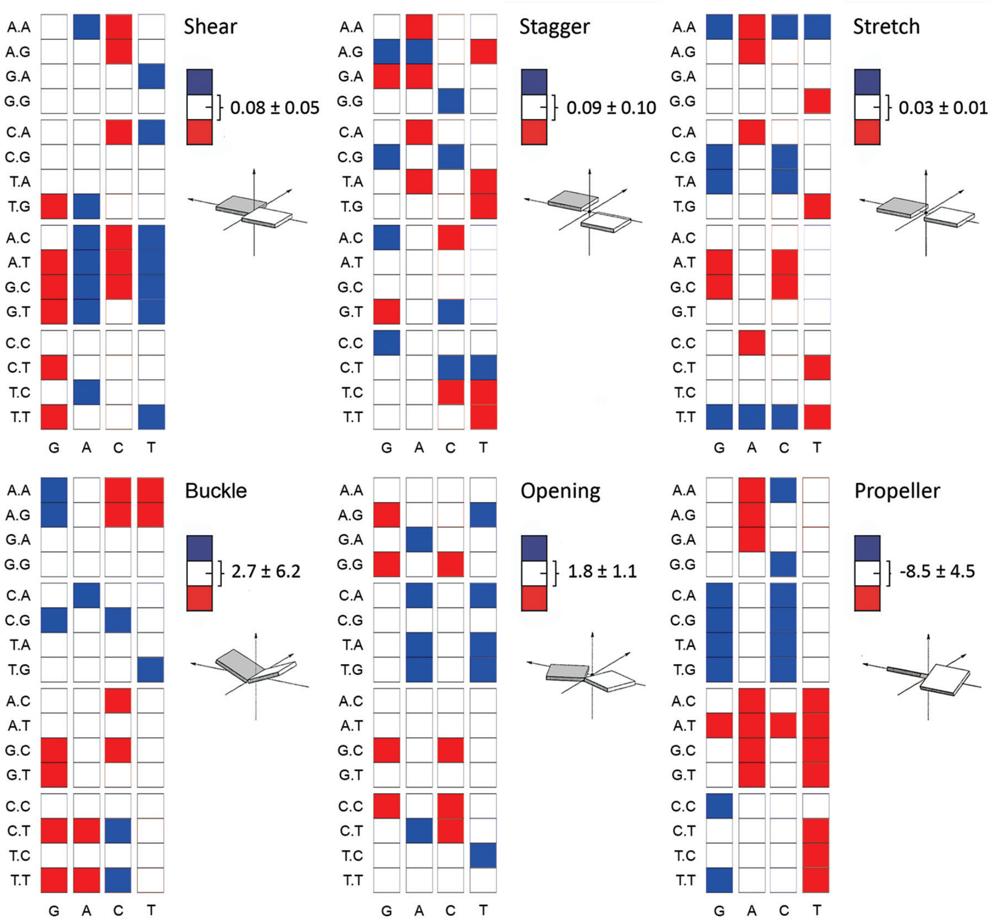

Figure 2. Average values of intra-basepair helical coordinates of the central base-pair (x-axis) in all possible 64 trinucleotide sequence contexts (y-axis). Results obtained from the miniABC$_{BSC1}$-K simulations. The global averages (white) are over all sequence contexts and standard deviations reflect the variation among trinucleotide contexts. The blue squares mean that a specific base-pair has an average value above the global average plus one standard deviation, while red squares mean an average value below the global average minus one standard deviation.

the appropriate averages along our MD simulation time series, and over all instances of dinucleotide (or NN, nearest neighbor) or tetranucleotide (NNN, next nearest neighbor) contexts. Only a few general NN rules can be observed for the first moments: (i) As suggested by Calladine roll (YR) > roll (RY) and twist (RY) > twist (YR); (ii) YR basepair steps typically have higher than normal slide and roll; (iii) RY steps typically have lower than normal slide and roll and (iv) YY and RR steps have lower than normal tilt values. Any further rules concerning the average values of helical inter-basepair coordinates need to be formulated as the averages for the central junction or step in a specific tetranucleotide sequence context due to strong nonlocal sequence dependence, at least in part due to tetranucleotide dependent anharmonic effects (Figure 3 and discussion below).

Backbone polymorphism

The flexibility of DNA backbone is linked to rotations around seven torsion angles (α, β ,γ, δ, ε, ζ and χ, with δ in the present analysis being replaced by the sugar phase angle P), which in some cases move in a concerted way (e.g. α/γ and ε/ζ) between well-defined conformational sub-states. The best-studied of the coupled transitions is the so-called BI/BII transition, which occurs due to the concerted rotation of the ε/ζ torsions. BI→BII transitions are believed to be functionally relevant. They occur in some high-resolution crystal structures (65,66) and are also detected in ^{31}P NMR spectra (67,68). Results in Supplementary Figure S9 show that the BII state is much more frequent than expected from simulations performed using previous force fields, matching NMR estimates for equivalent sequences (69). Very interestingly (see Figure 4, and Supplementary Tables S4 and S5), the BI/BII equilibrium is strongly dependent on the surrounding base sequence. For example, RR backbones (i.e. the ε and ζ torsions involving the phosphate between two purines) exhibit quite high BII percentages, especially in the presence of Y at the 5′ end of the corresponding tetranucleotide, while the YY backbones are typically biased towards the BI state, generating a strong asymmetry at RR:YY steps. While the general trends of BI/BII equilibria are robust with respect to changes in salt, a detailed analysis indicates the existence of subtle differences (5), which are especially visible for RR and YR basepair steps: in general, Na$^+$ increases the total percentage of the BII state (Figure 4), but reduces

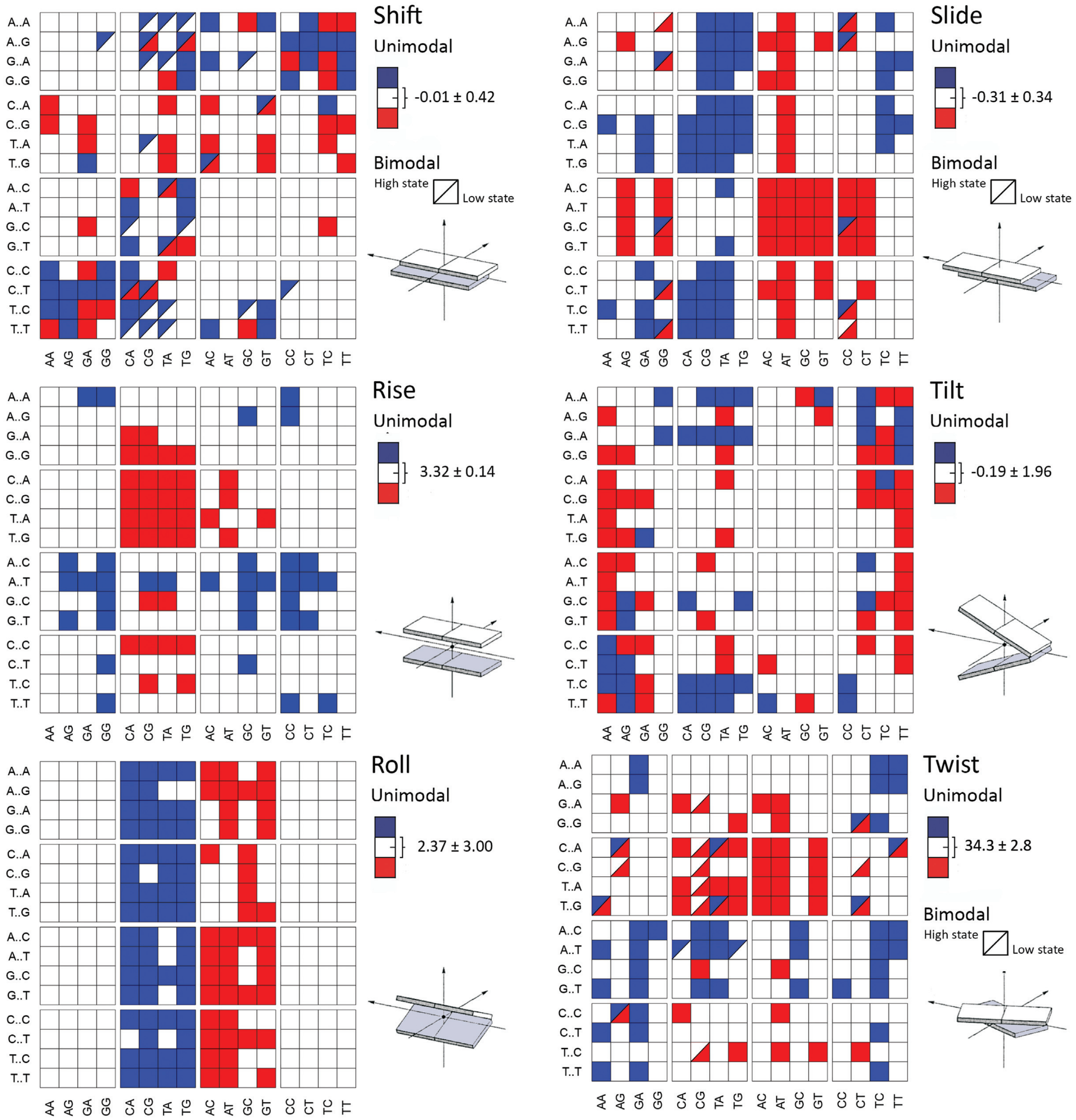

Figure 3. Average values of inter-basepair, or step, helical coordinates for the central junction (x-axis) in all possible 256 tetranucleotide contexts (y-axis). Results obtained from the miniABC$_{BSC1}$-K simulations. Tetranucleotides classified as bimodal (half-square) are polymorphic (i.e. they sample two clear conformational substates). The global averages (white), exhibited on the legend at the right of each squared-plot, were computed from the weighted-averages obtained through BIC (see Materials and Methods and Supplementary Methods), and while the standard deviations reflect the variation among the tetranucleotide sequences that share the same central basepair step. The blue squares mean that a specific step has an average value above the global average plus one standard deviation, while red squares mean an average value below the global average minus one standard deviation.

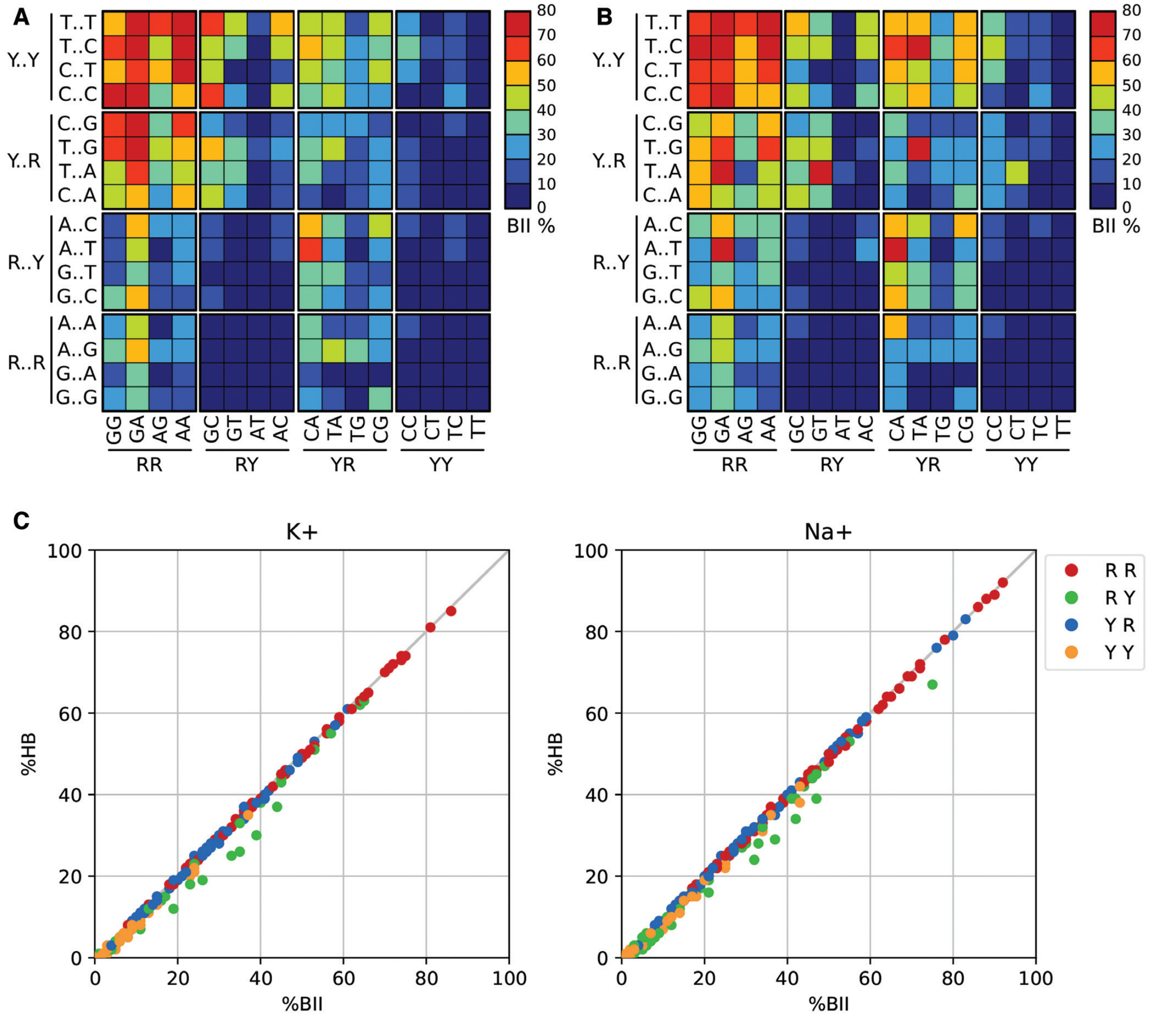

Figure 4. Sequence dependence of BII backbone conformations comparing K^+ and Na^+. (**A**) miniABC$_{BSC1}$-K BII percentages. (**B**) miniABC$_{BSC1}$-Na BII percentages. (**C**) Correlation between the percentage of BII (%BII, horizontal axis) and the occurrence of formation of the C–H···O H-bonds (%HB, vertical axis) at the central base step of each of the 256 possible tetranucleotide sequences, colour-coded according to the base type of the central base step.

its sequence-dependence, in perfect agreement with experimental data (70). As previously reported (4,5), we found a very strong correlation between BI→BII transitions and the formation of unconventional hydrogen bonds of the type CH—O, which are instrumental in mechanically coupling the movements detected in the backbone with those seen in the bases (see Figure 4, Supplementary Table S6).

In contrast to BI/BII dynamics, the α/γ conformational landscape is dominated by the canonical conformation, which, on average, represents ∼90% of the collected ensembles. Non-canonical conformers are more likely to appear in Na^+ simulations than with K^+ (Supplementary Tables S7 and S8). Transitions to non-canonical α/γ conformations are frequent, but the alternative states tend to have a short lifetime (on average we measured ∼500 transitions per μs per nucleotide, with an average residence time ∼5 ps). These brief transitions have little impact on the global conformational ensemble (9). No clear sequence-related rules can be determined for α/γ transitions, but, as expected, C and G nucleotides show longer-lived and more frequent α/γ transitions than A or T (8,9,71). Pseudo rotational Phase (P) angle analysis (Supplementary Figure S10) shows South (C2′-endo, ∼150°) conformations are dominant as expected, but East conformers are common, and sampling North states is not rare, especially for pyrimidines (9). As also expected, glycosidic torsions (χ) are always in the *anti* region (−180 to −90°), with purines sampling more frequently than pyrimidines the *high-anti* conformations (−90 to −30°; see Supplementary Figure S11). Finally, all nucleotides exhibit the same wide distribution for the β angle, spanning from 120° to 240°, with a strongly marked peak at the canonical value (180°) and a marginal population at ∼70° (g+, see Supplementary Figure S12), in good agreement with results from the data mining of X-ray structures (72).

The choreography of correlated motions in the DNA

The movements of the DNA duplex often involve concerted changes in conformational degrees of freedom, generating a complex choreography. As an example, puckering (measured by the pseudo rotational angle P) and glycosidic torsions (measured by the χ angle) are tightly coupled, and the population of East and North puckering leads to a marked displacement of χ to lower values (Supplementary Figure S13). Furthermore, χ and P torsions are coupled to the ε/ζ changes in a sequence-dependent manner (Supplementary Figure S14). Thus, in purines, the population of the BII state is coupled to a displacement of puckering to the East (P) and (χ) *high-anti* regions, while in pyrimidines the population of BII conformers leads only to a slight displacement to the *high-anti* region, without significant puckering changes.

When the conformational analysis is carried out at the basepair level, a pattern of sequence-dependent correlated movements emerges. All distinct trinucleotides show moderate-to-high correlations in shear-opening, shear-stretch, and stagger-buckle. The pattern of correlation is less clear for the remaining intra-basepair parameters, although several trinucleotides show stretch-opening correlations (Supplementary Figure S15). A more complex sequence-dependent picture of correlated movements can be obtained by analyzing the inter-basepair helical parameters (Supplementary Figure S16). For example, mild to strong correlations are found in shift-tilt, slide-twist, rise-tilt, shift-slide, and shift-twist movements for RR steps. For RY, weaker correlations can be found (depending on the tetranucleotide sequence-environment) in shift-tilt, slide-rise and roll-twist. Finally, YR basepair steps may exhibit moderate to strong correlations for shift-tilt, slide-twist, rise-twist and roll-twist (Supplementary Figure S16). Interestingly, for all the tetranucleotides, shift-slide and roll-twist always show negative correlations, while shift-tilt and slide-twist always show positive correlations. As expected, correlations also emerge when combining inter- and intra-helical parameters in the same analysis. Thus, a significant number of tetranucleotides show moderate to strong correlations of opening with shift, buckle with rise, and stagger with tilt (data not shown). It is also worth noting that the network of correlations extends to neighboring steps. As an example, twist in the central YR step of XYRR tetranucleotides is highly correlated with slide in the adjacent RR step (5,10), which again stresses the limitations of simple nearest neighbors interpretations of DNA conformational mechanics, and points the way to coarse grain models such as cgDNA (48), that encompass longer range coupling, with associated longer-range sequence-dependence of the observed means and many non-vanishing covariances.

Lastly, backbone and basepair conformations are connected in a complex way, with ε/ζ (BI/BII) being the major determinant of the polymorphism. Very often, tetranucleotides showing a simultaneous sampling of BI and BII conformations are those with bimodality in some helical parameter at the same step (70% of the bimodal inter-basepair helical parameters occur in steps with bimodal BI/BII distributions, see Figure 3 and Supplementary Tables S4 and S5). The BI/BII state also correlates with inter-basepair helical coordinates in neighboring junctions, explaining part of the geometrical constraints postulated by Calladine. For example, the increase in the percentage of BII at the central junction of a given tetranucleotide correlates with larger shift values at the same junction for all sequences (Supplementary Figures S17) and is also coupled to lower twist and slide values. The BI/BII ratio at a junction *i* also correlates with shift, twist and slide values at step *i* + 1 and *i* – 1 (Supplementary Figures S18 and S19), highlighting the subtle mechanical coupling between backbone and basepair step conformations within DNA (72).

All the observations made above can be unified in a global flexibility scheme for B-DNA (Figure 5), showing that all basepair junctions contain potentially polymorphic elements (BI/BII, shift, slide, or twist) that can lead to bimodal behavior depending on the specific tetranucleotide environment. The analysis we have carried out leads to a scheme with strong predictive power at the tetranucleotide level. As a single example, we can now say with confidence that when the choice of X and Y within an XYRY tetranucleotide leads to bimodality, this will be expressed in shift and twist, coupled with a low-to-moderate percentage of BII in the Watson strand. In contrast, when XRRX tetranucleotides are considered, bimodality will show up in either shift, slide or twist (or a combination thereof), coupled with a moderate-to-high percentage of BII in the Watson strand of the central junction.

Calladine–Dickerson rules and beyond

Calladine's principles were based on steric hindrance and stacking, which constrained a few helical parameters to some specific conformations (13). Dickerson transformed Callandine's principles in four simple sum functions (Σ_1 for twist), by which the expected local variation in twist, roll, torsion angle δ, and propeller twist could be explained from the DNA sequence (14). The Calladinec–Dickerson rules reproduced DDD (Drew–Dickerson Dodecamer), although the predictive power was limited since the value corresponding to one unit of Σ, and the center of the sum functions (i.e. 2.1° and 35.6° for $\Sigma_{1/DDD}$ respectively) could only be set *a posteriori* of knowing the 3D structure (14).

Modern force fields for DNA were fine-tuned over decades (20,24,30) and are now able to reproduce accurately helical conformations and backbone substates (25). Combined with a suitable library of sequences, the first moment (averages) of the equilibrium distributions of helical coordinates (36 distinct parameters if a basepair step is considered) for any sequence-dependent conformation could be predicted. This removes the need of previous knowledge of the 3D structure, extending Calladine–Dickerson's rules to any possible helical parameter and adding on top the polymorphic landscape of the backbone, the bases, and all the cross-correlations. We generate here ensembles for all the 256 possible tetranucleotide combinations from the 136 unique tetranucleotides found in the miniABC library, by inverting the shift and tilt values of the central step in complementary sequences (i.e. $\text{shift}_{\text{AGTT}} = -\text{shift}_{\text{AACT}}$). Similarly, ensembles for the 64 possible trinucleotides are generated from the first occurrence of the 32 unique trinucleotides with central purine.

Experimental validation of our extended set of rules to predict B-DNA conformation based on the sequence is still difficult to achieve, as DDD is the only sequence that was determined experimentally enough times, using significantly different techniques, protocols, and laboratory conditions, to have a view, yet limited, of the structural het-

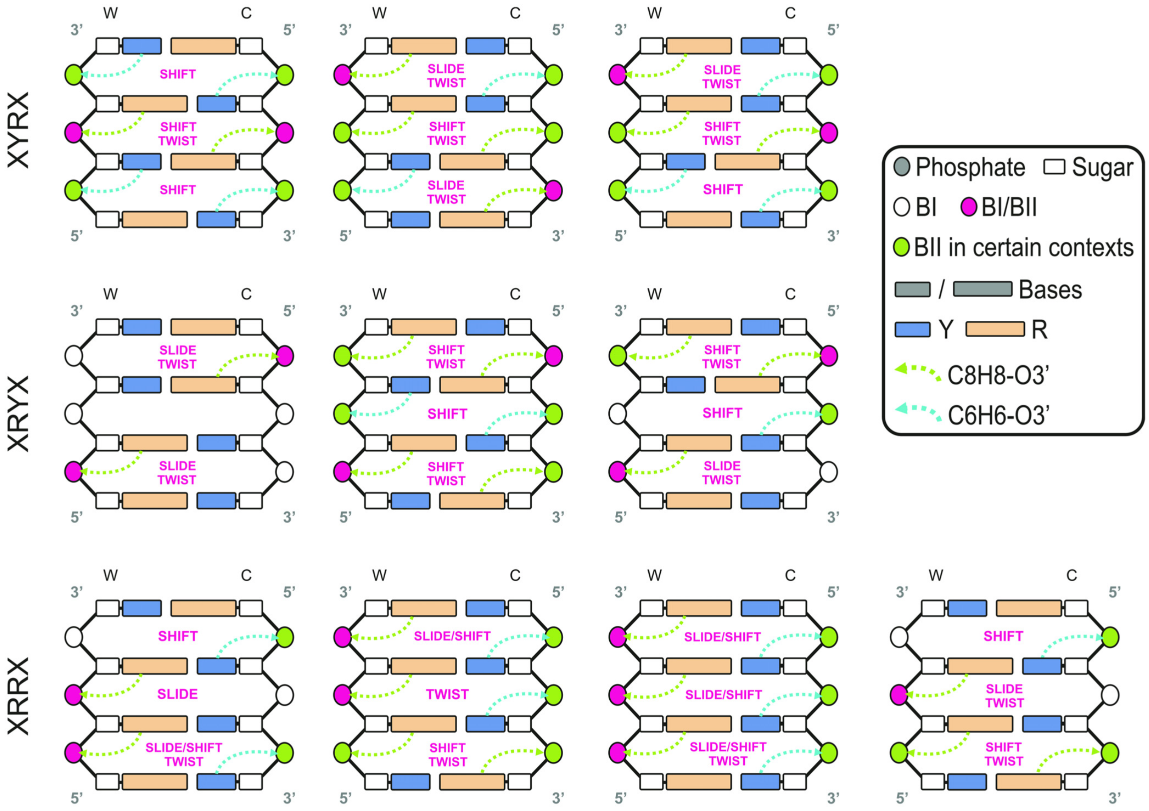

Figure 5. Schema of the polymorphic, or multi-well, landscape exhibited by B-DNA at the tetranucleotide level expressed in the purine (R)/pyrimidine (Y) alphabet, for which only 10 distinct combinations exist, but which still distinguish all possible behaviors. The only helical coordinates that exhibited multi-modality are shift, slide and twist, and each junction in the figure is marked with which coordinates can be multi-modal in it. There is a very high correlation between the occurrence of multi-modality and the formation of a noncanonical hydrogen bond in either the same or a neighboring junction, along with its associated BI/BII backbone transition (see text).

erogeneities of a given isolated B-DNA (9). Moreover, it's the only sequence for which two independent ^{31}P-NMR experiments were performed, and from which accurate BI% were obtained (68,73). Thus, we retrieve 93 structures from the PDB (9), plus the two ^{31}P-NMR experiments to compare the results of our predictive rules on the DDD sequence (Figure 6). Helical conformations predicted in terms of the inter-basepair parameters were in excellent agreement with the experimental ensemble and Calladine–Dickerson rules (see roll and twist in Figure 6), reproducing sequence-dependent features of this prototypical B-DNA sequence. Moreover, we correctly predicted base polymorphisms in shift and twist (5,10,18) and most importantly the backbone substates, in particular, BI/BII (Figure 6). Notwithstanding the foregoing, using our webserver (https://mmb.irbbarcelona.org/miniABC/), the average inter-basepair helical parameters of other B-DNA sequences could be predicted in few seconds giving results in good agreement with experiments determined by NMR and X-ray (Supplementary Figures S20-S27), and comparable to the ones obtained by microsecond long simulations (Supplementary Figures S26 and S27) (25).

The polymorphisms presented thorough this contribution, are dynamic structural heterogeneities that follow two main rules: (i) The crankshaft motion of the DNA backbone that precludes the existence of high BII% in two consecutive junctions (74), and explains the correlations found in consecutive helical parameters (Supplementary Figures S18 and S19) (63). (ii) The existence of conformational frustration, which implies that two or more bimodal elements cannot co-exist at the same time in a given substate. In this sense, when using the information from the second moment (covariance) a dynamic rule appears: If two consecutive high BII (>40%) backbones are predicted in the same strand, the one with the highest value should be kept in BII while the other remains in the BI substate (according to the corresponding value of each peak found by BIC in bimodal distributions). These simple dynamic rules allow to correctly model the structural heterogeneities of B-DNA, capturing the subtle conformational differences found by solution NMR structures for the DDD sequence (Figure 6).

CONCLUSIONS

The analysis of numerous molecular dynamics trajectories obtained with an accurate, last generation, force field has allowed us to derive some general rules concerning the equilibrium conformation distribution of B-DNA, which represent a significant step beyond Calladine–Dickerson earlier propositions. Specifically, we are now able to predict when significantly anharmonic distributions will arise as a function of the tetranucleotide sequence context:

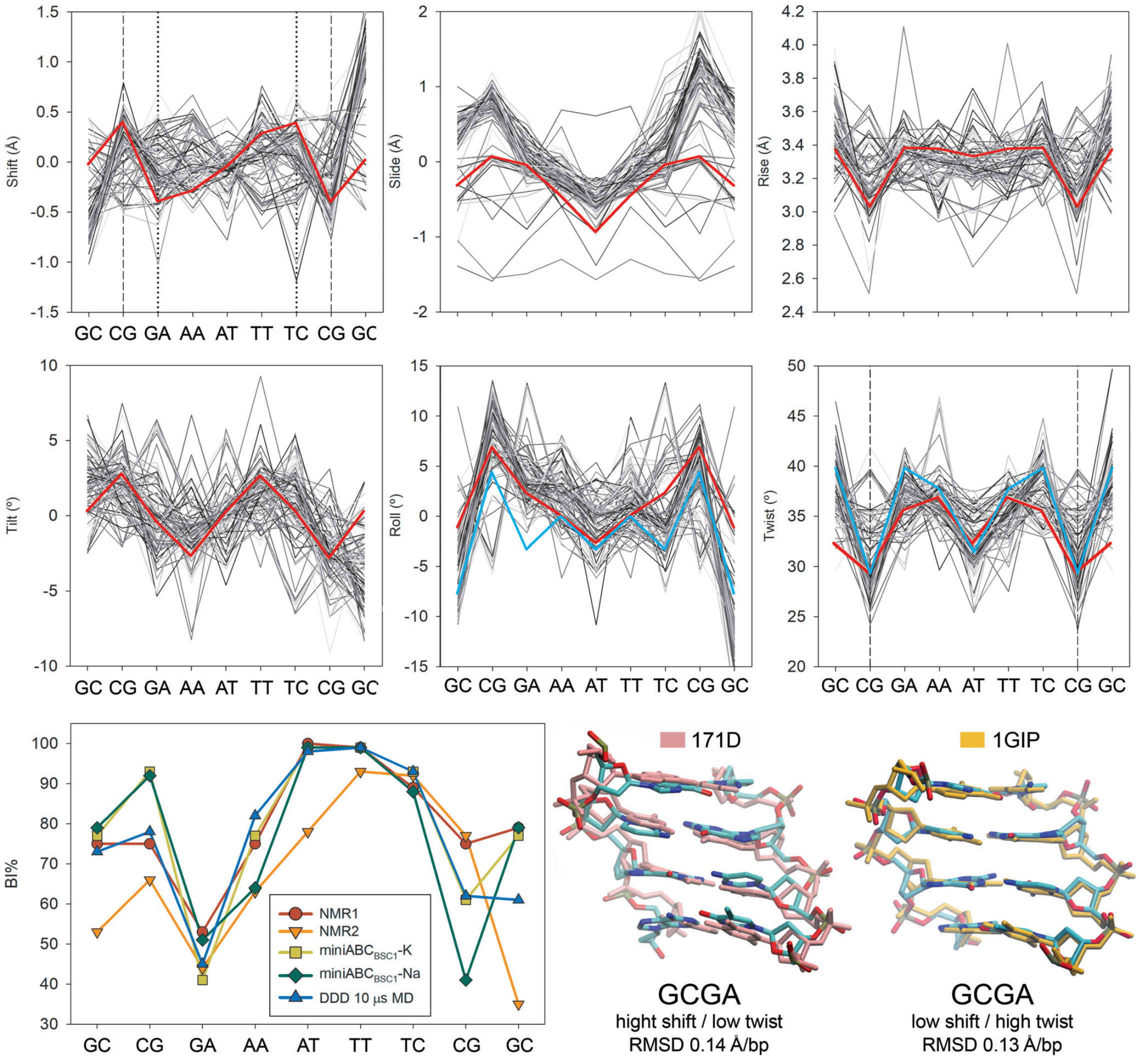

Figure 6. Comparison between experimental structures (X-ray and NMR) determined for the Drew−Dickerson Dodecamer (9), and conformations predicted by using the MD datasets produced herein. The six intra-basepair parameters were predicted (red lines) and compared with all the experimental structures (gray lines) and Calladine–Dickerson prediction for Roll and Twist (cyan line). Vertical dashed lines represent predicted binormal/bimodal steps, while vertical dotted lines represent steps with clear multi-peaked distributions although not bimodal according to Helguerro (see Materials and Methods). In the last row, predicted BI% (yellow and green) were compared with ^{31}P-NMR gold-standard measurements and multi-microsecond long MD simulation of the same sequence using PARMBSC1 force field (9). NMR1 stands for the work by Schwieters *et al.* (74), and NMR2 from Tian *et al.* (68). 3D superposition showing the capability of the miniABC library and our set of rules to predict two conformational substates observed experimentally for DDD at the highly polymorphic GCGA tetranucleotide (NMR structures with PDB id: 171D, and 1GIP). The RMSD was measured after aligning the backbones between the experimental structures and the two substates captured by the miniABC library.

- The first and second moments (average and covariance) of the equilibrium distributions of helical coordinates for DNA can only be understood in terms of nonlocal sequence-dependence contexts, to at least the trinucleotide level for intra-basepair coordinates, and the tetranucleotide level for inter-basepair coordinates.
- A harmonic model of DNA dynamics will not be able to accurately predict third and higher moments of the equilibrium distribution because significant anharmonic movements arise frequently. In fact, the distribution of many inter-basepair coordinates is significantly binormal and, in a non-negligible number of cases, actually bimodal (i.e. multi-peaked). Such bimodality and the relative population of corresponding local minima of the free energy is dependent on the tetranucleotide context. Slide for GG, twist for CG and AG, and shift for YR are the most common steps and helical coordinates exhibiting bimodality, with the tetranucleotides most commonly enhancing bimodality being AGGA, GGGA, ACGG, GCGA, GCGG and GTAA.
- Backbone torsional changes are coordinated in pairs (α/γ, P/χ and ε/ζ). Movements in α/γ lead to the generation of short-lived non-canonical states, which can, however, be populated in the presence of ligands, as it was previously observed for protein-DNA complexes (7). Changes in sugar puckering to the East region leads to lower χ values, while coordinated changes in the ε/ζ pair lead to the BI/BII polymorphism with coupled impacts

on helical parameters. Both ε/ζ and P/χ couplings exhibit sequence dependence.

- The BI/BII conformational change is coupled to the cationic atmosphere surrounding DNA, and to the formation of non-canonical CH—O base-backbone hydrogen bonds. BI/BII transitions are especially prevalent for YRRX sequences and often are associated with bimodality in helical coordinate distributions at the basepair step level. They are a major source of polymorphism in B-DNA. In general, the population of the BII state is coupled to large shift, and low slide and twist at the same junction, but distant and more complex correlations exist between BI/BII conformational states and the helical conformation of neighbouring steps.
- Helical parameters at a given step are not independent but show a complex backbone-mediated pattern of dependencies. For example, shift-slide and roll-twist always show negative correlations, while shift-tilt and slide-twist always show positive correlations. On the contrary, correlations between slide-rise and tilt-roll vary as a function of base sequence. Moreover, helical coordinate correlations may extend to neighboring base pairs as a function of the local sequence.
- Although the differences are subtle, the polymorphic landscape delivered by the simulations in NaCl is slightly richer: (i) non-canonical α/γ conformers are more likely to appear in Na^+ simulations than with K^+; (ii) in general, Na^+ increases the total percentage of the BII substate, but reduces its sequence-dependence; (iii) the appearance of clear bimodality in the inter-basepair helical coordinates is more common in Na^+ simulations than with K^+ (3% in miniABC$_{BSC1}$-K^+ versus 5% in miniABC$_{BSC1}$-Na^+).
- Calladine's principles and Dickerson's sum functions for twist/roll/δ/propeller can now be transformed into quantitative predictions for all the structural features (helical conformations and backbone substates) of canonical DNA sequences. These extended rules have been implemented in a web server that predicts the average conformation of any B-DNA sequence, in terms of the average helical parameters, base and backbone polymorphisms, and P/χ conformations (see https://mmb.irbbarcelona.org/miniABC/).
- Furthermore, using the predictive cgDNA coarse-grained model (and its dinucleotide dependent parameter sets fit to MD simulations), the nonlocal sequence-dependent first (average) and second (covariance) helical coordinate moments can be computed interactively for an arbitrary sequence on the cgDNAweb (75) server http://cgdnaweb.epfl.ch/, including interactive visualization of the expected or ground state conformation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

P.D.D. is a PEDECIBA (Programa de Desarrollo de las Ciencias Básicas) and SNI (Sistema Nacional de Investigadores, Agencia Nacional de Investigación e Innovación, Uruguay) researcher. A.B. and J.W. are La Caixa PhD fellows (UB and IRB Barcelona, Spain). M.O. is an ICREA (Institució Catalana de Recerca i Estudis Avançats) researcher. The authors thank the Ascona B-DNA Consortium for the trajectories in the standard set (μABC). Calculations were performed in the Laboratory for Computation and Visualization of Mathematics and Mechanics at the EPFL, Lausanne, and at IRB Barcelona.

Author contributions: The miniABC sequence library was designed by M.P. and R.L. Simulations were performed by A.S.P. with the assistance of D.P., A.H. and P.D.D. All co-authors were involved in producing results and further discussions. P.D.D. integrated all the results and was the scientific coordinator for the project. P.D.D., J.H.M. and M.O. discussed the analysis and wrote the manuscript with contributions from all the co-authors. G.B. and J.W. built the miniABC web server and web page. The original idea of the project came from P.D.D., R.L., J.H.M. and M.O.

FUNDING

Spanish Ministry of Science [BFU2014-61670-EXP, BFU2014-52864-R]; Catalan SGR; Instituto Nacional de Bioinformática; European Research Council (ERC SimDNA); European Union's Horizon 2020 research and innovation program [676556]; European Union's Horizon 2020 research; BioExcel and MuG EU-projects; Biomolecular & Bioinformatics Resources Platform [ISCIII PT 13/0001/0030] co-funded by the Fondo Europeo de Desarrollo Regional (to M.O.); MINECO Severo Ochoa Award of Excellence, Government of Spain (awarded to IRB Barcelona); CNRS and the ANR project CHROME [ANR-12-BSV5-0017-01 to R.L.]; Swiss National Science Foundation [200020_163324 to J.H.M.]. Funding for open access charge: European Union's Horizon 2020 research and innovation program [676556].

Conflict of interest statement. None declared.

REFERENCES

1. Neidle,S. (2008) *Principles of Nucleic Acid Structure*. Academic Press, London.
2. Fuller,W., Forsyth,T. and Mahendrasingam,A. (2004) Water-DNA interactions as studied by X-ray and neutron fibre diffraction. *Philos. Trans. R. Soc. B Biol. Sci.*, **359**, 1237–1248.
3. Hartmann,B., Piazzola,D. and Lavery,R. (1993) BI-BII transitions in B-DNA. *Nucleic Acids Res.*, **21**, 561–568.
4. Balaceanu,A., Pasi,M., Dans,P.D., Hospital,A., Lavery,R. and Orozco,M. (2017) The role of unconventional hydrogen bonds in determining BII propensities in B-DNA. *J. Phys. Chem. Lett.*, **8**, 21–28.
5. Dans,P.D., Faustino,I., Battistini,F., Zakrzewska,K., Lavery,R. and Orozco,M. (2014) Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.*, **42**, 11304–11320.
6. Zgarbová,M., Jurečka,P., Lankaš,F., Cheatham,T.E., Šponer,J. and Otyepka,M. (2017) Influence of BII backbone substates on DNA twist: a unified view and comparison of simulation and experiment for All 136 distinct tetranucleotide sequences. *J. Chem. Inf. Model.*, **57**, 275–287.
7. Várnai,P., Djuranovic,D., Lavery,R. and Hartmann,B. (2002) Alpha/gamma transitions in the B-DNA backbone. *Nucleic Acids Res.*, **30**, 5398–5406.
8. Pérez,A., Luque,F.J. and Orozco,M. (2007) Dynamics of B-DNA on the microsecond time scale. *J. Am. Chem. Soc.*, **129**, 14739–14745.

9. Dans,P.D., Danilāne,L., Ivani,I., Dršata,T., Lankaš,F., Hospital,A., Walther,J., Pujagut,R.I., Battistini,F., Gelpí,J.L. *et al.* (2016) Long-timescale dynamics of the Drew-Dickerson dodecamer. *Nucleic Acids Res.*, **44**, 4052–4066.
10. Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F., Laughton,C. *et al.* (2014) μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–12283.
11. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem*, **79**, 233–269.
12. Boehr,D.D., Nussinov,R. and Wright,P.E. (2009) The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.*, **5**, 789–796.
13. Calladine,C.R. (1982) Mechanics of sequence-dependent stacking of bases in B-DNA. *J. Mol. Biol.*, **161**, 343–352.
14. Dickerson,R.E. and Klug,A. (1983) Base sequence and helix structure variation in B and A DNA. *J. Mol. Biol.*, **166**, 419–441.
15. Fratini,A. V, Kopka,M.L., Drew,H.R. and Dickerson,R.E. (1982) Reversible bending and helix geometry in a B-DNA dodecamer: CGCGAATTBrCGCG. *J. Biol. Chem.*, **257**, 14686–14707.
16. Cheatham,T.E. III, Brooks,B.R. and Kollman,P.A. (2001) Molecular modeling of nucleic acid structure. *Curr. Protoc. Nucleic Acid Chem.*, **6**, 7.5.1–7.5.12.
17. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 11163–11168.
18. Dans,P.D., Pérez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678.
19. Pérez,A., Luque,F.J. and Orozco,M. (2012) Frontiers in molecular dynamics simulations of DNA. *Acc. Chem. Res.*, **45**, 196–205.
20. Dans,P.D., Walther,J., Gómez,H. and Orozco,M. (2016) Multiscale simulation of DNA. *Curr. Opin. Struct. Biol.*, **37**, 29–45.
21. Beveridge,D.L., Barreiro,G., Byun,K.S., Case,D.A., Cheatham,T.E., Dixit,S.B., Giudice,E., Lankas,F., Lavery,R., Maddocks,J.H. *et al.* (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps. *Biophys. J.*, **87**, 3799–3813.
22. Dixit,S.B., Beveridge,D.L., Case,D.A., Cheatham,T.E., Giudice,E., Lankas,F., Lavery,R., Maddocks,J.H., Osman,R., Sklenar,H. *et al.* (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II. Sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.*, **89**, 3721–3740.
23. Lavery,R., Zakrzewska,K., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dixit,S., Jayaram,B., Lankas,F., Laughton,C. *et al.* (2010) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.*, **38**, 299–313.
24. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A. *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–59.
25. Dans,P.D., Ivani,I., Hospital,A., Portella,G., González,C. and Orozco,M. (2017) How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res.*, **45**, 4217–4230.
26. Pérez,A., Marchán,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A., Orozco,M., McCammon,J.A., McCammon,J.A., Gelin,B.R. *et al.* (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, **92**, 3817–3829.
27. Mintseris,J. and Eisen,M.B. (2006) Design of a combinatorial DNA microarray for protein-DNA interaction studies. *BMC Bioinformatics*, **7**, 429.
28. Medvedev,P., Georgiou,K., Myers,G. and Brudno,M. (2007) Computability of models for sequence assembly. In: *Algorithms in Bioinformatics*. Springer, Berlin, Heidelberg, pp. 289–301.
29. Orenstein,Y. and Shamir,R. (2013) Design of shortest double-stranded DNA sequences covering all k-mers with applications to protein-binding microarrays and synthetic enhancers. *Bioinformatics*, **29**, 71–79.
30. Pérez,A., Marchán,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, **92**, 3817–3829.
31. Case,D., Babin,V., Berryman,J., Betz,R., Cai,Q., Cerutti,D., Cheatham,T. III, Darden,T., Duke,R., Gohlke,H. *et al.* (2014) AMBER14.
32. Salomon-Ferrer,R., Götz,A.W., Poole,D., Le Grand,S. and Walker,R.C. (2013) Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.*, **9**, 3878–3888.
33. Arnott,S. and Hukins,D.W.L. (1972) Optimised parameters for A-DNA and B-DNA. *Biochem. Biophys. Res. Commun.*, **47**, 1504–1509.
34. Berendsen,H.J.C., Grigera,J.R. and Straatsma,T.P. (1987) The missing term in effective pair potentials. *J. Phys. Chem.*, **91**, 6269–6271.
35. Dang,L.X. (1995) Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-crown-6 ether: a molecular dynamics study. *J. Am. Chem. Soc.*, **117**, 6954–6960.
36. Darden,T., York,D. and Pedersen,L. (1993) Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, **98**, 10089.
37. Ryckaert,J.-P., Ciccotti,G. and Berendsen,H.J. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, **23**, 327–341.
38. Hospital,A., Andrio,P., Cugnasco,C., Codo,L., Becerra,Y., Dans,P.D., Battistini,F., Torres,J., Gõni,R., Orozco,M. *et al.* (2016) BIGNASim: A NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.*, **44**, D272–D278.
39. Roe,D.R. and Cheatham,T.E. (2013) PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.*, **9**, 3084–3095.
40. Hospital,A., Faustino,I., Collepardo-Guevara,R., González,C., Gelpí,J.L. and Orozco,M. (2013) NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res.*, **41**, W47–W55.
41. Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
42. Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.
43. Kass,R.E. and Raftery,A.E. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773–795.
44. Schilling,M.F., Watkins,A.E. and Watkins,W. (2002) Is human height bimodal? *Am. Stat.*, **56**, 223–229.
45. de Helguero,F. (1904) Sui Massimi Delle curve dimorfiche. *Biometrika*, **3**, 84.
46. Petkevičiūtė,D., Pasi,M., Gonzalez,O. and Maddocks,J.H. (2014) cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA. *Nucleic Acids Res.*, **42**, e153.
47. Gonzalez,O., Pasi,M., Petkevičiūtė,D., Glowacki,J. and Maddocks,J.H. (2017) Absolute versus relative entropy parameter estimation in a coarse-grain model of DNA. *Multiscale Model. Simul.*, **15**, 1073–1107.
48. Gonzalez,O., Petkevičiūtė,D and Maddocks,J.H. (2013) A sequence-dependent rigid-base model of DNA. *J. Chem. Phys.*, **138**, 055102.
49. Glass,G.V. and Hopkins,K.D. (2008) *Statistical Methods in Education and Psychology*. 3rd edn., Allyn & Bacon, Boston.
50. Pearson,K. (1895) Note on regression and inheritance in the case of two parents. *Proc. R. Soc. London*, **58**, 240–242.
51. R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.
52. Oliphant,T.E. (2007) Python for scientific computing. *Comput. Sci. Eng.*, **9**, 10–20.
53. Hunter,J.D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.
54. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.
55. Mitchell,J.S., Glowacki,J., Grandchamp,A.E., Manning,R.S. and Maddocks,J.H. (2017) Sequence-dependent persistence lengths of DNA. *J. Chem. Theory Comput.*, **13**, 1539–1555.

56. Andreatta,D., Sen,S., Pérez Lustres,J.L., Kovalenko,S.A., Ernsting,N.P., Murphy,C.J., Coleman,R.S. and Berg,M.A. (2006) Ultrafast dynamics in DNA: 'fraying' at the end of the helix. *J. Am. Chem. Soc.*, **128**, 6885–6892.
57. Priyakumar,U.D. and MacKerell,A.D. (2006) NMR imino proton exchange experiments on duplex DNA primarily monitor the opening of purine bases. *J. Am. Chem. Soc.*, **128**, 678–679.
58. Pasi,M. and Lavery,R. (2016) Structure and dynamics of DNA loops on nucleosomes studied with atomistic, microsecond-scale molecular dynamics. *Nucleic Acids Res.*, **44**, 5450–5456.
59. Lankas,F., Lavery,R. and Maddocks,J.H. (2006) Kinking occurs during molecular dynamics simulations of small DNA minicircles. *Structure*, **14**, 1527–1534.
60. Altan-Bonnet,G., Libchaber,A. and Krichevsky,O. (2003) Bubble dynamics in double-stranded DNA. *Phys. Rev. Lett.*, **90**, 138101.
61. Zeida,A., Machado,M., Dans,P. and Pantano,S. (2012) Breathing, bubbling, and bending: DNA flexibility from multimicrosecond simulations. *Phys. Rev. E*, **86**, 021903.
62. McAteer,K. and Kennedy,M.A. (2000) NMR evidence for base dynamics at all TpA steps in DNA. *J. Biomol. Struct. Dyn.*, **17**, 1001–1009.
63. Balaceanu,A., Buitrago,D., Walther,J., Hospital,A., Dans,P.D. and Orozco,M. (2019) Modulation of the helical properties of DNA: next-to-nearest neighbour effects and beyond. *Nucleic Acids Res.*, **47**, 4418–4430.
64. Maehigashi,T., Hsiao,C., Woods,K.K., Moulaei,T., Hud,N.V. and Williams,L.D. (2012) B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. *Nucleic Acids Res.*, **40**, 3714–3722.
65. Djuranovic,D. and Hartmann,B. (2003) Conformational characteristics and correlations in crystal structures of nucleic acid oligonucleotides: evidence for sub-states. *J. Biomol. Struct. Dyn.*, **20**, 771–788.
66. Madhumalar,A. and Bansal,M. (2005) Sequence preference for BI/BII conformations in DNA: MD and crystal structure data analysis. *J. Biomol. Struct. Dyn.*, **23**, 13–27.
67. Heddi,B., Foloppe,N., Bouchemal,N., Hantz,E. and Hartmann,B. (2006) Quantification of DNA BI/BII backbone states in solution. Implications for DNA overall structure and recognition. *J. Am. Chem. Soc.*, **128**, 9170–9177.
68. Tian,Y., Kayatta,M., Shultis,K., Gonzalez,A., Mueller,L.J. and Hatcher,M.E. (2009) ^{31}P NMR investigation of backbone dynamics in DNA binding sites. *J. Phys. Chem. B*, **113**, 2596–2603.
69. Ben Imeddourene,A., Elbahnsi,A., Guéroult,M., Oguey,C., Foloppe,N. and Hartmann,B. (2015) Simulations meet experiment to reveal new insights into DNA intrinsic mechanics. *PLOS Comput. Biol.*, **11**, e1004631.
70. Heddi,B., Foloppe,N., Oguey,C. and Hartmann,B. (2008) Importance of accurate DNA structures in solution: the Jun–Fos model. *J. Mol. Biol.*, **382**, 956–970.
71. Dršata,T., Pérez,A., Orozco,M., Morozov,A. V, Sponer,J. and Lankaš,F. (2013) Structure, stiffness and substates of the Dickerson-Drew dodecamer. *J. Chem. Theory Comput.*, **9**, 707–721.
72. Svozil,D., Kalina,J., Omelka,M. and Schneider,B. (2008) DNA conformations and their sequence preferences. *Nucleic Acids Res.*, **36**, 3690–3706.
73. Schwieters,C.D. and Clore,G.M. (2007) A physical picture of atomic motions within the Dickerson DNA dodecamer in solution derived from joint ensemble refinement against NMR and large-angle X-ray scattering data. *Biochemistry*, **46**, 1152–1166.
74. Calladine,C.R. (2004) Understanding DNA: The Molecule & How It Works. Elsevier Academic Press.
75. De Bruin,L. and Maddocks,J.H. (2018) cgDNAweb: a web interface to the cgDNA sequence-dependent coarse-grain model of double-stranded DNA. *Nucleic Acids Res.*, **46**, W5–W10.