

## A Coarse Grained Model for Atomic-Detailed DNA Simulations with Explicit Electrostatics

Pablo D. Dans, Ari Zeida, Matías R. Machado, and Sergio Pantano\*

*Institut Pasteur de Montevideo, Mataojo 2020, CP 11400 Montevideo, Uruguay*

Received December 4, 2009

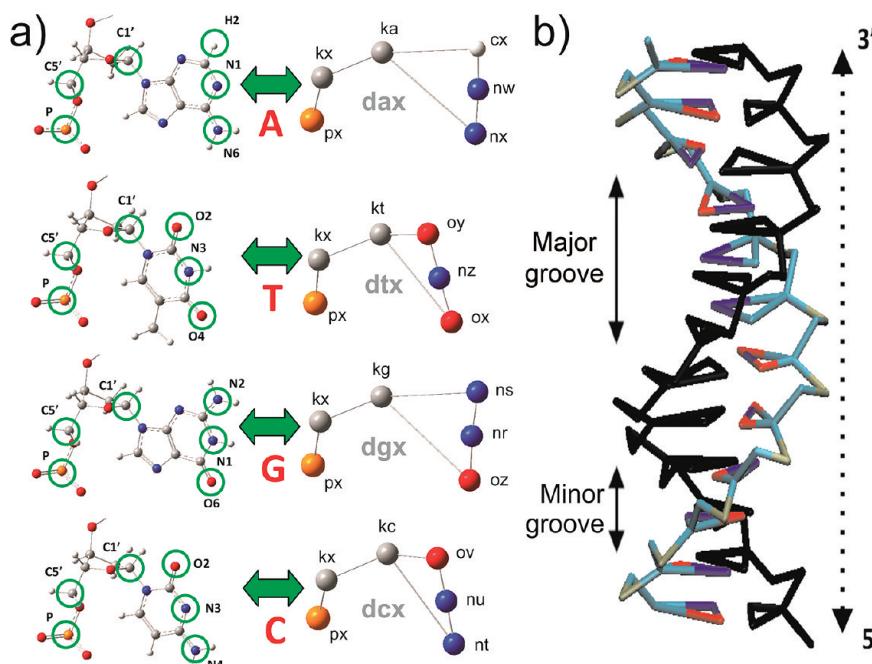
**Abstract:** Coarse-grain (CG) techniques allow considerable extension of the accessible size and time scales in simulations of biological systems. Although many CG representations are available for the most common biomacromolecules, very few have been reported for nucleic acids. Here, we present a CG model for molecular dynamics simulations of DNA on the multi-microsecond time scale. Our model maps the complexity of each nucleotide onto six effective superatoms keeping the “chemical sense” of specific Watson–Crick recognition. Molecular interactions are evaluated using a classical Hamiltonian with explicit electrostatics calculated under the framework of the generalized Born approach. This CG representation is able to accurately reproduce experimental structures, breathing dynamics, and conformational transitions from the A to the B form in double helical fragments. The model achieves a good qualitative reproduction of temperature-driven melting and its dependence on size, ionic strength, and sequence specificity. Reconstruction of atomistic models from CG trajectories give remarkable agreement with structural, dynamic, and energetic features obtained from fully atomistic simulation, opening the possibility to acquire nearly atomic detail data from CG trajectories.

### Introduction

Computer simulations have become a reliable tool for the study of structure and dynamics of soft condensed matter systems, as they expose molecular insights that can be difficult or impossible to obtain with experimental techniques. The continuous motivation to expand the limits imposed by the available computer power has prompted scientists to develop simplified representations that reduce the complexity, size, and conformational degrees of freedom of molecular systems while keeping the physical essence of the interactions that rule their behavior.<sup>1</sup> The remarkable improvement in accuracy and reliability achieved by the so-called coarse-grain (CG) representations, together with the development of new algorithms and computer power, offers currently the possibility to reach biologically relevant time scales and system sizes (see ref 2 for an exhaustive review of the latest developments in CG techniques applied to molecular systems). A wide variety of CG representations are available for the most common biological macromolecules, including

highly complex lipid–protein systems (see, for instance refs 3 and 4). Nevertheless, only a few implementations have been reported for nucleic acids. Among these applications, notable success has been achieved in the description of DNA structure, dynamics, and melting.<sup>5–8</sup> At the base level, some interesting DNA models inspired us in developing our CG model. Zhang and Collins described the B-DNA as a sequence of rigid bodies (base-ribose) connected by flexible rods. Depending on the type of nucleic base (A/T or G/C), four to five centroids were used in the contraction scheme. Molecular dynamics simulations of thermal melting transition were performed using DNA fragments of 100 base pairs (bp).<sup>9</sup> Tepper and Voth developed a DNA model with explicit solvent particles using 14 uniformly distributed centroids per base pair, covalently linked to reproduce the spontaneous formation of the double helix.<sup>5</sup> In the model by Knotts et al.,<sup>6</sup> each base was reduced to three interaction sites with *ad hoc* potentials for stacking and base pairing. This model successfully reproduced salt-dependent melting, bubble formation, and rehybridization. Using wavelet projection to obtain the effective CG potential between effective centroids, the overall deformation response of a DNA

\* Corresponding author. Tel.: +598-2522 0910. Fax: +598-2522 0910. E-mail: spantano@pasteur.edu.uy.



**Figure 1.** Mapping scheme between atomistic and CG models. (a) Circles highlight the coordinates of the elements from the all-atom representation preserved in the CG model. The residue, superatom, and connectivity are displayed. (b) CG representation of a 12-mer double helix DNA in the canonical B-form that illustrates grooves and 5'-3' direction (black strand).

molecule was achieved with molecular dynamics (MD) techniques.<sup>7</sup> Representing the DNA as a worm-like polymer and using the “rigid base pair model”, homogeneous elastic properties were reproduced by fitting the model against experimental data.<sup>8</sup> In the Mergell et al. model of DNA, each base pair was represented by a rigid ellipsoid linked to the backbone by semirigid harmonic springs.<sup>10</sup> Recently, CG models of DNA were devoted to protein–DNA docking, by optimizing the interaction surface between the macromolecular partners.<sup>11</sup> Similarly, simplified Go-models for RNA have accomplished the description of folding dynamics under varying temperatures and mechanical stretches.<sup>12,13</sup> With a less detailed representation, RNA<sup>14</sup> and also DNA<sup>15</sup> molecules were reduced to only one centroid per nucleotide to study the packing dynamics of a virus genome inside the protein capsid. In this last DNA study, an implicit solvent approach was used to mimic the biological environment.<sup>15</sup> These kinds of models have also been applied with success to the description of large molecular aggregates such as nucleosomes and ribozymes.<sup>16–22</sup>

In this contribution, we present a new CG model for MD simulation of nucleic acids ruled by a Hamiltonian function identical to that used by the most popular MD simulation packages. Electrostatic interactions are treated within the framework of the generalized Born model for implicit solvation.

The model reproduces canonical structures as well as conformational transitions from the A to B form of DNA. We obtain also a good reproduction of the temperature, size, and sequence-specific and ionic strength driven melting. The breathing dynamics of poly(AT) domains were compared with experiments raising comparable life times for end-fraying and also internal hydrogen bonds disruption at the base pair level. Reconstruction of all-atom trajectories from

CG MD runs shows a high-quality reproduction of geometrical features with maximum deviations on the order of 2–3 Å with respect to the experimental structures and/or all-atom simulations.

## Methods

**Coarse Grain Mapping.** Our CG model reduces the complexity of a nucleotide to six effective interaction sites (hereafter called superatoms) for each type of canonical nucleotide in DNA (A, T, C and G). This defines four different coarse-grained bases (dax, dtx, dcx, and dgx), which map to the all-atom nucleotides as illustrated in Figure 1a retaining the “chemical sense” of the interactions. Each of the six superatoms was placed on the Cartesian coordinates of one element in the all-atom representation and condensed the molecular information from its atomic neighborhood. The number of superatoms chosen retains the Watson–Crick interaction sites and preserves the asymmetry in the backbone, the identity of the minor and major grooves, as well as the 5'-3' polarity of the DNA strands (see Figure 1b). Under this scheme, the total mass of the individual atoms of the real nucleotides, including hydrogen, is condensed onto the superatoms, as shown in Table 1.

Phosphate groups are represented by the px superatoms placed on the position of the corresponding phosphorus. The position of the C5' atom was used to place the superatom kx, which serves to establish the 5'-3' direction of each DNA strand and allows for the formation of the major and minor grooves (see Figure 1b). The kn superatom (where kn = ka, kt, kc, or kg) lays at the position of the C1' atom. The superatoms that participate in the Watson–Crick interactions are placed in the same position as the corresponding atoms preserving the molecular specificity between

**Table 1.** Masses, Charges, and Lennard-Jones Parameters Assigned to the Superatoms

superatoms <sup>a</sup>	mass	atoms represented <sup>b</sup>	charges (e)	Lennard-Jones	
				$\varepsilon$ (kcal/mol)	$\sigma$ (Å)
px	78.97	P+O1P+O2P+O5'	-1.00	0.2000	2.6000
kx	73.07	C5'+C4'+C3'+O3'+O4'	0.00	0.1094	2.4080
ka	41.05	C1'+C2'+N9	0.00	0.1094	1.9080
nx	40.03		0.35	0.1900	1.8240
nw	40.03	(C8+N7+C5+C4+C6+N6+N1+N3+C2) <sup>c</sup>	-0.35	0.1900	1.8240
cx	40.03		0.00	0.1094	1.9080
kt	41.05	C1'+C2'+N1	0.00	0.1094	1.9080
ox	37.03		-0.35	0.2400	1.6612
nz	37.03	(C6+C5+O4+C4+N3+O2+C2+C) <sup>c</sup>	0.70	0.1900	1.8240
oy	37.03		-0.35	0.2400	1.6612
kg	41.05	C1'+C2'+N9	0.00	0.1094	1.9080
oz	45.71		-0.70	0.3100	1.6612
nr	45.71	(C8+N7+C5+C4+N3+C2+N1+C6+O6+N2) <sup>c</sup>	0.35	0.2600	1.8240
ns	45.71		0.35	0.2600	1.8240
kc	41.05	C1'+C2'+N1	0.00	0.1094	1.9080
nt	32.03		0.70	0.2600	1.8240
nu	32.03	(C6+C5+C4+N4+N3+O2+C2) <sup>c</sup>	-0.35	0.2600	1.8240
ov	32.03		-0.35	0.3100	1.6612

<sup>a</sup> The types of the superatoms match those included in the coordinate and topology files that are available from the authors upon request.

<sup>b</sup> Hydrogen atoms are omitted for brevity. Their masses are added to the corresponding heavy atoms. <sup>c</sup> The sum of the masses is equally distributed among the three superatoms.

both DNA strands. In this sense, all-atom Watson–Crick hydrogen bonds are shrunk to two-point electrostatic interactions in the CG model.

This scheme leads to an easy mapping/back-mapping from all-atom to CG representation and vice versa. Using internal coordinates and canonical distances, angles, and dihedrals from the B-form of Arnott et al.,<sup>23</sup> we can recover the complete all-atom picture. Dynamic events in the ps–ns time scale can be followed within a multi-microsecond trajectory calculated at the CG level. To this aim, we developed an algorithm that uses as input the instantaneous position of three superatoms to infer the Cartesian coordinates of the atoms in the neighborhood in each MD frame. A Fortran 90 implementation of the homemade algorithm is provided in Table S1 as Supporting Information. The reconstruction to the all-atom picture is made in three steps proceeding from the base to the phosphate moiety (see Table S2 in the Supporting Information for a pseudo-code explaining the algorithm). Since we have less information about the sugar conformation and the dihedrals involved in the phosphodiester bond, a loss of accuracy of the back-mapped coordinates in the backbone region can be expected (see Figure S1 in the Supporting Information). To correct the positioning of the sugar moiety and the distances of the phosphodiester bonds, 150 steps of geometric optimization were performed on each frame after the complete CG to all-atom reconstruction (see Figure S2 in the Supporting Information).

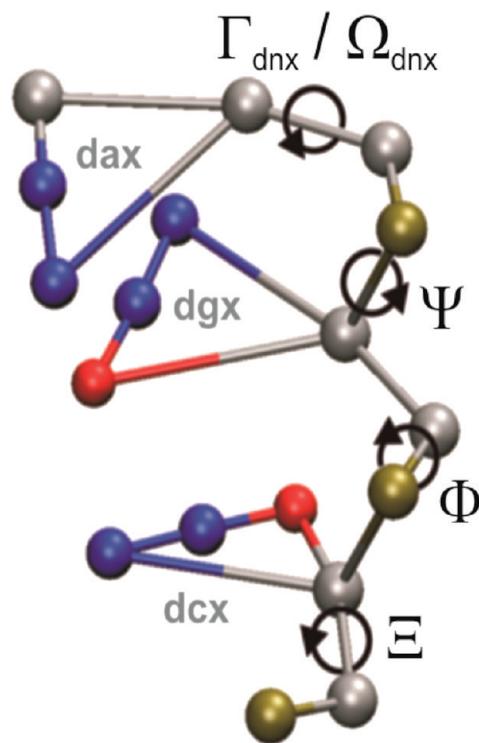
**Parameterization.** With the aim of maximizing the transferability between different MD packages, our model employs a widely used Hamiltonian function:

$$U = \sum_{\text{bonds}} k_b (r_{ij} - r_{\text{eq}})^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_{\text{eq}})^2 + \sum_{\text{dihedrals}} \frac{V_k}{2} [1 + \cos(n_k \varphi - \gamma_k^{\text{eq}})] + \sum_l \sum_{l>m}^N \left\{ 4\varepsilon \left[ \left( \frac{\sigma}{r_{lm}} \right)^{12} - \left( \frac{\sigma}{r_{lm}} \right)^6 \right] + \frac{q_l q_m}{\epsilon r_{lm}} \right\} \quad (1)$$

where  $k_b$  is the bond stretching constant,  $r_{ij} = r_i - r_j$ , and  $r_{\text{eq}}$  is the equilibrium bond distance between two linked elements.  $k_\theta$  is the bond angle constant.  $\theta$  is the instantaneous angular value defined by three successive elements, and  $\theta_{\text{eq}}$  is the equilibrium bond angle.  $V_k$  is the height of the torsional barrier;  $n_k$  is its periodicity.  $\varphi$  is the torsion angle defined by four consecutively bonded elements, and  $\gamma_k^{\text{eq}}$  is the phase angle. In the fourth term, the sum runs over all the particles of the system ( $N$ ). This term corresponds to the Lennard-Jones and Coulombic potentials, in which  $\varepsilon$  is the maximum depth of the function and  $\sigma$  is the zero energy point or van der Waals diameter. While the values of  $\varepsilon$  were used as free parameters, those of  $\sigma$  for the backbone superatoms were set to roughly match the excluded volume of the groups of atoms represented (see Table 1). Superatoms participating in the base preserve the  $\sigma$  values coming from the corresponding heavy atoms to avoid artifacts that could disrupt the intra-base-pair step (rise). Lastly,  $q_{l,m}$  is the charge of each superatom, and  $\epsilon$  is the vacuum permittivity.

Hydration and ionic strength effects were taken into account using the generalized Born (GB) model<sup>24</sup> for implicit solvation as implemented in AMBER.<sup>25</sup> The Born effective radii were fixed to 1.5 Å for all superatoms.

In the present model, the equilibrium bond distances and bond angles were taken from the canonical B-form of Arnott et al.<sup>23</sup> The bond stretching and bond angle constants were fixed to 400 kcal/mol·Å<sup>2</sup> and 75 kcal/mol·rad<sup>2</sup> for all bonds and angles, respectively (eq 1). The torsional barrier for the three dihedral angles of the backbone was fixed to 10 kcal/mol (see  $\Phi$ ,  $\Xi$ , and  $\Psi$  in Figure 2 and Table 2). The periodicity of dihedral angles was set to nearly reproduce the canonical conformations of the B-form of Arnott et al.<sup>23</sup> To complete the model, two more torsions,  $\Gamma_{\text{dnx}}$  and  $\Omega_{\text{dnx}}$ , that act on the same bond as  $\Omega$  were added (where dnx stands for each of the four bases: dax, dtx, dgx, and dcx). The parameters for the  $\Gamma_{\text{dnx}}$  and  $\Omega_{\text{dnx}}$  dihedral angles, which can be visualized in Figure 2, are specific for each nucleic base. All the torsional parameters used in our model are displayed in Table 2.



**Figure 2.** Dihedral angles used in the CG model. Three dihedrals account for the backbone movements for which the parameters are the same regardless of the nucleobase ( $\Phi = kn\text{-}px\text{-}kx\text{-}kn$ ,  $\Xi = px\text{-}kx\text{-}kn\text{-}px$ , and  $\Psi = kx\text{-}kn\text{-}px\text{-}kx$  where  $kn = ka, kt, kc$ , or  $kg$ ). The dihedral angles  $\Xi$ ,  $\Omega_{dnx}$ , and  $\Gamma_{dnx}$  act on the same bond but are defined using different superatoms ( $dnx = dax, dtx, dgx, dcx$ ). See Table 2 for dihedral angles definition.

**Benchmark System: The Drew–Dickerson Dodecamer.** To validate the structural, dynamical, and energetic behavior of our CG scheme, the results presented in the first part of this contribution correspond to the Drew–Dickerson dodecamer of DNA (also called the EcoRI dodecamer),<sup>26–28</sup> which was used as a benchmark system. This dodecamer of sequence 5'-d(CGCGAATTCTCGCG)-3' has been largely studied by means of experimental and theoretical works, giving rise to a solid bibliographic base to compare our results.<sup>29–33</sup> As the starting structure for the CG simulation (labeled DDcgB), the Drew–Dickerson dodecamer was built<sup>25</sup> in the canonical B-form of Arnott et al.<sup>23</sup> During simulation, nonbonded interactions were calculated up to a

cutoff of 18 Å within the GB approximation, and the salt concentration was set to 0.15 M. Temperature was controlled using a Langevin thermostat<sup>34,35</sup> with a friction constant of 50 ps<sup>-1</sup>, which approximates the physical collision frequency for liquid water.<sup>36</sup> The random seed generator of the stochastic force was randomly changed every restart of the simulation (every 1 μs) to avoid quasi-periodic oscillations. The temperature was raised linearly from 0 to 298 K in 5 ns. After that point, production runs of 5 μs were performed, and snapshots were recorded for analysis every 50 ps using a time step of 5 fs to integrate the classical equation of motion. To avoid the fraying of the helix ends frequently observed in long MD simulations,<sup>37</sup> loose harmonic restraints of 3.0 kcal/mol·Å<sup>2</sup> were added to preserve the Watson–Crick hydrogen bonds of the capping base pairs.

To compare our results with state-of-the-art molecular dynamic simulations, the same sequence was built in the Arnott B-form,<sup>23</sup> solvated with explicit water molecules, and surrounded by K<sup>+</sup> and Cl<sup>-</sup> ions to mimic the physiological conditions (this system was labeled DDaaB). The all-atom molecular dynamic simulation of the unconstrained Drew–Dickerson dodecamer was performed using the parm99<sup>38</sup> force-field with the correction proposed by Orozco and co-workers for nucleic acids (parmbsc0).<sup>39</sup> Ions were treated with the same force-field. The final system contained 36 K<sup>+</sup>, 14 Cl<sup>-</sup>, and 3926 TIP3P water molecules<sup>40</sup> in a truncated octahedral box. Initially, the water molecules and ions were relaxed by 1000 steps of energy minimization imposing harmonic restraints of 25 kcal/mol·Å<sup>2</sup> to DNA. Subsequently, four energy minimization runs were performed (with the same number of steps) where the restraints on DNA were gradually reduced from 20 to 5 kcal/mol·Å<sup>2</sup>. All optimizations and equilibration MD simulations were performed using constant volume. Long-range interactions were treated using the PME approach<sup>41</sup> with a 12 Å direct space cutoff. The last optimized structure was taken as the starting point for the MD simulations. The entire system was then heated from 0 to 300 K during a 200 ps MD run with harmonic restraints of 5.0 kcal/mol·Å<sup>2</sup> imposed to DNA at a constant volume. Final temperature and a constant pressure of 1 atm were then reached by coupling the system to the Berendsen thermostat and barostat, respectively.<sup>42</sup> Fifty nanoseconds of production MD simulation were performed in the isobaric–isothermal ensemble. An integration time step of 2 fs was used, and all

**Table 2.** Torsional Parameters Used in eq 1 for the CG-DNA Model<sup>a</sup>

dihedral	torsional parameters											
	$V_1^b$	$V_2$	$V_3$	$V_4$	$n_1$	$n_2$	$n_3$	$n_4$	$\gamma_1^{eq}$	$\gamma_2^{eq}$	$\gamma_3^{eq}$	$\gamma_4^{eq}$
kn <sup>c</sup> -px-kx-kn ( $\Phi$ ) <sup>c</sup>	10.0				8				161.0			
px-kx-kn-px ( $\Xi$ )	10.0				8				-153.2			
kx-kn-px-kx ( $\Psi$ )	10.0				4				-29.3			
px-kx-ka-nx ( $\Omega_{dax}$ )	10.0	6.0	7.0	10.0	1	7	2	1	118.0	47.0	20.0	-220.0
px-kx-ka-cx ( $\Gamma_{dax}$ )	6.0	4.0	2.0		1	3	4		65.0	145.0	130.0	
px-kx-kt-ox ( $\Omega_{dtx}$ )	10.0	5.0	7.0	10.0	1	8	2	1	117.0	47.0	20.0	-140.0
px-kx-kt-oy ( $\Gamma_{dtx}$ )	6.0	4.0	2.0		1	3	4		65.0	145.0	130.0	
px-kx-kg-oz ( $\Omega_{dgx}$ )	10.0	6.5	7.0	10.0	1	6	2	1	110.0	90.0	20.0	-220.0
px-kx-kg-oz ( $\Gamma_{dgx}$ )	6.0	4.0	2.0		1	3	4		65.0	145.0	130.0	
px-kx-kc-nt ( $\Omega_{dcx}$ )	10.0	5.0	7.0	10.0	1	8	2	1	117.0	47.0	20.0	-140.0
px-kx-kc-ov ( $\Gamma_{dcx}$ )	6.0	4.0	2.0		1	3	4		65.0	135.0	130.0	

<sup>a</sup> See Figure 2 for a comprehensive identification of the  $\Phi$ ,  $\Xi$ ,  $\Psi$ ,  $\Omega_{dnx}$ , and  $\Gamma_{dnx}$  angles. <sup>b</sup> See third term in eq 1. <sup>c</sup> Where  $kn = ka, kt, kc$ , or  $kg$ .

bond lengths involving hydrogen atoms were restrained using the SHAKE algorithm.<sup>43</sup>

Using the *ptraj* utility of AMBER,<sup>25</sup> root mean square deviations (RMSD) were calculated on all the superatoms/atoms of each residue. The mobility of the bases relative to the backbone was evaluated by comparing atomic B-factors against experimental data. We calculated the quotient between the B-factors of the phosphate atom/superatom and the central heavy atom/superatom engaged in the Watson–Crick interaction (N1 for purines and N3 for pyrimidines). The CG trajectories were back-mapped to all-atom representation and, together with the state-of-the-art MD simulations, analyzed with the program Curves 5.1<sup>44</sup> to monitor the effects of thermal fluctuations upon the major determinants of the B-DNA molecular structure. Root mean square fluctuations (RMSF) and time evolution were calculated for selected helical parameters. The *anal* module of AMBER<sup>25</sup> was used to calculate the interaction energies between bases, strands, GC pairs, and AT pairs in terms of electrostatic and van der Waals contributions. When analyzing back-mapped trajectories, in all the cases, only a discontinuous 50-ns-long trajectory containing the final 10 ns of each microsecond was taken into account for shortness. For comparison purposes, calculated properties were also obtained for crystallographic and averaged NMR derived data (PDB structures 1BNA<sup>45</sup> and 2DAU,<sup>46</sup> respectively).

All MD simulations were carried out using the *sander* module of AMBER 10.<sup>25</sup> Molecular drawings were performed with VMD 1.8.6.<sup>47</sup>

**DNA Melting.** The CG model was tested to reproduce thermal melting for several systems analyzing the effect of variable length, GC content, and ionic strength of the medium. The sequences chosen were taken from the recently determined experimental work by Owczarzy and co-workers:<sup>48</sup>

- (i) 5'-d(ATCGTCTGGA)-3' (seq10)
- (ii) 5'-d(TACTAACATTAACCA)-3' (seq15a)
- (iii) 5'-d(GCAGTGGATGTGAGA)-3' (seq15b)
- (iv) 5'-d(GCGTCGGTCCGGGCT)-3' (seq15c)
- (v) 5'-d(AGCTGCAGTGGATGTGAGAA)-3' (seq20)

Separated runs were carried out for ionic strengths of 0.07, 0.12, 0.22, and 1.0 M. The melting protocol was the same for each sequence studied and consisted of 3.0  $\mu$ s of MD simulation in which the temperature was raised 100 °C in five steps of 20 °C. Each step consisted of 0.1  $\mu$ s of heating followed by 0.5  $\mu$ s simulated at constant temperature. No restraints were added to the capping base pairs.

To define a melting criterion, hydrogen bonds between base pairs were considered to exist if the distance between the corresponding “acceptor” and “donor” superatom was less than 4.0 Å. The characteristic melting temperature is reached when 50% of the base pairs are in an open state. To generate the melting curves, the percentage of the opened base pairs within the sequence was calculated for each frame of the simulation. Adjacent averaging every 500 frames was performed to clean out the noise. Averaged points were sorted from lowest to highest temperatures, and a sigmoid fit with the Gompertz 4 parameters equation was applied:

$$y_0 + ae^{-e^{-(T-T_0)/b}} \quad (2)$$

This procedure yields one single continuous function of temperature. In eq 2,  $T_0$  is the abscissa of the inflection point, which corresponds to the calculated melting temperature. The regression coefficients for all the sigmoid fits were always >0.8. Results were integrally obtained from the total CG trajectories. Notice that the back-mapping procedure was not applied.

**The A to B Transition.** The Drew–Dickerson sequence was also built in the A-form of Arnott et al.<sup>23</sup> to test the capability of the model to reproduce a conformational transition from the A to the B form (DDcgA). Five microseconds of coarse grained MD simulations were run under the same conditions used in the DDcgB system. RMSDs with respect to the experimental and canonical B-form structures, pitch, and minor and major groove width were calculated to evaluate the structural transition.

**DNA Breathing Dynamics.** Finally, we studied the breathing movement of the Drew–Dickerson dodecamer and a 29-bp-long double-stranded DNA: 5'-d(GGCGCCCAATAT-AAAATATTAAAATGCGC)-3'. The sequence contains a GC clamp domain (G1 to C7) and a long AT track that corresponds to a breathing domain (A8 to A24). The simulation conditions were fixed to roughly match the experimental work by Altan-Bonnet and co-workers.<sup>49</sup> The most relevant difference resided in the fact that the sequence used by Altan-Bonnet et al. contained a thymine tetraloop to avoid the separation of both strands. However, since the structure of this loop is unknown, we decided to replace it by loose harmonic restraints of 3.0 kcal/mol·Å<sup>2</sup> to preserve the Watson–Crick hydrogen bonds of the last base pair (5'-C<sub>29</sub>-3' in strand1 and 5'-G<sub>1</sub>-3' in strand2).

The criterion to define the base opening/closing was identical to that established for melting. MD simulations of 4  $\mu$ s at 37 °C with an ionic strength of 0.1 M were performed.

## Results and Discussion

A major goal for molecular simulations is not only the reproduction of stable trajectories of molecular systems oscillating around equilibrium conformations but also to achieve the capacity to explore the accessible conformational space and evolve toward more stable conformations. In the following paragraphs, we provide some examples of the performance of our model to reproduce the structure, energetics and dynamics of stable trajectories around equilibrium configurations, melting of DNA, conformational transitions, and breathing dynamics.

**Benchmark System: CG Model vs All-Atom.** All simulations started with the canonical B-form and were stable along all the simulation time. A first measure of the quality of the CG model can be obtained from a direct comparison between the whole trajectories of CG and all-atom representations. To this aim, we calculated the RMSD using all the superatoms in the CG model and the corresponding atoms in the all-atom trajectory (according to the mapping presented in Figure 1). We found that the intrinsic fluctuations during CG and all-atom schemes were very similar. Furthermore, the structural models obtained from both simulations with respect to the experimental structures are practically identical

**Table 3.** Structural Comparison between CG and All-Atom Simulations<sup>a</sup>

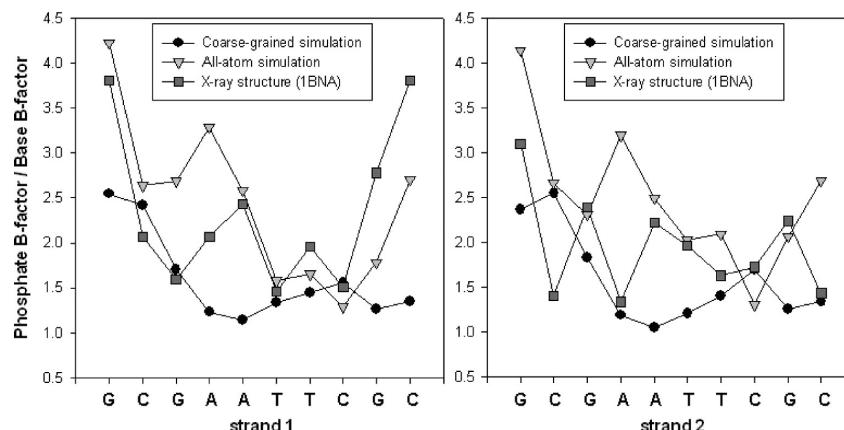
	mean during MD trajectory	starting conformer (B form)	X-ray (1BNA <sup>45</sup> )	NMR (2DAU <sup>46</sup> )
DDcgB	1.0 ± 0.3	1.8 ± 0.3	2.3 ± 0.3	3.1 ± 0.3
DDaaB	1.6 ± 0.4	2.8 ± 0.4	2.6 ± 0.4	2.7 ± 0.4

<sup>a</sup> RMSD are calculated over 5 μs and 50 ns for the CG and all-atom trajectories, respectively. Values are reported in Angstroms.

(Table 3). Only subtle differences appear when comparing both trajectories against the reference structures.

To analyze the internal flexibility of the dodecamer, B-factors were calculated for selected groups of atoms/superatoms and were compared with the values coming from the X-ray experiments (PDB structure 1BNA). Absolute B-factors calculated from the all-atom trajectory differ significantly from those determined using the CG approach and the X-ray experiments. Only global qualitative trends for the structure as a whole could be obtained. However, the B-factors of the phosphorus elements relative to those of atoms belonging to the base moiety are good descriptors of the relative mobility of different segments of the nucleobases. A comparison between these values indicates that the all-atom simulation (DDaaB) always has the highest mobility, while the coarse-grained version (DDcgB) always presents the lowest (Figure 3). As shown, the relative values were always greater than 1.0 for all the systems, pointing out, as expected, the higher mobility of the backbone with respect to the base. In general, we observe that the relative mobility is lower in the CG model. This can be related to the reduced number of degrees of freedom or to a nonoptimal mass distribution.

**Benchmark System: Back-Mapped CG Model vs All-Atom.** Despite these encouraging results, it becomes difficult to establish a direct comparison between both simulations. Therefore, we sought to extract atomistic information from our CG model. To this end, we back-mapped the last 10 ns of each microsecond from our CG trajectory (DDcgB). This generated an atomistic noncontiguous 50-ns-long trajectory that is directly comparable with that of the all-atom simulation (DDaaB).



**Figure 3.** Higher mobility of phosphate groups. B-factors for the phosphorus atoms/superatoms relative to the central elements in the Watson–Crick interaction region along both strands. The coarse-grained (DDcgB) and the all-atom simulation (DDaaB) were compared to the experimental B-factors obtained from the X-ray structure with the PDB code 1BNA.

**Table 4.** Structural Comparisons for the Drew–Dickerson Structure d(CGCGAATTCGCG)<sub>2</sub><sup>a,b</sup>

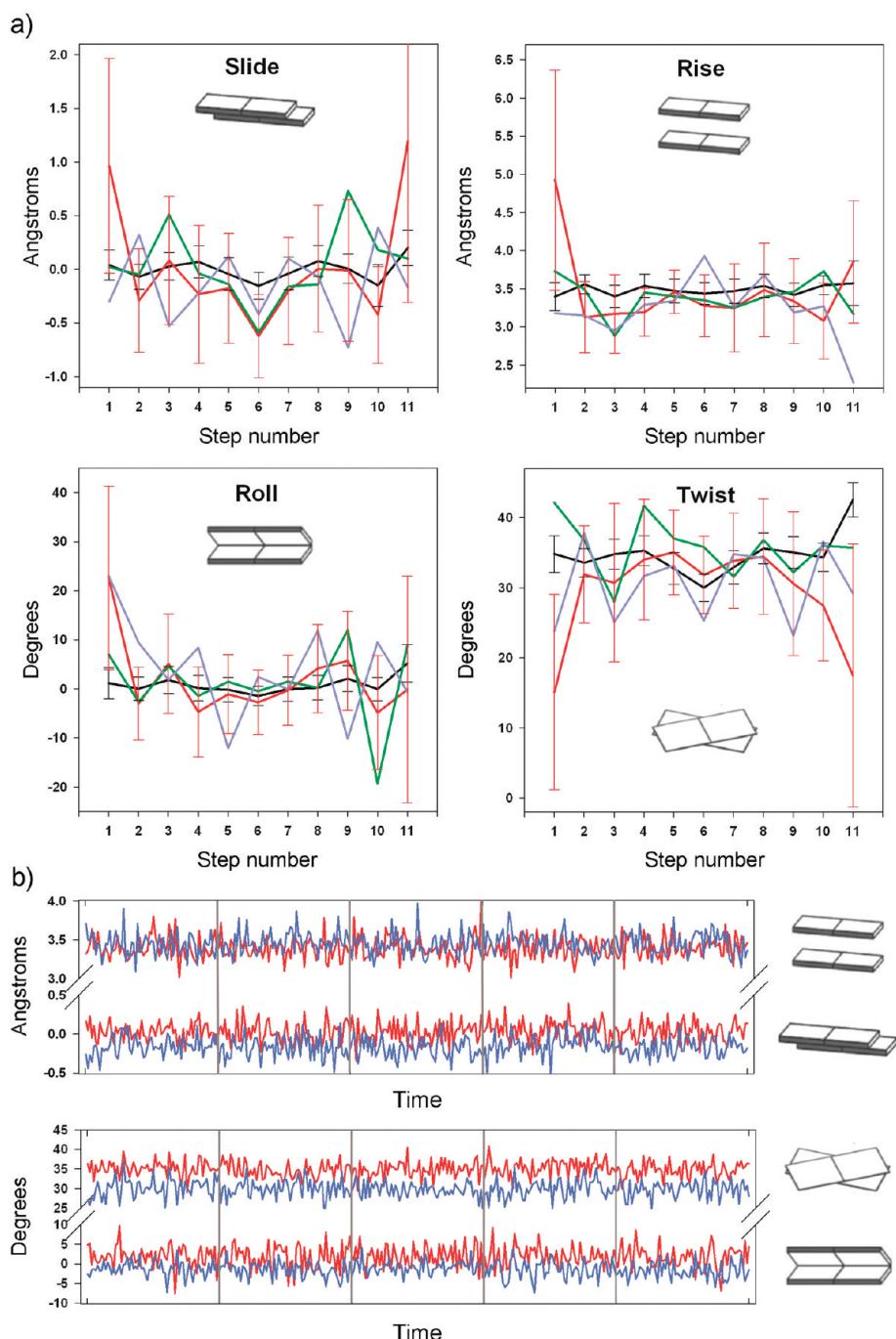
	DDcgB	DDaaB	Arnott-A	Arnott-B	1BNA	2DAU
DDcgB			6.5	1.8	2.3	3.1
DDaaB			5.6	3.0	2.8	2.8
Arnott-A	1.7	2.0		6.3	6.0	4.8
Arnott-B	0.9	1.5	1.5		1.4	3.4
1BNA	1.3	1.2	1.9	0.9		3.3
2DAU	1.5	1.4	1.9	1.5	1.6	

<sup>a</sup> Heavy-atom RMSD between the specified structures. <sup>b</sup> The upper-right portion represents RMSD fit measured in Å calculated over all the atoms. The lower-left portion represents RMSD fit calculated for the four base pairs underlined in the heading, i.e., residues 3–6 and 19–22.

**Structural and Dynamical Comparison.** Table 4 presents a comparative view of both simulations against the canonical A and B conformations and two experimental structures. The averaged RMSD for the DDaaB simulation was 2.8 Å apart from both the crystallographic (1BNA) and NMR (2DAU) structures and 3.0 Å with respect to the canonical B-form. Analogously, the family of structures obtained with the CG model remained 2.3 Å, 3.1 Å, and 1.8 Å apart from the X-ray structure, NMR structure, and canonical B-form, respectively (upper-right portion of Table 4).

If we consider the averaged RMSD calculated for the selected inner four base pairs (residues 3–6 and 19–22), the values are almost the same between DDcgB and DDaaB with respect to both experimental structures (lower-left portion in Table 4). We can conclude that the differences between all-atom and back-mapped CG simulations are rather subtle, and that both simulations sample very similar or equivalent conformational spaces.

A more stringent evaluation of the quality of the B-form reached by the CG model can be obtained from a comparison of the fluctuations of some selected helical parameters (Figure 4). RMSFs were calculated for the Slide, Rise, Roll and Twist, which are the most distinctive base pairs parameters between the A and B canonical forms (Figure 4a). The large fluctuations observed in the helix ends of DDaaB were not present in DDcgB due to the loose harmonic restraints imposed to preserve the Watson–Crick hydrogen bonds of the capping base pairs in the implicit solvent simulation.



**Figure 4.** Selected helical parameters. (a) RMSF of the Slide, Rise, Roll, and Twist. The red line corresponds to DDaaB and the black line to DDcgB. Experimental structures 1BNA and 2DAU are represented by the green and the blue lines, respectively. Average values and standard deviations are plotted in Angstroms for the Slide and Rise and in degrees for the Roll and Twist parameters. The values are presented along the helix from the 5' to 3' direction (*x* axis). (b) The same helical parameters for two selected intra-base steps (C3/G4 in blue and A6/T7 in red) were plotted along 50 noncontiguous nanoseconds of the back-mapped DDcgB simulation.

Although the fluctuations about the mean values were in general somewhat larger in DDaaB versus DDcgB, the averages exhibited similar trends, especially in the Slide and Twist parameters. Compared to the all-atom simulation, the coarse-grained model exhibited a similar sequence-dependent trend in the Slide and Twist parameters for the CG, GA, AA, AT, TT, and TC dinucleotides (DNA steps 3–8 in Figure 4a).

A more dynamical picture of the structural stability can be acquired following the instantaneous values of the helical

parameters during the simulation time. The same selected helical parameters are plotted against time for the back-mapped noncontiguous 50 ns trajectory. For the sake of brevity and clarity, only the C3/G4 and A6/T7 dinucleotides are plotted in Figure 4b. A first global inspection of Figure 4b illustrates the stability of the simulation, as no drift could be observed in the values of the parameters against the simulation time. The Rise and Slide fluctuated around the canonical values, and the Roll showed a distinctive behavior between the C3/G4 and A6/T7 dinucleotides comparable with

**Table 5.** Comparison of Averaged<sup>a</sup> Electrostatic and van der Waals (VdW) Interactions

	electrostatic (kcal/mol)		VdW (kcal/mol)	
	DDcgB	DDaaB	DDcgB	DDaaB
St1 <sup>b</sup> vs St2	1449 ± 38	1434 ± 68	-66 ± 5	-72 ± 8
G4-C21 bp	4 ± 3	6 ± 2	-2 ± 1	-1 ± 1
A5-T20 bp	19 ± 2	11 ± 2	-2 ± 1	-1 ± 1

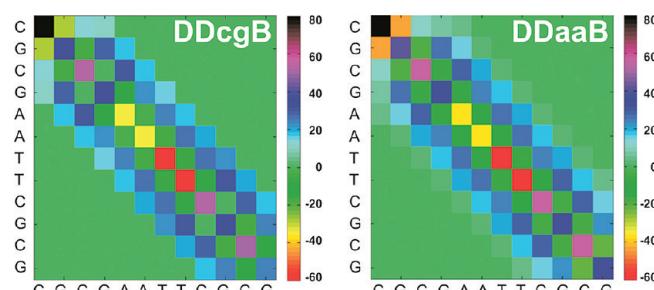
<sup>a</sup> The averages were calculated over 50 contiguous (DDaaB) or noncontiguous (DDcgB) nanoseconds. <sup>b</sup> St1 stands for strand 1 and St2 for strand 2.

that observed in the all-atom MD simulation.<sup>50</sup> The slight separation between the Twist and Roll traces observed in Figure 4b may suggest a sequence-specific behavior. To shed light on this issue, an exhaustive and systematic study of the helical parameters for all the possible unique combinations of dinucleotides and tetranucleotides (for a total of 146 possible combinations) should be carried out and compared against recent results coming from molecular dynamic simulations.<sup>50,51</sup> Such study is clearly beyond the scope of the present contribution.

**Energetic Comparison.** In order to further validate the back-mapping procedure and obtain further support on the equivalence between the conformational spaces sampled by the CG and atomistic models, we compared the nonbonded interaction terms of the energy. Calculations were done averaging the results in vacuum using in both cases the same force field (parm99) applied to the all-atom MD and back-mapped trajectories. Comparisons for the van der Waals (VdW) and electrostatic components of the interaction energy between (i) the two strands, (ii) the bases of a GC pair, and (iii) the bases of an AT pair are shown in Table 5.

In light of the correspondent values within the standard deviations, the electrostatic and VdW interactions between DNA strands were virtually the same for both simulations. The good correspondence between both nonbonded interaction terms points out that the conformational space sampled by the CG model was energetically compatible with the state-of-the-art molecular dynamics. Note that the electrostatic contributions in Table 5 are always positive numbers since we computed the Coulombic interaction between two negatively charged strands. When comparing selected GC or AT base pairs, some subtle differences in the averaged electrostatics arise between both approaches. In our back-mapped CG model, the GC base pairs are slightly more stable, whereas the AT base pairs showed an opposite trend. Aimed at acquiring a more global picture, we looked at the electrostatic interactions per residue. For this task, we computed a 12 × 12 electrostatic interactions matrix. The results are presented as an interaction map in Figure 5. A very good correlation between both maps can be observed, providing further support for the compatibility between both approaches.

**DNA Melting.** Experimentally, the melting temperature ( $T_0$ ) can be defined for an ensemble of double-stranded DNA molecules as the temperature at which half of the population is in the double-helical state and half in “random-coil” states. This type of definition, which is a good approximation for short DNA sequences, matches with the assumption that



**Figure 5.** Color map of the averaged electrostatic interaction between the 12 nucleotides within the same strand. Comparison between the back-mapped coarse grained (DDcgB) and the all-atom (DDaaB) simulations. The color scale ranges from -60 to +80 kcal/mol, which are the lower and upper boundary values in the all-atom simulation. It must be noticed that these values were obtained from an effective force field and must not be taken as real energies. The average was calculated over 50 contiguous (DDaaB) or noncontiguous (DDcgB) nanoseconds.

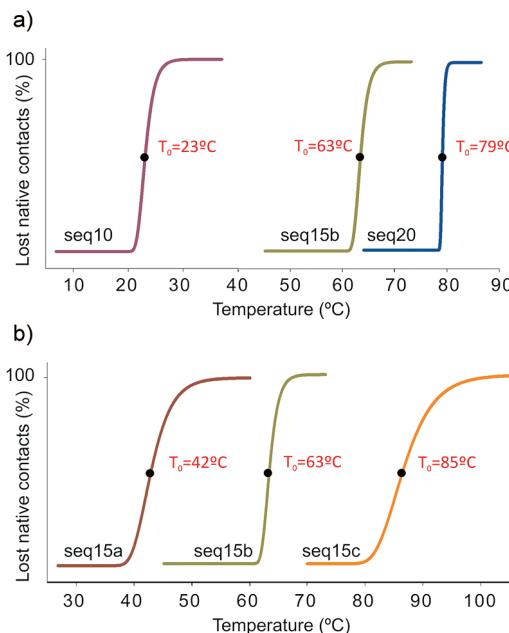
melting occurs in a two-state transition. The melting temperature is highly dependent on the length of the double-stranded DNA. Furthermore, because GC base-pairing is generally stronger than AT base-pairing, the amount of guanine and cytosine (called the “GC content”) can be estimated by measuring the temperature at which DNA melts.  $T_0$  also depends on the salt concentration or ionic strength of the surrounding medium, as a higher electrostatic screening reduces the mutual repulsion between the negatively charged backbones of each strand in the macromolecule. In other words,  $T_0$  can be used as an indirect measurement of the thermodynamic stability of a double-stranded DNA filament. In terms of the modeling, a good reproduction of the melting process may be indicative of a well-balanced energetic representation of the molecule under study.

To analyze the energetic features of the CG model, we followed the melting process of five sequences of different lengths, varying also the GC content and the ionic strength according to the Debye–Hückel screening parameter  $\kappa$ .<sup>52</sup> Our results were compared with recent experimental determinations for the same DNA sequences under nearly the same conditions.<sup>48</sup> No back-mapping was performed, as the fraction of native contacts can be measured directly from the CG trajectories.

We studied the length and GC-content dependence of the melting behavior for double-stranded DNA in implicit solvation. Melting temperatures were obtained from single simulations of double-stranded DNA where the temperature was raised in discrete steps of 20° to determine the melting point.

At first glance, good qualitative agreement can be found. As expected, increasing the base pairs number produced a higher  $T_0$  (Figure 6a). Similarly, a higher GC content shifts the  $T_0$  to higher temperatures (Figure 6b). However, in light of standard deviations in the temperature measurement (Table 6), the results could be considered rather qualitative.

There was no variation in  $T_0$  for seq15b at 0.07, 0.12, and 0.22 salt concentrations, for which the calculated melting point was always 63 °C (see Table 6). The only significantly



**Figure 6.** Fitted melting curves. (a) Sequences containing 10 (seq10), 15 (seq15b), and 20 (seq20) bp and 50–53% GC content. (b) Sequences with 15 bp for which the GC content is 20% (seq15a), 53% (seq15b), and 80% (seq15c), respectively. The inflection points (see eq 2) that determine the melting temperatures are indicated with black dots. Notice that the melting curves were obtained after a fitting procedure (see Methods). The numeric values along with the corresponding standard deviations are displayed in Table 6.

different  $T_0$  was obtained at a 1.0 M salt concentration. This is probably due to the way in which the salt effects are incorporated into the GB model. In practice, the linearized Debye–Hückel approximation gives salt effects that are somewhat larger than those predicted by more accurate methods.<sup>52</sup> Saturation of salt effects takes place near 1.0 M, and the best fit with more accurate Poisson–Boltzmann estimations occurs for values from 0.1 to 0.4 M.<sup>52</sup> Previous MD simulations of nucleic acid structures carried out with either a 0.1 or 0.2 M salt concentration showed almost identical results.<sup>53</sup> Recent work describing the melting reaction in DNA hexamers using the same force field (parm99 with the Perez and co-workers modification<sup>39</sup>) and more accurate all-atom simulations for sampling of the free energy landscape also gave only qualitative results.<sup>54</sup>

The aim of this last set of simulations discussed was to test the qualitative dependence of the melting point upon variations of different factors. A precise determination of

the melting temperature would need a better sampling such as, for instance, that performed by Knotts and co-workers.<sup>6</sup> They used replica exchange methods to achieve a more quantitative determination. We decided to not perform this kind of calculation, as there is a rather large arbitrariness in the molecular level definition of the melting point. For instance, a small variation (even of tenths of an angstrom) in the cutoff criteria for a native contact between two interacting bases can significantly shift the position of the melting points.

A clear advantage of using MD simulations is that the dynamic behavior of the melting process can be followed on the molecular scale. Thus, sequence- and location-dependent initiation and propagation of the steps that leads to DNA denaturation can be analyzed in detail. In all the sequences studied here, the melting of the helix started from the termini and proceeded toward the center (as an example, the movie for seq15b at 0.12 M is provided in the Supporting Information). This suggests that the loss of internal Watson–Crick interactions has a high-energy cost if the terminal base pairs are still formed as observed in other all-atom simulation work,<sup>54</sup> making internal fraying less frequent.

**The A to B Transition.** A celebrated result of effective force fields was the capability to reproduce complex conformational changes such as the A to B transition in duplex DNA.<sup>55,56</sup> Therefore, we faced the challenge of reproducing with our CG model the transition from the canonical A to B form, which is the physiologically more stable conformation of double-stranded DNA.

We prepared the same Drew–Dickerson dodecamer studied in the previous section but in the canonical A-form. To follow the A→B transition along the simulation, we calculated the RMSD of all the superatoms with respect to the corresponding atoms in the canonical B-form (see mapping scheme in Figure 1) and the two experimental structures. The results for 5  $\mu\text{s}$  of simulation are shown in Figure 7. The conformational transition took place progressively in a relatively long time window, arriving at final state after nearly 1.2  $\mu\text{s}$  (Figure 7a).

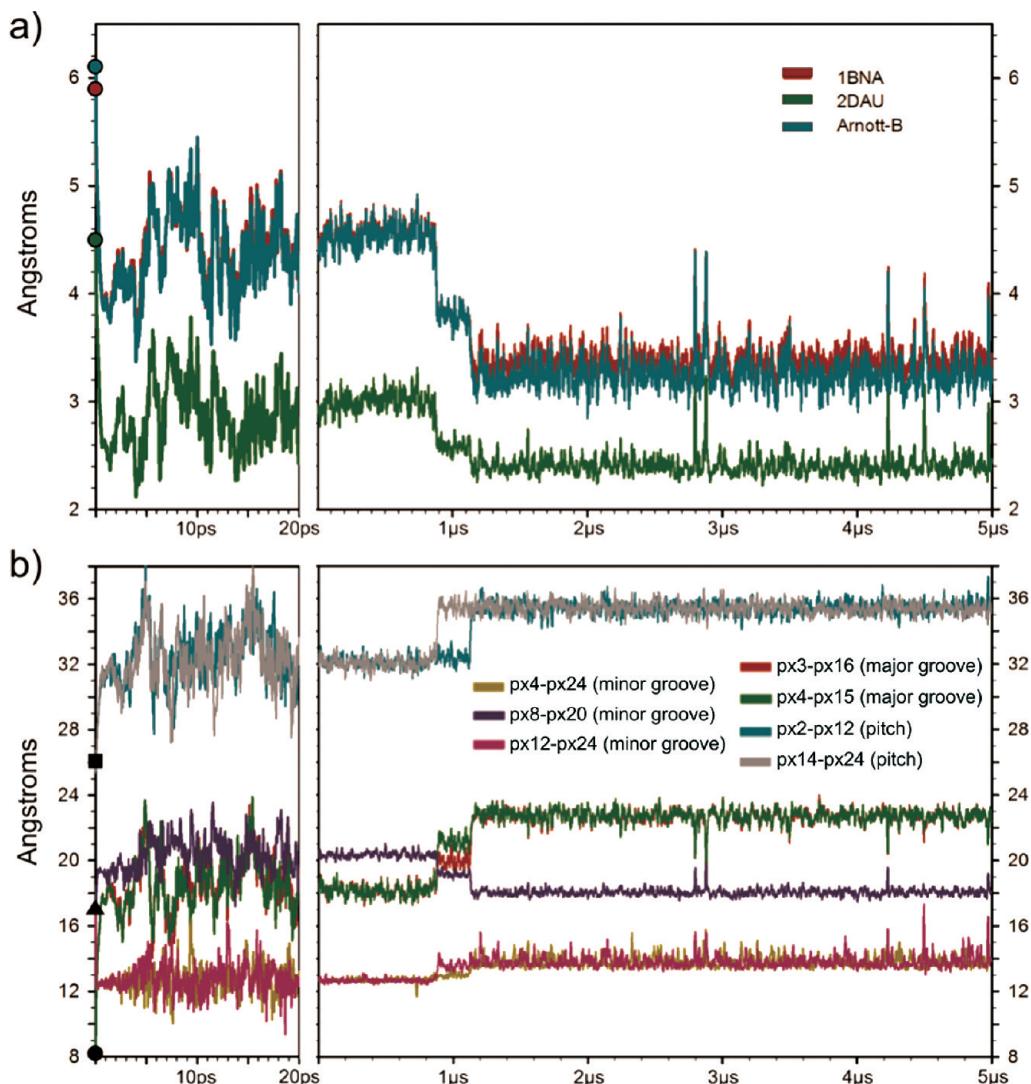
The final RMSD value reached after the transition was 3.3 Å with respect to the canonical B-form, e.g., a value comparable with the deviations obtained from atomistic simulations of duplex B-DNA using the generalized Born approximation.<sup>37</sup>

To reach the final B-form structure (between 1.2 and 5  $\mu\text{s}$ ), the conformational transition occurred in three steps:

**Table 6.** Reference Names and DNA Sequences Used in the Melting Experiments for which the GC Content and Salt Concentrations Are Indicated

reference name	DNA sequence (5'-3')	GC content (%)	salt concentration (M)	$T_0$ exptl <sup>a</sup> (°C)	$T_0$ calcd (°C)	st. dev.
seq10	ATCGTCTGGAA	50	0.12	37.4	23	25
seq15a	TACTAACATTAACATA	20	0.12	40.4	42	20
seq15b	GCAGTGGATGTGAGA	53	0.07 0.12 0.22 1.00	51.2 54.8 58.0 63.3	63 63 63 100	22 22 22 26
seq15c	GCGTCGGTCGGGCT	80	0.12	67.7	85	25
seq20	AGCTGCAGTGGATGTGAGAA	50	0.12	63.5	79	19

<sup>a</sup> Taken from ref 48.



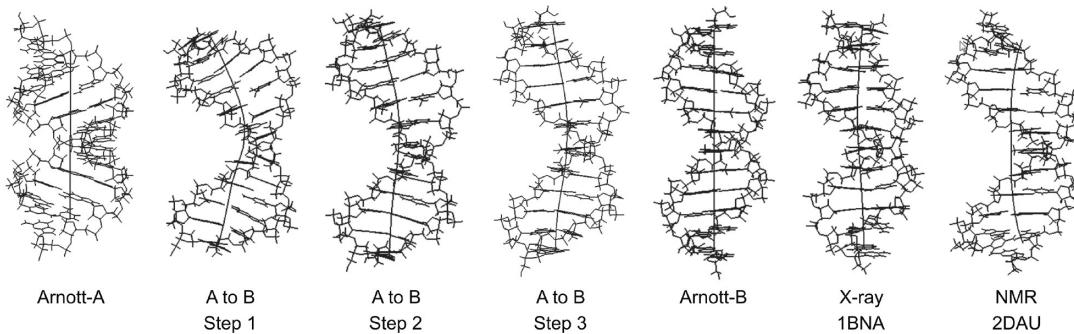
**Figure 7.** Time evolution of the A to B conformational transition. (a) RMSD using as a reference the canonical B-form (Arnott-B, blue line) and the X-ray and NMR structures (1BNA, dark red line, and 2DAU, green line, respectively). Colored dots indicate the RMSD of the initial conformer with respect to the reference structures. (b) Time evolution of selected distances (pitch, minor and major grooves) during simulation (color codes are indicated in the picture). Black squares, triangles, and circles indicate the starting values for pitch and minor and major grooves, respectively. In both cases, the data shown in the left panels correspond to instantaneous values, while data presented in the right panels correspond to a running average every 200 frames.

(i) In the first few picoseconds (left panel in Figure 7a), the initial structure (canonical A-form) underwent an abrupt conformational change that mainly affected the width of the major groove and, in a second degree, the overall pitch (see Figure 7b). On average, the major groove went from 8 to 18 Å and the pitch from 26 to 32 Å. These changes gave rise to a first cluster of structures 2.6 Å apart from the canonical A-form that remained stable during the first ~900 ns (step 1 in Figure 8). Using the generalized Born model, Tsui and Case<sup>53</sup> showed the convergence from an A-form DNA to a cluster of structures near the B-form within 20 ps of simulation. The quick transition was characterized by the rapid increase of the major groove and the end-to-end length (pitch). The same behavior was observed in the first 20 ps of the CG simulation (Figure 7b). Obviating that the DNA sequence is not strictly the same, visual inspection of the final structure obtained by Tsui and Case<sup>53</sup> after the transition looks very similar to the first cluster of structures obtained

in the first picosecond of our CG model (compare the second structure in Figure 8 with Figure 9 in ref 53).

(ii) The following ~300 ns were characterized by a second cluster of structures 3.3 Å apart from the initial structure (first shoulder in Figure 7a). As shown in Figure 7b, the major groove continued to increase from 18 to 21 Å. This movement was followed by a decrease in the wideness of the minor groove measured in the central part of the sequence (from residues 8 and 20, dark blue line). In this case, the pitch underwent an asymmetric transformation to first rearrange the 3'-5' strand; subsequently the 5'-3' strand changed its value from 32 to 35 Å (a value very near the 34 Å of the canonical B-form).

(iii) Finally, between 1.2 μs and the end of the simulation, a last cluster of conformers 3.0 Å apart from the reference structure could be found. To reach this last state, the pitch in the 5'-3' strand went to a final value of 35 Å. The major groove experienced a subsequent increase accompanied by



**Figure 8.** Comparison between back-mapped snapshots and atomistic structures. The conformers labeled steps 1–3 correspond to back-mapped representative snapshots from the conformational A to B transition: steps 1 (0–900 ns), 2 (900–1200 ns), and 3 (1.2–5.0  $\mu$ s). The DNA axis was calculated with the Curves program.<sup>44</sup>

a  $\sim$ 1 Å narrowing in the minor groove. Note that, along the 5  $\mu$ s of simulation, the minor groove measured in the extremity of the sequence (between residues 4–24 and 12–16) only underwent slight changes.

In short, the A  $\rightarrow$  B transition can be characterized by global changes in the major structural determinants of double-helical DNA (pitch and groove measurement) in a way that reminds the motion of a “crankshaft”. Worth notice is the presence of some peaks in the RMSD after 2  $\mu$ s of simulation. These correspond to little shifts between the two strands in the AT track that produce transient changes in the minor and major grooves. This behavior was only observed in the central tract and can be associated with breathing movements in the double helix (see next section).

As shown in Figure 8, the conformational changes seem to begin in the central part of the double helix and propagate to the ends, in the same way reported by Cheatham and Kollman in the first simulation on the A to B transition of DNA using all-atom simulations in explicit solvent.<sup>53,55</sup>

The comparison of the A to B transition with the work of Tsui and Case<sup>53</sup> appears to be relevant in the context of the actual time scale sampled by our CG scheme. This is always a complicated issue when dealing with CG simulations, as it is expected that the reduction of degrees of freedom translates to a flattening of the conformational space. The putative correspondence between our work and that of Tsui and Case seems to suggest some equivalence between both simulation schemes. However, the correspondence in the conformational transition may be an artifact of the model that is parametrized to reproduce the B-DNA. To further explore this issue, we sought to test our model against experimental data for which characteristic times ranging from picoseconds to hundreds of microseconds have been reported.

**DNA Breathing Dynamics.** The microsecond time scale for the full A to B transition begs the question of the correspondence between the real and simulated times. Some insights about this issue can be obtained from a comparison with published simulations on the microsecond time scale. Along the CG simulations of the Drew–Dickerson dodecamer, some transient base pair opening events occurred during the trajectory, especially at the AT pairs. The average lifetime of an open base pair is typically on the order of few picoseconds, but some opening events last for hundreds of picoseconds. These results are in very good agreement with

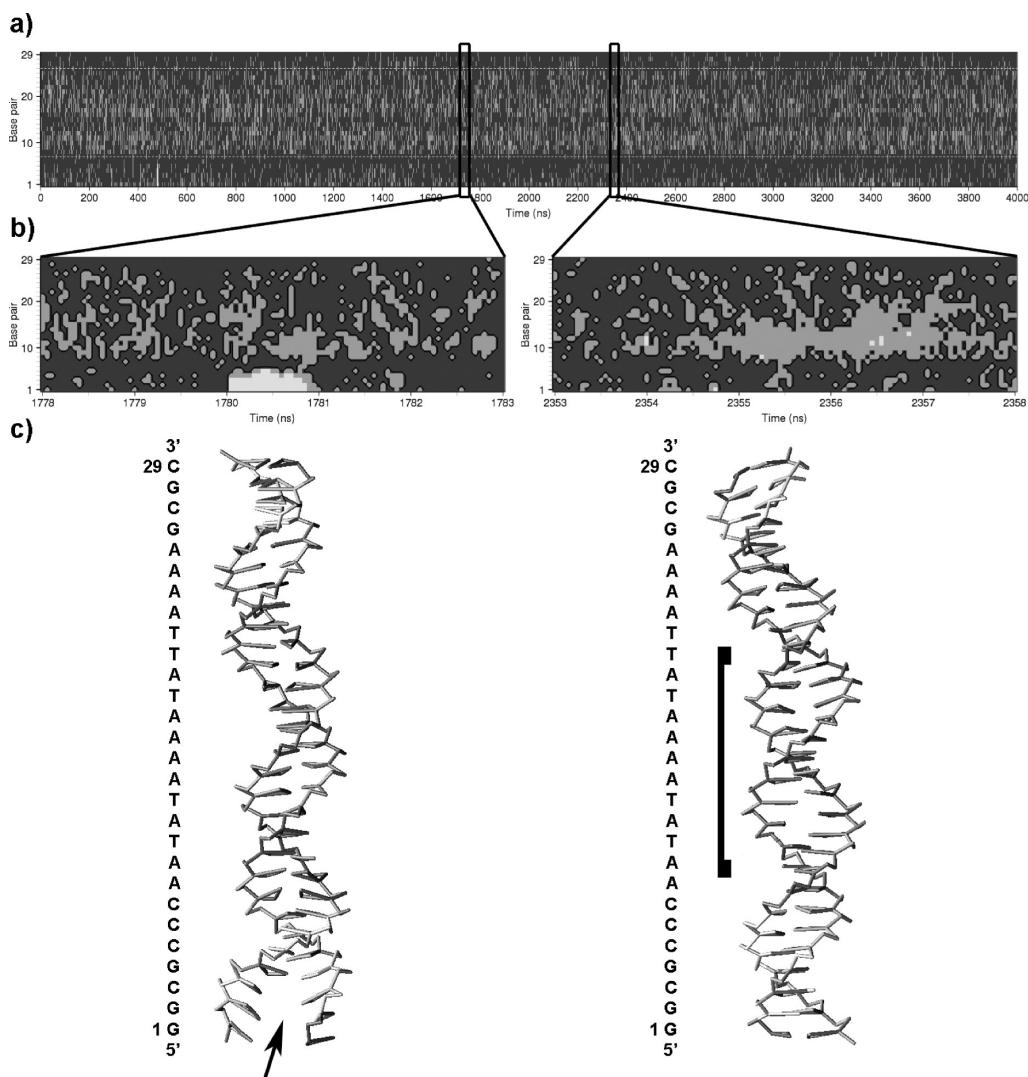
the work of Perez and co-workers,<sup>57</sup> who performed the atomistic simulation of the Drew–Dickerson over 1.2  $\mu$ s.

Aimed at directly comparing our model with well established experimental results and acquiring a more global perspective, we sought to perform the simulation of a 29-bp-long double-stranded DNA trying to mimic the laboratory conditions.<sup>49</sup> Base pair opening/closing dynamics have been reported for this kind of system on time scales ranging from picoseconds to nanoseconds<sup>58</sup> to hundreds of microseconds.<sup>49</sup> This would allow us to set the time frame of our simulations within a time scale window of near 8 orders of magnitude, covering (i) end-fraying, (ii) breathing, i.e., opening/closing of internal base pairs, and (iii) bubble formation, i.e., temporary opening of internal base pairs implying a partial loss of the double-helical structure.

Following the criterion to define an open state (see the Methods), we calculated the instantaneous state of each base pair (open/close) for each frame of the simulation and the time and sequence extension of those events. As was expected, significantly fewer open states were found in the GC clamp region compared to the AT domain (Figure 9a). Fraying events typically involved few base pairs (typically one or two, Figure 9b) that relax reaching the closed state in dozens to hundreds of picoseconds. This effect is compatible with X-ray,<sup>59</sup> NMR,<sup>60,61</sup> and computer<sup>31,61</sup> studies indicating that fraying is largely confined to the last two base pairs. The CG model also agrees with time-resolved Stokes shifts spectroscopy measurements that restrict the base-opening time to the range of dozens of picoseconds to a few nanoseconds.<sup>62</sup> Nevertheless, during the 4  $\mu$ s of simulation, we found two events where the end-fraying spread even up to the sixth base pair (Figure 9b,c).

In the AT domain, a nearly continuous breathing dynamic was found along the simulation (Figure 9a), registering several opening/closing events. These events remained in the open state on the nanosecond time scale (see Figure 9b right). The global deformation and the time scale are well comparable with the NMR imino-proton exchange measurements.<sup>58</sup> In this technique, only slight opening of the base pairs, as those observed in the CG model, would be sufficient for the reaction to occur.

Notably, simultaneous opening/closing events with extensions from 2 to 10 consecutive base pairs were frequently observed (Figure 9b). Although with a much shorter time



**Figure 9.** Breathing dynamics of the 29-bp-long double-stranded DNA. Base pairs (y axis) are plotted versus time (x axis) in nanoseconds. (a) Overview of the breathing along the trajectory. Dark gray color represents closed state base pairs (inter base distance lower than 4 Å). Open states were divided into two ranges: from 4 to 6 Å (light gray) and more than 6 Å (white). White dashed lines delimit the AT breathing domain.<sup>49</sup> (b) Five nanosecond closeups of the trajectory. (c) Representative structures of the end-fraying at the GC clamp (left) and AT breathing domain (right). Fraying and breathing are evidenced with an arrow and square bracket, respectively.

range, these results agree with multiexponential kinetics inferred from fluorescence relaxation times in an analogous molecular system for which opening/closing times of 20–100  $\mu$ s were reported.<sup>49</sup> It is worth note that these data were obtained from fluorescence quenching experiments, which require a significant distortion in the double-helical structure (bubble formation) in order to be detectable. Such large deformations were never observed along our simulations.

The correspondence with previous theoretical work<sup>57</sup> and NMR studies<sup>58</sup> suggests that the time scale sampled by our model may roughly match the real one. Should this be true, a simulation time on the order of milliseconds would be needed to properly sample the  $\sim$ 100  $\mu$ s process of bubble formation reported for 29-bp-long double-stranded DNA.<sup>49</sup> Alternatively, the absence of large deformations in our CG simulations could be related to the relative stiffness in the torsional parameters used. A larger number of simulations

on different systems and comparison against experimental data are needed to further clarify this point.

## Conclusions

We presented herein a nontopological CG model for MD simulations of DNA with explicit electrostatics that offers the possibility to fully recover the atomistic information. Back-mapped CG trajectories gave geometries with maximum deviations of a few angstroms from experimental values, which may be compatible with all-atom simulations offering a considerable speedup. Coarse-grained simulations were carried out in a single node with eight Intel Xeon 2.66 GHz cores at a rate of  $\sim$ 100  $\mu$ s/superatom/day. At this rate, we performed 1  $\mu$ s of the coarse-grained simulation using the Drew–Dickerson system in  $\sim$ 1.5 days. Around 850 days would be needed to run 1  $\mu$ s of the all-atom simulation described herein. Globally, a speedup by a factor of nearly

600 is granted using the CG model. An advantage of the present contribution is that many of the published CG simulation schemes are implemented in *ad hoc* codes or require tailor-made modifications of standard simulation packages, which are often difficult to access and/or operate for the general public. A notable exception of this is the MARTINI force field.<sup>63</sup> The evaluation of the interactions using a classical Hamiltonian allows for a straightforward porting to any other publicly available MD simulation package (topologies and parameters files in AMBER format are available from the authors upon request).

Although the sampling time remains a not completely solved issue, this kind of implementation may open new alternatives to the study of dynamic properties of nucleic acids at longer time scales and for larger systems.

Finally, we would like to stress the fact that the results showed here cover only applications where DNA exists near its B-form. Clearly, Hoogsteen and sugar-edge pairs are out of reach for the present model. This begs the question of whether noncanonical structural motifs can be also well described (structure of telomeres, circular DNA, etc.). This is particularly relevant for the case of RNA where a multiplicity of structural motifs is present (bulges, wobbles, hairpins, and internal loops, etc.). Work is currently ongoing in our group to expand the description to these more challenging cases.

**Acknowledgment.** This work was supported by ANII (Agencia Nacional de Investigación e Innovación), Programa de Apoyo Sectorial a la Estrategia Nacional de Innovación—INNOVA URUGUAY (Agreement n8 DCI - ALA/2007/19.040 between Uruguay and the European Commission), and Grant FCE\_60-2007. M.R.M. is a beneficiary of the National Fellowship System of ANII.

**Supporting Information Available:** Fortran 90 implementation of the homemade algorithm needed for the reconstruction of the CG trajectories is provided. A pseudo-code version explaining the homemade algorithm and two figures illustrating its accuracy (before and after the energy minimization) are also provided along with a movie of the melting process for seq15b at a 0.12 M salt concentration. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Klein, M. L.; Shinoda, W. Large-scale molecular dynamics simulations of self-assembling systems. *Science* **2008**, *321* (5890), 798–800.
- (2) Voth, G. A. *Coarse-Graining of Condensed Phase and Biomolecular Systems*, 1st ed.; Taylor & Francis Group: New York, 2009; pp 1–455.
- (3) Treptow, W.; Marrink, S. J.; Tarek, M. Gating motions in voltage-gated potassium channels revealed by coarse-grained molecular dynamics simulations. *J. Phys. Chem. B* **2008**, *112* (11), 3277–3282.
- (4) Ollila, O. H.; Risselada, H. J.; Louhivuori, M.; Lindahl, E.; Vattulainen, I.; Marrink, S. J. 3D pressure field in lipid membranes and membrane-protein complexes. *Phys. Rev. Lett.* **2009**, *102* (7), 078101.
- (5) Tepper, H. L.; Voth, G. A. A coarse-grained model for double-helix molecules in solution: spontaneous helix formation and equilibrium properties. *J. Chem. Phys.* **2005**, *122* (12), 124906.
- (6) Knotts, T. A.; Rathore, N.; Schwartz, D. C.; de Pablo, J. J. A coarse grain model for DNA. *J. Chem. Phys.* **2007**, *126* (8), 084901.
- (7) Chen, J.-S.; Teng, H.; Nakano, A. Wavelet-based multi-scale coarse graining approach for DNA molecules. *Finite Elem. Anal. Des.* **2007**, *43*, 346–360.
- (8) Becker, N. B.; Everaers, R. From rigid base pairs to semi-flexible polymers: coarse-graining DNA. *Phys. Rev. E.: Stat. Nonlin. Soft. Matter Phys.* **2007**, *76* (2 Pt 1), 021923.
- (9) Zhang, F.; Collins, M. A. Model simulations of DNA dynamics. *Phys. Rev. E* **1995**, *52* (4), 4217–4224.
- (10) Mergell, B.; Ejtehadi, M. R.; Everaers, R. Modeling DNA structure, elasticity, and deformations at the base-pair level. *Phys. Rev. E* **2003**, *68*, 021911.
- (11) Poulain, P.; Saladin, A.; Hartmann, B.; Prévost, C. Insights on Protein-DNA Recognition by Coarse Grain Modelling. *J. Comput. Chem.* **2008**, *29*, 2582–2592.
- (12) Hyeon, C.; Thirumalai, D. Mechanical unfolding of RNA hairpins. *Proc. Natl. Acad. Sci. U. S. A* **2005**, *102* (19), 6789–6794.
- (13) Hyeon, C.; Thirumalai, D. Forced-unfolding and force-quench refolding of RNA hairpins. *Biophys. J.* **2006**, *90* (10), 3410–3427.
- (14) Zhang, D.; Konecny, R.; Baker, N. A.; McCammon, J. A. Electrostatic interaction between RNA and protein capsid in cowpea chlorotic mottle virus simulated by a coarse-grain RNA model and a Monte Carlo approach. *Biopolymers* **2004**, *75* (4), 325–337.
- (15) Forrey, C.; Muthukumar, M. Langevin Dynamics Simulations of Genome Packing in Bacteriophage. *Biophys. J.* **2006**, *91*, 25–41.
- (16) Voltz, K.; Trylska, J.; Tozzini, V.; Kurkal-Siebert, V.; Langowski, J.; Smith, J. Coarse-grained force field for the nucleosome from self-consistent multiscaling. *J. Comput. Chem.* **2008**, *29* (9), 1429–1439.
- (17) Hyeon, C.; Dima, R. I.; Thirumalai, D. Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. *Structure* **2006**, *14* (11), 1633–1645.
- (18) Tan, R. K. Z.; Petrov, A. S.; Harvey, S. C. YUP: A Molecular Simulation Program for Coarse-Grained and Multiscaled Models. *J. Chem. Theory Comput.* **2006**, *2*, 529–540.
- (19) Korolev, N.; Lyubartsev, A. P.; Nordenskiöld, L. Computer modeling demonstrates that electrostatic attraction of nucleosomal DNA is mediated by histone tails. *Biophys. J.* **2006**, *90* (12), 4305–4316.
- (20) Wocjan, T.; Klenin, K.; Langowski, J. Brownian Dynamics Simulation of DNA Unrolling from the Nucleosome. *J. Phys. Chem. B* **2009**, *113* (9), 2639–2646.
- (21) Langowski, J. Polymer chain models of DNA and chromatin. *Eur. Phys. J. E. Soft. Matter* **2006**, *19* (3), 241–249.
- (22) Langowski, J.; Heermann, D. W. Computational modeling of the chromatin fiber. *Semin. Cell Dev. Biol.* **2007**, *18* (5), 659–667.
- (23) Arnott, S.; Campbell-Smith, P. J.; Chandrasekaran, R. *Handbook of Biochemistry and Molecular Biology*, 3rd Nucleic Acids ed.; CRC Press: Cleveland, OH, 1976; Vol. II, pp 411–422.

- (24) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.* **1996**, *100*, 19824–19839.
- (25) AMBER 10; University of California: San Francisco, CA, 2008.
- (26) Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E. Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **1981**, *78* (4), 2179–2183.
- (27) Dickerson, R. E.; Drew, H. R. Structure of a B-DNA dodecamer. II. Influence of base sequence on helix structure. *J. Mol. Biol.* **1981**, *149* (4), 761–786.
- (28) Drew, H. R.; Dickerson, R. E. Structure of a B-DNA dodecamer. III. Geometry of hydration. *J. Mol. Biol.* **1981**, *151* (3), 535–556.
- (29) McConnell, K. J.; Beveridge, D. L. DNA structure: what's in charge. *J. Mol. Biol.* **2000**, *304* (5), 803–820.
- (30) Phan, A. T.; Leroy, J. L.; Gueron, M. Determination of the residence time of water molecules hydrating B'-DNA and B-DNA, by one-dimensional zero-enhancement nuclear Overhauser effect spectroscopy. *J. Mol. Biol.* **1999**, *286* (2), 505–519.
- (31) Young, M. A.; Ravishanker, D.; Beveridge, D. L. A 5-ns Molecular Dynamics Trajectory for B-DNA: Analysis of Structure, Motions, and Solvation. *Biophys. J.* **1997**, *73*, 2313–2336.
- (32) Denisov, V. P.; Carlstrom, G.; Venu, K.; Halle, B. Kinetics of DNA hydration. *J. Mol. Biol.* **1997**, *268* (1), 118–136.
- (33) Duan, Y.; Wilkosz, P.; Crowley, M.; Rosenberg, J. M. Molecular dynamics simulation study of DNA dodecamer d(CGCGAATTCGCG) in solution: conformation and hydration. *J. Mol. Biol.* **1997**, *272* (4), 553–572.
- (34) Pastor, R. W.; Brooks, B. R.; Szabo, A. An analysis of the accuracy of Langevin and molecular dynamics algorithms. *Mol. Phys.* **1988**, *65*, 1409–1419.
- (35) Wu, X.; Brooks, B. R. Self-guided Langevin dynamics simulation method. *Chem. Phys. Lett.* **2003**, *381*, 512–518.
- (36) Izaguirre, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D. Langevin stabilization of molecular dynamics. *J. Chem. Phys.* **2001**, *114*, 2090–2098.
- (37) Cheatham, T. E., III; Case, D. A. *Computational Studies of RNA and DNA*; Springer: Dordrecht, The Netherlands, 2006; pp 45–71.
- (38) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules. *J. Comput. Chem.* **2000**, *21*, 1049–1074.
- (39) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., III; Laughton, C. A.; Orozco, M. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **2007**, *92* (11), 3817–3829.
- (40) Jorgensen, W. L. Transferable Intermolecular Potential Functions for Water, Alcohols, and Ethers. Application to Liquid Water. *J. Am. Chem. Soc.* **1981**, *103*, 335–340.
- (41) Darden, T. A.; York, D.; Pedersen, L. Particle mesh Ewald: An  $N \cdot \log(N)$  method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (42) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3691.
- (43) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (44) Lavery, R.; Sklenar, H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.* **1988**, *6* (1), 63–91.
- (45) Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E. Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **1981**, *78* (4), 2179–2183.
- (46) Denisov, A. Y.; Zamaratski, E. V.; Maltseva, T. V.; Sandstrom, A.; Bekiroglu, S.; Altman, K. H.; Egli, M.; Chattopadhyaya, J. The solution conformation of a carbocyclic analog of the Dickerson-Drew dodecamer: comparison with its own X-ray structure and that of the NMR structure of the native counterpart. *J. Biomol. Struct. Dyn.* **1998**, *16* (3), 547–568.
- (47) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.
- (48) Owczarzy, R.; You, Y.; Moreira, B. G.; Manthey, J. A.; Huang, L.; Behlke, M. A.; Walder, J. A. Effects of sodium ions on DNA duplex oligomers: improved predictions of melting temperatures. *Biochemistry* **2004**, *43* (12), 3537–3554.
- (49) Altan-Bonnet, G.; Libchaber, A.; Krichevsky, O. Bubble dynamics in double-stranded DNA. *Phys. Rev. Lett.* **2003**, *90* (13), 138101.
- (50) Lavery, R.; Zakrzewska, K.; Beveridge, D.; Bishop, T. C.; Case, D. A.; Cheatham, T., III; Dixit, S.; Jayaram, B.; Lankas, F.; Laughton, C.; Maddocks, J. H.; Michon, A.; Osman, R.; Orozco, M.; Perez, A.; Singh, T.; Spackova, N.; Sponer, J. 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.* **2010**, *38* (1), 299–313.
- (51) Dixit, S. B.; Beveridge, D. L.; Case, D. A.; Cheatham, T. E., III; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Sklenar, H.; Thayer, K. M.; Varnai, P. 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.* **2005**, *89* (6), 3721–3740.
- (52) Srinivasan, J.; Trevathan, M. W.; Beroza, P.; Case, D. A. Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. *Theor. Chem. Acc.* **1999**, *(101)*, 426–434.
- (53) Tsui, V.; Case, D. A. Molecular Dynamics Simulations of Nucleic Acids with a Generalized Born Solvation Model. *J. Am. Chem. Soc.* **2000**, *122*, 2489–2498.
- (54) Piana, S. Atomistic simulation of the DNA helix-coil transition. *J. Phys. Chem. A* **2007**, *111* (49), 12349–12354.
- (55) Cheatham, T. E., III; Kollman, P. A. Observation of the A-DNA to B-DNA transition during unrestrained molecular dynamics in aqueous solution. *J. Mol. Biol.* **1996**, *259* (3), 434–444.
- (56) Soliva, R.; Luque, F. J.; Alhambra, C.; Orozco, M. Role of sugar re-puckering in the transition of A and B forms of DNA

- in solution. A molecular dynamics study. *J. Biomol. Struct. Dyn.* **1999**, *17* (1), 89–99.
- (57) Pérez, A.; Luque, F. J.; Orozco, M. Dynamics of B-DNA on the Microsecond Time Scale. *J. Am. Chem. Soc.* **2007**, *129*, 14739–14745.
- (58) Gueron, M.; Leroy, J. L. Studies of base pair kinetics by NMR measurement of proton exchange. *Methods Enzymol.* **1995**, *261*, 383–413.
- (59) Holbrook, S. R.; Kim, S. H. Local mobility of nucleic acids as determined from crystallographic data. I. RNA and B form DNA. *J. Mol. Biol.* **1984**, *173* (3), 361–388.
- (60) Fujimoto, B. S.; Willie, S. T.; Reid, B. R.; Schurr, J. M. Position-dependent internal motions and effective correlation times for magnetization transfer in DNA. *J. Magn Reson. B* **1995**, *106* (1), 64–67.
- (61) Kojima, C.; Ono, A.; Kainosh, M.; James, T. L. DNA duplex dynamics: NMR relaxation studies of a decamer with uniformly <sup>13</sup>C-labeled purine nucleotides. *J. Magn. Reson.* **1998**, *135* (2), 310–333.
- (62) Andreatta, D.; Sen, S.; Pérez Lustres, J. L.; Kovalenko, A. E. N. P. M. C. J.; Coleman, R. S. B. M. A. Ultrafast Dynamics in DNA: “Fraying” at the End of the Helix. *J. Am. Chem. Soc.* **2006**, *128*, 6885–6892.
- (63) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **2007**, *111* (27), 7812–7824.

CT900653P