Ab initio determination of coarse-grained interactions in double-stranded DNA

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:41384025

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
**Ab initio** determination of coarse-grained interactions in double-stranded DNA

Chia Wei Hsu,1 Maria Fyta,1,2 Greg Lakatos,1,3 Simone Melchionna,3,4 and Efthimios Kaxiras1,3,a)

1Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA
2Department of Physics, Technical University of Munich, Garching 85748, Germany
3School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA
4IPCF-CNR, Istituto Processi Chimico-Fisici, Consiglio Nazionale delle Ricerche, Università La Sapienza, P.le A. Moro 2, 00185 Rome, Italy

(Received 12 December 2011; accepted 13 August 2012; published online 12 September 2012)

We derive the coarse-grained interactions between DNA nucleotides from *ab initio* total-energy calculations based on density functional theory (DFT). The interactions take into account base and sequence specificity, and are decomposed into physically distinct contributions that include hydrogen bonding, stacking interactions, backbone, and backbone-base interactions. The interaction energies of each contribution are calculated from DFT for a wide range of configurations and are fitted by simple analytical expressions for use in the coarse-grained model, which reduces each nucleotide into two sites. This model is not derived from experimental data, yet it successfully reproduces the stable B-DNA structure and gives good predictions for the persistence length. It may be used to realistically probe dynamics of DNA strands in various environments at the μs time scale and the μm length scale. © 2012 American Institute of Physics. [http://dx.doi.org/10.1063/1.4748105]

I. INTRODUCTION

Biological systems exhibit high degrees of complexity that are essential to the functions they perform. The DNA double helix is one such example: the properties of this macromolecule are directly influenced by its conformational variability as well as by environmental factors that include counterions, impurities, and temperature, as it performs a wide variety of vital cellular functions such as transcription and replication.1 A full account of such biological functions must rely on a realistic description of the physical processes that underlie them. Fine-grained calculations of DNA at the atomic level2,3 can provide this level of detailed description, but they are restricted to very small systems of order tens of base pairs and time scales of order ns, whereas most biological processes involve DNA behavior at the scale of more than a hundred base-pairs and take place at the μs time scale and beyond. To probe these biologically relevant processes, a realistic and efficient coarse-grained model of DNA is necessary. Examples of crucial functions under current investigation that could benefit from a coarse-grained model of DNA include the translocation of DNA through nanopores4,5 in the context of ultra-fast electronic sequencing, DNA capture in and ejection from nanoscale capsules/wells, and the study of the interplay between histones and DNA during mitosis.6

The main theoretical challenge in biological systems is to bridge the scales between the atomistic and the macroscopic without wasting computational resources on uninteresting behaviors, such as the internal dynamics within a base which practically never changes shape, or the motion of solvent molecules far from the biomolecule. Computational methods of this type have been used before, for example, in the study of protein dynamics.7 Multiscale simulations have also been successfully applied to the study of the electronic behavior and electron localization in stretched dry DNA.8 Our ultimate goal is to enable the study of variations in the DNA structure using multiscale approaches that do not sacrifice accuracy while achieving high efficiency. In the present work we focus on the first stage toward this goal, namely, the development of a coarse-grained model capable of accurately reproducing the structure of double-stranded DNA (ds-DNA) in solution, and simple enough to be efficiently combined with multiscale simulation techniques.

Many coarse-grained models of DNA have been proposed in the past few years.8–21 Most of them are constructed in a “top-down” fashion,8,9,11 where the interaction potentials are chosen to reproduce certain sets of experimental data. Bead-string models10 have been used to study diffusion and structural relaxation of single strands. Rigid base-pair models11 have been used to describe the elastic properties of DNA.22,23 The three-site-per-nucleotide model by Knotts and coworkers14 captures the salt-dependent melting of DNA, and has been extended to include solvent-induced attraction between DNA strands24 that helped to gain insights on hybridization25–27 and on certain sequence-dependent effects;28 this model has also been adapted to describe the mechanical denaturation of long DNA29 and to include explicit solvent molecules.30 Starr and coworkers proposed a simple model that captures the basics of hybridization,13 and this model has been used to study Holliday junctions31 and the self-assembly of DNA-linked nanoparticles.32 Ouldridge and coworkers proposed another model that is sequence independent, but can reproduce several structural, mechanical, and thermodynamic properties.
DNA.\textsuperscript{16,33} Another family of models starts from all-atom empirical force fields, and construct the coarse-grained model potentials from bottom-up.\textsuperscript{18–21} The model by Savelyev and coworkers\textsuperscript{18,34} was parametrized by matching moments of observables in the Hamiltonian. Other approaches to a bottom-up construction include minimizing difference between the all-atom and coarse-grained potentials,\textsuperscript{19} imposing molecular bonding geometry constraints,\textsuperscript{20} and inverting the Boltzmann function to get the coarse-grained potentials.\textsuperscript{21}

Each of these models has its regime of validity, depending on what experimental data were used for the model derivation. In this work, we develop a minimal model of ds-DNA that incorporates sequence specificity and has realistic mechanical robustness to bending and untwisting forces. We seek a model that is chemically accurate, yet not based on empirical observations. For these purposes, we take a bottom-up approach, and construct the potentials of the coarse-grained system directly from first-principles calculations. We divide the interaction potentials into independent parts that come from physically distinct contributions, and for each contribution we carry out \textit{ab initio} calculations to find the functional forms of the potentials. We impose no \textit{a priori} assumption on the functional forms of the potentials—these are determined based on results of the \textit{ab initio} calculations. The present form of the model has its limitations that we will discuss, but it is flexible for future improvements.

This paper is organized as follows: in Sec. II we describe the methodology employed for the first-principles calculations. In Sec. III we present the \textit{ab initio} data for each energy contribution and the analytical forms of the corresponding potentials. The implementation and performance of the coarse-grained model is described in Sec. IV. Finally, we present validations of the model in Sec. V, and conclude in Sec. VI.

II. METHODOLOGY OF \textit{AB INITIO} CALCULATIONS

We carry out the first-principles calculations using density functional theory (DFT).\textsuperscript{35} In our DFT calculations, we do not deal with environmental factors such as solvent molecules and ions, and the calculations are carried out at the ground state (zero temperature). The environment and temperature factors are added \textit{a posteriori} in the coarse-grained model through electric field screening and through Brownian dynamics. A more accurate approach will include temperature dependence in the coarse-graining procedure, but this is a challenging task and is not yet taken into account in the current work. Presence of the water molecules and ions may also affect the energetics, but their effects are not investigated in this work.

We use SIESTA,\textsuperscript{36} an electronic structure code based on atomic-like orbitals, to carry out the DFT calculations. This approach has been previously applied in similar studies of gas-phase DNA bases\textsuperscript{37} and successfully reproduces properties such as optical response in comparison to available experimental data.\textsuperscript{38} We use the Troullier-Martins scheme\textsuperscript{39} to obtain pseudopotentials to eliminate the core electrons from the calculation and to produce a smoother valence charge density. We use a basis of double-$\zeta$ polarized atomic orbitals for all the atoms involved (13 orbitals for C, N, O, and P; 5 orbitals for H). An auxiliary real space grid equivalent to a plane-wave cutoff of 100 Ry is used for the calculation of the Hartree and exchange-correlation energies. For geometry optimization, a structure is considered fully relaxed when the magnitude of force on every atom is smaller than 0.04 eV/Å.

We use the generalized gradient approximation (GGA) with the PBE exchange-correlation functional\textsuperscript{40} to describe the backbone-base and the inter-base-pair interactions, as this functional is known to describe covalent and hydrogen bonds well.\textsuperscript{41,42} Interactions within the backbone are treated with the local-density approximation (LDA),\textsuperscript{43} which is adequate for describing small radial and angular deviations of covalent bonds from their equilibrium values. The interaction between stacked base-pairs has a large contribution from long-ranged van der Waals (vdW) forces and exhibits an elaborate energy landscape that depends sensitively on the geometry.\textsuperscript{44,45} Local or semi-local exchange-correlation functionals cannot describe such long-range effects;\textsuperscript{46} they do not reproduce the $\sim r^{-6}$ behavior at large separation $r$ that is characteristic of vdW interactions, and usually underestimate the stacking distance between two planar structures. There exists empirical corrections that add the vdW effects to the energies obtained from DFT calculations,\textsuperscript{47} but we find that such correction still leads to underestimation of the stacking distance. Therefore, we employ a non-empirical long-range vdW density functional developed by Dion \textit{et al.}\textsuperscript{48} to carry out calculations for the stacked base-pairs interactions.

The calculated interaction energy between two constituents may be susceptible to the basis set superposition error (BSSE), which results in unphysical lowering of the interaction energy when the two constituents come close to each other. To correct for BSSE, we take the full counterpoise approach\textsuperscript{49} at each separation $r$, we optimize the dimer geometry and carry out four additional calculations (constituent A along and constituent B along, with and without ghost orbitals), and obtain the BSSE-corrected interaction energy as

\begin{equation}
E(r) = E_{AB}(r) - E_A^*(r) - E_B^*(r) + E_A(r) + E_B(r)
- E_A^* - E_B^*,
\end{equation}

where the subscript denotes the geometry: AB is the optimized dimer geometry, A and B are its constituent geometries, and A* and B* are the individually optimized constituent geometries; the prime indicates that ghost orbitals of the other constituent are used in the energy calculation.

III. CONSTRUCTION OF MODEL POTENTIALS

We coarse-grain each nucleotide into two interaction sites: one for the base and one for the sugar-phosphate backbone. The base site is identified with the position of the nitrogen atom (N1 for pyrimidines; N9 for purines) that connects the base to the sugar, and the backbone site is identified with the position of the sugar C1’ atom. This representation is illustrated in Fig. 1. This choice of coarse-grained site coordinates allows for unambiguous determination of bonding distance, bonding angles, etc., that enter the model variables. It also
facilitates direct comparison to experimental structures during validation of the model.

To derive the effective interaction between these coarse-grained sites, we follow earlier work\cite{hwa} and decompose the total interaction energy into contributions that have distinct physical meanings: the hydrogen bond between complementary bases $E_{hb}$, the stacking energy between neighboring base-pairs $E_{st}$, contributions from the backbone $E_{bb}$, and electrostatic interaction between the charged phosphate groups $E_{el}$:

$$E_{total} = E_{hb} + E_{st} + E_{bb} + E_{el}.$$  

(2)

These are physically distinct contributions, and in our model they are treated as independent and additive. Each contribution depends on several variables, and the effect of these variables are not taken as independent.

The choice of the coarse-grained sites and the decomposition of the interaction potentials has determined the structure of the model. The remaining construction of the model is to find explicit functional forms for each contribution, which we address in the following.

A. Hydrogen bonding

We consider first the interaction between two complementary bases: adenine-thymine (AT) or guanine-cytosine (GC), which comes from hydrogen bonds. We consider its dependence on the base-to-base separation and on the relative angles between the planes of the two bases. The angular dependence keeps the two bases coplanar and maintains the correct base-pair geometry.

We calculate the interaction energy as a function of the distance $r_{hb}$ between the pyrimidine N1 atom and the purine N9 atom (see Fig. 1), by starting from the energy-minimized geometry and varying $r_{hb}$ by translating the two bases parallel to the direction of the hydrogen bonds. At each $r_{hb}$ value, we optimize the geometry while fixing the four atoms that correspond to the coarse-grained sites (N1 and H1 of the pyrimidine, and N9 and H9 of the purine) to preserve $r_{hb}$ and the relative angles. At small $r_{hb}$ values, the atoms are also constrained on the base-pair plane to prevent the two bases from rotating out of plane. The effect of flipping angles and of non-planarity will be examined separately.

Figure 2 shows the calculated interaction energy, which can be described by the universal binding energy relation (UBER)\cite{hu}:

$$E_{hb}(r_{hb}) = E_0(1 + a^*e^{-a^*}), \quad a^* = (r_{hb} - r_0)/l,$$  

(3)

with its minimum at $r_{hb} = r_0$, $E_{hb}(r_{hb}) = E_{0}$, $E_{0}$, $r_0$, and $l$ are given in Table I. The energy at the minimum is $-0.64$ eV for AT and $-1.17$ eV for GC, in agreement with the values $-0.60$ eV and $-1.17$ eV, respectively, obtained in previous calculations of the DNA hydrogen bonds.\cite{fron} We do not parametrize the base-to-base interaction between mismatched base-pairs, but doing so will be a straightforward extension of the current work.

Next we examine the effect of non-planarity of the bases: when the two planes of the bases are not aligned, the hydrogen bond weakens. We measure non-planarity by the dihedral angle $\theta_d$ between the planes of the two bases. Starting from the energy-minimized geometry ($\theta_d = 0$), we vary $\theta_d$ by rotating the two bases in opposite directions around the line from the pyrimidine N1 atom to the purine N9 atom. To keep $\theta_d$ fixed, we optimize the geometry while holding the 6-fold ring of the

![FIG. 1. Schematic of an adenine-thymine (AT) base-pair showing the geometric variables in the hydrogen bond potential in (a) the atomic structure calculations and (b) the coarse-grained model. The distance $r_{hb}$, the two flip-angles $\phi^{(i)}$, and the dihedral angle $\theta_d$ (angle between the two base planes, not shown) are used to describe interaction between the two bases. The vectors $r_{CC}$ and $r_{CN}^{(i)}$ are used to define the normal vectors of the two bases and the dihedral angle. In this figure and Figs. 4 and 7, the colored spheres represent C (cyan), H (white), N (blue), O (red), and P (brown) atoms.]

![FIG. 2. Hydrogen bonding energy versus distance $r_{hb}$ between two complimentary bases. Inset shows the dependence on the dihedral angle $\theta_d$ between the two bases. Symbols are BSSE-corrected results from DFT calculations, and lines are fittings to Eqs. (3) and (4) with parameters given in Table I.]

<table>
<thead>
<tr>
<th>bp</th>
<th>$E_0$ (eV)</th>
<th>$r_0$ (Å)</th>
<th>$l$ (Å)</th>
<th>$k_d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td>-0.639</td>
<td>8.866</td>
<td>0.703</td>
<td>1.800</td>
</tr>
<tr>
<td>GC</td>
<td>-1.165</td>
<td>9.018</td>
<td>0.727</td>
<td>1.288</td>
</tr>
</tbody>
</table>

TABLE I. Fitting parameters for the radial dependence of the hydrogen bond interaction in Eq. (3) and the dihedral angle dependence in Eq. (4).
two bases fixed. The resulting energies are shown in the inset of Fig. 2. We fit the interaction energy with the expression

\[ E_{\text{hb}}(\theta_d) = E_0 e^{k_d \cos(\theta_d)}, \]  

where \( E_0 \) is the same as in Eq. (3); values of the parameter \( k_d \) are given in Table I.

The interaction between the two bases also depends on the in-plane angles of the two bases. In our coarse-grained model, this is described by the flip-angle \( \phi_{\text{hb}} \) for the two bases \( (i = 1, 2) \), as defined as the base1-sugar1-sugar2 angle for \( i = 1 \) and the base2-sugar2-sugar1 angle for \( i = 2 \) (see Fig. 1(b)). The flip-angles in the ground-state configuration, \( \phi_{\text{hb}}^{(i,0)} \), are listed in Table II. Starting from the energy-minimized geometry \( (\phi_{\text{hb}}^{(i,0)}) \), we consider either rotating the pyrimidine around the H1 atom, or rotating the purine around the H9 atom. The anchoring hydrogen atom mimics the backbone, which in physiological conditions remains relatively stationary when base flipping occurs. To keep \( \phi_{\text{hb}} \) fixed at each rotated configuration, we optimize the geometry while fixing the pyrimidine atoms N1 and H1, and the purine atoms N9 and H9. For rotation inward, the atoms are constrained on the base-pair plane to prevent the two bases from rotating out of plane.

The resulting energies are plotted as a function of \( \Delta \phi_{\text{hb}}^{(i)} = \phi_{\text{hb}}^{(i)} - \phi_{\text{hb}}^{(i,0)} \) in Fig. 3. When \( \Delta \phi_{\text{hb}}^{(i)} \) is positive, the energy is attractive and decays to zero. When \( \Delta \phi_{\text{hb}}^{(i)} \) is negative, the two bases repel each other. We describe these two types of behavior separately, and fit the base-flipping interaction energy to the expression

\[ E^{(i)}_{\text{hb}}(\phi_{\text{hb}}^{(i)}) = E_{\text{hb}}^{(a,i)}(\Delta \phi_{\text{hb}}^{(i)}) + E_{\text{hb}}^{(b,i)}(\Delta \phi_{\text{hb}}^{(i)}) \quad (i = 1, 2), \]

where \( E_{\text{hb}}^{(a,i)} \) is described by an exponential and \( E_{\text{hb}}^{(b,i)} \) is described by a harmonic spring

\[ E_{\text{hb}}^{(b,i)} = E_0 \left[ \frac{(\Delta \phi_{\text{hb}}^{(i)})^2}{2\sigma^2} \right] \Theta(-\Delta \phi_{\text{hb}}^{(i)}), \]  

and \( \Theta \) is the step function. Again, \( E_0 \) is the same as in Eq. (3). The values of the parameter \( \sigma \) are listed in Table II.

Up to this point we have the interaction potential between the two complementary bases as a function of \( r_{\text{hb}}, \theta_d, \phi_{\text{hb}}^{(1)}, \) and \( \phi_{\text{hb}}^{(2)} \), individually while keeping other variables fixed at the equilibrium values. Here we assume that the interaction depends only on these four variables; even in this case, the general dependence is not trivial, as the four variables are not independent. For example, when \( \theta_d \) or \( \Delta \phi_{\text{hb}}^{(1)} \) is large, the hydrogen bond is basically broken, and there should no longer be a strong dependence on \( r_{\text{hb}} \). To capture the interdependences between the variables while keeping a reasonably simple form for the interaction, we define the following functions

\[ r_{\text{hb}}-d(\theta_d) = E_{\text{hb}}-d(\theta_d)/E_0, \]

\[ \tau_{\text{hb}}^{(i)}(\phi_{\text{hb}}^{(i)}) = E_{\text{hb}}^{(a,i)}(\Delta \phi_{\text{hb}}^{(i)})/E_0 \quad (i = 1, 2), \]

and take the final expression of the hydrogen bonding energy to be

\[ E_{\text{hb}}(r_{\text{hb}}, \theta_d, \phi_{\text{hb}}^{(1)}, \phi_{\text{hb}}^{(2)}) = E_{\text{hb}}^{(a)}(r_{\text{hb}})r_{\text{hb}}-d(\theta_d) \prod_{i=1}^{2} \tau_{\text{hb}}^{(i)}(\phi_{\text{hb}}^{(i)}) + \sum_{i=1}^{2} E_{\text{hb}}^{(b,i)}(\phi_{\text{hb}}^{(i)}). \]  

Note that the repulsive part of the flipping interaction \( E_{\text{hb}}^{(b,i)} \) is treated as additive since it does not serve to weaken the bond; this also ensures that the modulation functions \( \tau_{\text{hb}}^{(i)} \) take values strictly between 0 and 1. Equation (9) is reduced to Eqs. (3)–(5) for close-to-minimum geometries, i.e.,

\[ E_{\text{hb}}(r_{\text{hb}}, 0, \phi_{\text{hb}}^{(1),0}, \phi_{\text{hb}}^{(2),0}) = E_{\text{hb}}^{(a)}(r_{\text{hb}}), \]

\[ E_{\text{hb}}(r_0, \theta_d, \phi_{\text{hb}}^{(1),0}, \phi_{\text{hb}}^{(2),0}) = E_{\text{hb}}^{(a)}(\theta_d), \]

\[ E_{\text{hb}}(r_0, 0, \phi_{\text{hb}}^{(1),0}, \phi_{\text{hb}}^{(2),0}) = E_{\text{flip}}(\phi_{\text{hb}}^{(1),0}). \]

Therefore, we expect that close to equilibrium, Eq. (9) serves as a good approximation to the interaction energy between the two bases.

**TABLE II.** Fitting parameters for the flip-angle interaction in Eqs. (6) and (7), and the backbone base-sugar-sugar angle potential in Eq. (18). \( k_{\text{bas}} \) is in \( 10^{-4} \text{eV/deg}^2 \); all angles are in degrees.

<table>
<thead>
<tr>
<th>Base</th>
<th>( \phi_{\text{hb}}^{(i,0)} )</th>
<th>( \sigma^{(i)} )</th>
<th>( k_{\text{bas}} )</th>
<th>( \theta_d^{(i)} )</th>
<th>( \theta_d^{(i)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>54.53</td>
<td>18.67</td>
<td>3.489</td>
<td>94.17</td>
<td>61.36</td>
</tr>
<tr>
<td>T</td>
<td>55.93</td>
<td>16.82</td>
<td>4.689</td>
<td>92.67</td>
<td>63.40</td>
</tr>
<tr>
<td>G</td>
<td>52.69</td>
<td>25.57</td>
<td>4.165</td>
<td>90.30</td>
<td>63.69</td>
</tr>
<tr>
<td>C</td>
<td>54.87</td>
<td>22.43</td>
<td>6.178</td>
<td>91.55</td>
<td>66.37</td>
</tr>
</tbody>
</table>
B. Stacking interactions

We turn next to the interaction between two stacked base-pairs. We consider the dependence on the stacking distance \( r_{st} \) between the two base-pairs first. The two base-pairs are relaxed and stacked in parallel. Following usual conventions,\(^5\) we define the axial point of each base-pair to be the 1:2 weighted center of the green atoms. The atoms used to define the twist angle \( \theta_{tw} \) are colored in green, and the axis of the double helix (shown as a dotted circle) is defined as the 1:2 weighted center of the green atoms. In (b), the planes of the two base-pairs are shown. The twist angle is reduced for clarity of illustration. The four vectors \( \mathbf{d}_1, \mathbf{d}_1', \mathbf{d}_2, \) and \( \mathbf{d}_2' \), as indicated are used to define the stacking distance \( r_{st} \) in the coarse grained representation. See main text for description of the other geometric variables involved in stacking.

There are ten different combinations of stacking, known as the ten Watson-Crick nearest-neighbors. For each of these ten combinations, we vary \( r_{st} \) from 2.5 Å to 5.0 Å and calculate the interaction energy without further relaxation. The results are shown in Fig. 5. At this range of \( r_{st} \), the energy variation is well described by the expression

\[
E_{st,z}(r_{st}) = -\epsilon \left[ 5 \left( \frac{r_m}{r_{st}} \right)^6 - 6 \left( \frac{r_m}{r_{st}} \right)^5 \right], \tag{11}
\]

which has a minimum at \( r_{st} = r_m \). \( E_{st,z} = \epsilon \). The resulting values of the parameters from fitting the ab initio values with this expression are listed in Table III. It may seem surprising that the attractive part of the interaction behaves as \( r^{-5} \). This is no coincidence: the attraction between the two base-pairs is due to the vdW force, which behaves as \( r^{-6} \) for two point particles and as \( r^{-4} \) for two thin sheets (assuming additivity of the vdW interaction). The range of \( r_{st} \) being considered here is comparable to the radius of the base-pairs (about 4 Å), so we expect a power in between the two, i.e., \( r_{st}^{-5} \). We find this is indeed the case; expressions with any other integer power of \( r_{st} \) lead to poor fitting. Again, we do not parametrize the stacking interaction for mismatched base-pairs, which can be a straightforward extension to the current work.

TABLE III. Fitting parameters of the stacking interaction in Eq. (11) for the ten Watson-Crick nearest-neighbors. In our notation, GC-AT indicates the stacking of GC base-pair and AT base-pair with GA and TC following the 3′–5′ direction.

<table>
<thead>
<tr>
<th>bps</th>
<th>( r_m (\text{Å}) )</th>
<th>( \epsilon ) (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT-AT</td>
<td>3.550</td>
<td>-0.567</td>
</tr>
<tr>
<td>GC-GC</td>
<td>3.580</td>
<td>-0.477</td>
</tr>
<tr>
<td>GC-AT</td>
<td>3.566</td>
<td>-0.555</td>
</tr>
<tr>
<td>AT-TA</td>
<td>3.546</td>
<td>-0.530</td>
</tr>
<tr>
<td>GC-CG</td>
<td>3.498</td>
<td>-0.632</td>
</tr>
<tr>
<td>GC-TA</td>
<td>3.543</td>
<td>-0.549</td>
</tr>
<tr>
<td>AT-GC</td>
<td>3.535</td>
<td>-0.538</td>
</tr>
<tr>
<td>CG-GC</td>
<td>3.537</td>
<td>-0.565</td>
</tr>
<tr>
<td>TA-GC</td>
<td>3.613</td>
<td>-0.529</td>
</tr>
<tr>
<td>TA-AT</td>
<td>3.668</td>
<td>-0.513</td>
</tr>
</tbody>
</table>
resulting energies $E_{tw}$ are shown in Fig. 6. This energy term has a complex dependence on $\theta_{tw}$, defying a simple analytical expression. Given its periodicity, we fit $E_{tw}$ with a Fourier series:

$$E_{tw}(\theta_{tw}) = a_0 + \sum_{n=1}^{7} [a_n \cos(n\theta_{tw}) + b_n \sin(n\theta_{tw})].$$  \hspace{1cm} (12)

The resulting values of the coefficients in this expansion are given in Table IV. For AT-AT and GC-GC stacking, $E_{tw}(-\theta_{tw}) = E_{tw}(\theta_{tw})$ by symmetry, and so the coefficients $b_n$ are zero. We note that although Eq. (12) involves many trigonometric functions, the higher terms can be obtained from trigonometric addition rules or from the Chebyshev method. Therefore, its computational cost is similar to that of Eq. (11).

The stacking interaction also depends on the tilt-angle $\theta_{tl}$, which we define as the angle between the normal vectors of the two base-pair planes. The two base-pairs can tilt in a variety of ways, and so this dependence can be complex. We find that close to the optimal stacked configuration (at small $\theta_{tl}$, and $r_{ss} = r_{0s}$, $\theta_{tw} = 36^\circ$), the interaction energy dependence can be approximately described by

$$E_{st}(\theta_{tl}) = \varepsilon \cos^2(\theta_{tl}),$$ \hspace{1cm} (13)

where $\varepsilon$ is the same as in Eq. (11). This expression also has the physical meaning that the energy is at a minimum in the case of parallel ($\theta_{tl} = 0$) and anti-parallel ($\theta_{tl} = \pi$) stacking, and is zero when the two base-pairs are perpendicular to each other ($\theta_{tl} = \pm \pi/2$).

The interaction between two stacked base-pairs depends on $r_{ss}$, $\theta_{tw}$, and $\theta_{tl}$. To capture all three dependences and their correlations with a reasonably simple form, we define the following functions:

$$\tau_{tw}(\theta_{tw}) = E_{tw}(\theta_{tw})/\varepsilon, \hspace{0.2cm} \tau_{tl}(\theta_{tl}) = E_{tl}(\theta_{tl})/\varepsilon$$ \hspace{1cm} (14)

and take the final expression of the stacking interaction to be

$$E_{st}(r_{ss}, \theta_{tw}, \theta_{tl}) = E_{st-tw}(r_{ss}) \tau_{tw}(\theta_{tw}) \tau_{tl}(\theta_{tl}),$$ \hspace{1cm} (15)

similar to our treatment of the hydrogen bond. Equation (15) involves much simplification, as the true interaction energy may depend on more than these three variables, and the dependence on $r_{ss}$, $\theta_{tw}$, and $\theta_{tl}$ may not factorize. By the same argument as for the hydrogen bonds, though, we expect Eq. (15) to be a good approximation close to equilibrium.

The stacking potential here is formulated as interaction between neighboring base-pairs, rather than between neighboring bases. This approach is sufficient when dealing with ds-DNA structures. However, we note that it may not be appropriate in processes like melting or hybridization that involve single-stranded DNA or broken base-pairs. In such situations, the stacking interaction should be reformulated as interaction between bases instead and be constructed in a similar fashion.

C. Backbone contribution

We next consider contributions from the backbone. First, to extract interactions within the sugar-phosphate backbone as distinct from any electrostatic or stacking interactions, we take the phosphate groups to be protonated and carry no charge, and replace the bases with terminating hydrogens. Starting from the energy-minimized geometry, we uniformly stretch or compress the backbone along the helical direction, as illustrated in Fig. 7(a), and allow the phosphate units to relax. The resulting energy $E_{bbs-ss}$ is shown in Fig. 8 as a function of $r_{ss}$, the distance between the C1’ atoms of neighboring sugars. The interaction energy per sugar-phosphate unit can be described by the expression

$$E_{bb-ss}(r_{ss}) = c_2(r_{ss} - r_{ss0})^2 + c_4(r_{ss} - r_{ss0})^4$$ \hspace{1cm} (16)

with $c_2 = 18.773$ eV/Å$^2$, $c_4 = 0.333$ eV/Å$^4$, and $r_{ss0} = 4.976$ Å.
The base is covalently bonded to the backbone. For this interaction, we consider only a base and a sugar, with the phosphate groups replaced by hydrogen atoms. Starting from the energy-minimized geometry, we translate the base and the sugar groups in opposite directions along the vector connecting the two. Then we hold the two atoms of the base-sugar covalent bond fixed, optimize the geometry, and calculate the interaction energy. We find results for the four bases to be nearly identical, so only results for guanine will be discussed. These are shown in Fig. 9, with a fit to the UBER expression that includes two additional terms

\[ E_{bb,c}(r_{CN}) = E_0(1 + a^* + f_2a^{*2} + f_3a^{*3})e^{-a^*}, \]
\[ a^* = (r_{CN} - r_0)/l, \]

where \( r_{CN} \) is the distance between the sugar C1’ atom and the base N atoms (N9 of purines or N1 of pyrimidines). The values of the parameters are: \( E_0 = -3.542 \text{ eV}, r_0 = 1.455 \text{ Å}, l = 0.400 \text{ Å}, f_2 = -0.132, \) and \( f_3 = 0.215. \)

The base also interacts with neighboring backbone groups, and this gives rise to the 5’/3’ asymmetry of the double-helix. We characterize this interaction using the angle between the backbone strand and the base. We consider a segment of single-stranded DNA, with three sugar groups connected by two protonated phosphate groups. The first and third bases are replaced with terminating hydrogen atoms, and the middle base is either A, C, G, or T (see Fig. 7(b)). From the energy-minimized geometry, we rotate the middle base around the vector \( r_{CC5} \times r_{CN} \) (both are shown in Fig. 7(b)), with the pivot point at C1’ of the middle sugar. For each rotated configuration, we relax the system while holding the connecting N atom of the base (N9 of purines or N1 of pyrimidines) and the C1’ atoms of the three sugars fixed. The resulting energy is shown in the inset of Fig. 9. We consider this energy to be a function of the two angles \( \theta_3 \) and \( \theta_5 \), where \( \theta_3 \) is the angle between \( r_{CN} \) and \( r_{CC5} \), and \( \theta_5 \) is the angle between \( r_{CN} \) and \( r_{CC5} \) (see Fig. 7(b)). We fit this energy term with the function

\[ E_{bb,b} = k_{bas} \frac{1}{2} \left[ (\theta_3 - \theta_3^{(0)})^2 + (\theta_5 - \theta_5^{(0)})^2 \right]. \]

where \( \theta_3^{(0)} \) and \( \theta_5^{(0)} \) are the angles in the initial configuration relaxed without constraint, and \( k_{bas} \) is obtained from fitting. The resulting values for these parameters are listed in Table II.

We treat Eqs. (16)–(18) as independent interactions, and so the total contribution of the backbone and its interaction with the base-pairs is given by

\[ E_{bb} = E_{bb,x} + E_{bb,c} + E_{bb,b}. \]

D. Electrostatics

All the potentials considered so far represent bonded or short-ranged interactions between components of DNA. The DFT calculations considered DNA as being composed of strictly neutral components in vacuum. However, under
physiological conditions, DNA is solvated in an aqueous electrolyte, and the phosphate groups in DNA are deprotonated. Therefore, we place a charge of \(-e\) on each sugar-phosphate group of the coarse-grained model, and include a Coulomb potential between these groups

\[
E_{\text{el}}(r) = \frac{e^2}{4\pi \varepsilon_0 \varepsilon(r)r}
\]

with a distance-dependent dielectric function \(\varepsilon(r)\), \(r\) being the distance between the charges and \(\varepsilon_0\) the dielectric constant in vacuum. This \(\varepsilon(r)\) plays two roles: (i) it incorporates the effects of ionic screening, and (ii) it accounts for the fact that closely spaced interacting groups are only partially solvated by the surrounding electrolyte. We use the following expression, closely related to the formulation of Ref. 52, for the dielectric function:

\[
\varepsilon(r) = \begin{cases} 
\varepsilon_{\text{int}}, & \text{for } r < r_0 \\
\varepsilon_{\text{int}} e^{\alpha(r-r_0)}, & \text{for } r_0 < r < r_1 \\
\varepsilon_{\infty} e^{\kappa r}, & \text{for } r > r_1 
\end{cases}
\]

where \(\varepsilon_{\infty} = 78\) is the dielectric constant of water, and \(\varepsilon_{\text{int}}\) is the dielectric constant in the interior of the DNA helix, which we take to be 3.52 The Debye length is \(\kappa^{-1}\), related to the ionic strength \(I\) through

\[
\kappa^{-1} = \sqrt{\frac{\varepsilon_0 \varepsilon \kappa T}{2 N_A e^2 I}},
\]

where \(\kappa T\) is the thermal energy and \(N_A\) is Avogadro’s number. For example, at [\(\text{Na}^+\)] = 0.1 M, the Debye length is 9.6 Å. The values \(r_0\) and \(r_1\) determine the boundary between unscreened and screened electrostatic interactions. The characteristic sizes of the chemical groups represented by our coarse-grained units range from about 2 Å to about 5 Å. Consequently we set \(r_0 = 4\) Å. Similarly, we choose \(r_1 = 13\) Å, approximately five times the mean water oxygen–water oxygen distance in bulk water,33 as the distance where the effective dielectric constant recovers the value predicted from the Debye–Hückel theory of screening in a bulk electrolyte. The value of \(\alpha\) is then chosen such that \(\varepsilon(r)\) is continuous. Between the two charged groups within the same base-pair, \(\varepsilon(r) = \varepsilon_{\infty}\) is used. Finally, since the electrostatic interaction decays exponentially at large distance, we truncate this interaction for distances above five times the Debye length. The simple electrostatics approach here is a first approximation, and is not meant to capture details such as the dependence of ion condensation on conformation,44 which will require explicit solvent molecules.

IV. IMPLEMENTATION AND PERFORMANCE OF THE COARSE-GRAINED MODEL

In Sec. III we derived the interactions between the coarse-grained sites. The total interaction energy is given by Eq. (2), and the different contributions include summations over all relevant interacting units. Specifically, \(E_{\text{ext}}\) includes a summation over all base-pairs, and \(E_{\text{stat}}\) includes a summation over all pairs of neighboring base-pairs. Of the three terms that comprise the \(E_{\text{bb}}\) contribution, \(E_{\text{bb},r}\) includes a summation over all pairs of neighboring backbone sites, \(E_{\text{bb,c}}\) includes a summation over all pairs of connected backbone and base sites, and \(E_{\text{bb,b}}\) includes a summation over all bases. Finally, \(E_q\) includes a summation over all pairs of backbone sites.

Next, we relate the geometrical variables to the positions of the coarse-grained sites. In the coarse-grained representation, the base site is identified with the pyrimidine N1 atom or the purine N9 atom, and the backbone site is identified with the sugar C1’ atom. Then the distance and angle variables \(r_{\text{bb}}, \phi_{\text{bb}}^{(1)}, \phi_{\text{bb}}^{(2)}, r_{\text{CN}}, \theta_3, \text{ and } \theta_5\) follow from the position of the coarse-grained sites. We define the remaining geometrical variables as follows. For each base-pair, we define the normal vector of base \(i\) to be \(n_i^{(1)} = r_{\text{CC}} \times r_{\text{CN}}^{(i)}\) with \(i = 1, 2\) corresponding to the two bases, and with \(r_{\text{CC}}\) and \(r_{\text{CN}}^{(i)}\) defined in Fig. 1(b). The dihedral angle \(\theta_d\) of this base-pair is given by \(\cos \theta_d = \mathbf{n}_1 \cdot \mathbf{n}_2\), with the hat denoting unit vectors. We define the average normal of this base-pair to be \(\mathbf{n} = \mathbf{n}_1 + \mathbf{n}_2\), and let the tilt-angle \(\theta_{\text{tl}}\) between base-pair \(j\) and base-pair \(j + 1\) be given by \(\cos \theta_{\text{tl}} = \mathbf{n}_j \cdot \mathbf{n}_{j+1}\). For the stacking distance \(r_{\text{st}}\), we take the average distance of the four sites in base-pair \(j + 1\) to the plane of base-pair \(j\). Specifically, we compute the four vectors \(\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \text{ and } \mathbf{d}_4\), shown in Fig. 4(b), project them onto the two normal vectors as \(z_i = \mathbf{d}_i \cdot \mathbf{n}_1^{(i)}\) and \(z_i' = \mathbf{d}_i' \cdot \mathbf{n}_1^{(i)}\) for \(i = 1, 2, \text{ and } \text{take } r_{\text{st}} = (z_1 + z_1' + z_2 + z_2')/4\). Finally, the twist angle \(\theta_{\text{tw}}\) is defined as the angle between \(r_{\text{CC},j}\) and \(r_{\text{CC},j+1} - (r_{\text{CC},j+1} \cdot \mathbf{n}_1^{(j+1)} \mathbf{n}_1^{(j+1)}\) which is the projection of \(r_{\text{CC},j+1}\) onto the plane of base-pair \(j\).

We implement this model for calculations in the microcanonical ensemble (constant number of particles \(N\), volume \(V\), and energy \(E\)) and the canonical ensemble (constant \(N\), \(V\), and temperature \(T\)) with implicit solvent Brownian dynamics. We choose the mass of each site to be the total mass of the group of atoms that it represents: 178 amu for the backbone site, 134 amu for base A, 125 amu for T, 150 amu for G, and 110 amu for C. Multiple-time-step integrators are used for time propagation: the RESPA algorithm55 is used for the \(\text{NVE}\) ensemble, and the multiple-time-step stochastic integrator56 is used for Brownian dynamics. The Coulomb interaction between different base-pairs is treated as the slow-varying component of the Hamiltonian, integrated with a time-step of 20 fs, while all the other forces are categorized into the fast-varying component of the Hamiltonian, integrated with a time-step of 6.67 fs (i.e., three divisions per large time-step). This choice of time-steps ensures high stability: when running in the \(\text{NVE}\) ensemble at 300 K, the total energy is conserved to within \(5 \times 10^{-4}\) eV per base-pair for arbitrary sequences at several conditions tested.

As a measure of the performance of this coarse-grained model, a 20 fs step of the stochastic integrator for a DNA molecule consisting of 250 base-pairs at 0.1 M salt concentration takes 2.4 ms on a single Intel Xeon X5660 processor (2.80 GHz). A 100-ns simulation of such a strand takes 3 h 20 min. This performance exceeds the speed of all-atom simulation methodologies by many orders of magnitude, and is comparable to other coarse-grained models with similar complexity. The computation time scales linearly with the number
of base pairs, since the electrostatic interactions are damped. This performance makes simulations of ds-DNA in the \( \mu \text{m} \) length scale and \( \mu \text{s} \) time scale feasible.

V. MODEL VALIDATION

The different contributions in the model potentials have been derived from calculations of isolated units. Here we test the validity of the combined potential (additivity of the individual contributions), and its performance in larger systems (transferability of the potentials). Also, our \textit{a posteriori} inclusions of the ionic and temperature effects are crude, and their consequences will also be assessed.

A. Equilibrium structure

First, we test the structure prediction of the model with poly-AT DNA consisting of 250 base-pairs (bp) at temperature 300 K and salt concentration 0.1 M. When initialized with coordinates of the B-form DNA, the B-DNA structure is maintained. To test the robustness of our model potential, we also carried out simulations starting with a highly off-equilibrium structure—an elongated and uncoiled double strand—and followed the evolution. Figure 10 shows that in such case, the system is able to coil back to the B-DNA structure. Signatures of the B-DNA structure such as the major groove, minor groove, and the 10 bps-per-turn period are well reproduced. This demonstrates the model’s ability to predict the stable double helix structure, without resorting to Go-like potentials that are defined relative to a specific reference structure as in Ref. 14.

To be quantitative, we calculate the averages of many structural properties after the system equilibrates. These quantities are listed in Table V, with comparison to reference values derived from recent crystallographic data of a naturally occurring 16 bps oligomer. The average deviation of our model from experiment is about 2%. Given that no experimental data were used in the model construction, this close agreement with experiment is remarkable.

B. Persistence length

Next, we check the model’s ability to predict mechanical properties at long lengths by examining the persistence length \( l_p \), which gives an estimate of the bending rigidity of DNA. The persistence length can be extracted from the decay of the orientational correlation function

\[
\langle \hat{r}_i \cdot \hat{r}_j \rangle = e^{-s_{ij}/l_p},
\]

where \( \hat{r}_i \) is the unit tangent vector at base-pair \( i \), and \( s_{ij} \) is the arc length from base-pair \( i \) to base-pair \( j \). The tangent vectors and the arc length are evaluated along the axis of the double-helix of DNA. As the DNA axis is not an explicit interaction site in our coarse-grained model, we extrapolate its location by estimating the direction of the local helical axis and averaging the projection of the backbone sites onto the local axial direction. We consider 250-bp DNA with poly-AT sequence, poly-GC sequence, and a segment of the enterobacteriophage-\( \lambda \) genome (with GC content 0.47). Simulations at 300 K and 0.15 M salt concentration give \( l_p = 53 \text{ nm} \), 47 nm, and 41 nm, respectively, for the poly-AT, poly-GC, and the phage oligomer strand. These values are in close agreement with the experimental value of \( l_p \approx 50 \text{ nm} \) for double-stranded DNA. The fittings used to obtain these values are shown in the inset of Fig. 11.

We also test the validity of the implicit salt treatment by considering the salt dependence of the persistence length. This dependence has been shown experimentally\(^{59}\) to agree well with the nonlinear Poisson-Boltzmann prediction for uniformly charged cylinders\(^{60}\)

\[
l_p = l_0 + \frac{1}{4\kappa^2 l_0^2},
\]

where \( l_0 \) is the persistence length at infinite salt concentration, \( \kappa^{-1} \) is the Debye length given in Eq. (22), and \( l_0 \) is the Bjerrum length. The second term is the electrostatic contribution, and is inversely proportional to the salt concentration.
sudden extension, for pulling with 640 pN on the 5′ end.

Inset of Fig. 13 shows snapshots of the DNA before and after the sudden stretching and unwinding occur. The simulations capture the superstretching transition, although the critical force in our model is higher, and the amount of the sudden extension is less. The superstretching transition is a very stringent test for our computational model, as it involves strong departures from the equilibrium structure. Notwithstanding the neglect of solvent interactions and the athermal character, our coarse-grained potential still provides reasonable results under such critical conditions. Simulations also reveal that the superstretching is accompanied by unwinding of the double helix, as shown in Fig. 12. Above the critical force, our model shows that the DNA is twisted slightly in the left-handed direction.

D. Bubble formation

Although our potentials are formulated for ds-DNA, in some applications it is desirable to have a model that can account for broken hydrogen bonds. Here, we test the validity of our model when the double strand in a 50-bp poly-AT DNA is partially broken. The onset of melting can be observed in our model, but the melting temperature is overestimated. At temperature 550 K, the double-strand shows occasional unzipping; one instance is shown in Fig. 13(a). At 800 K, formation of bubbles can be observed; one instance is shown in Fig. 13(b). The overestimation of melting temperature may arise from several reasons. First, our stacking potential is not formulated for single strands or strands with broken hydrogen bonds. Second, the presence of water molecules may lower the energy of the broken-bond states, but this effect is not taken into account. Third, the coarse-grained potentials here are derived at zero temperature, which is stiffer than at finite temperature. However, the appearance of these intermediate molten states is still an indication that our model is stable beyond strictly ds-DNA structures.

VI. CONCLUSION

We have constructed the interaction potentials for a coarse-grained model of double-stranded DNA. The model potentials are derived from first-principles calculations for the DNA bases and base-pairs. Contributions to the potential include hydrogen bonding, base-stacking, backbone stretching, and interactions between bases and the backbone. Analytical functions are fitted to the computed energy, leading to a potential for DNA that is able to handle coarse-grained
configurations of random DNA sequences and can be used to model related biophysical processes. Unavoidably, some simplifying assumptions were used in the model formulation and construction. In order to use a minimal number of empirical terms, only the effect of ionic screening was added to the microscopically derived potential. At a later stage, such empirical term can be replaced by a more microscopic approach.63

A series of tests were performed, and they verified that the coarse-grained potential can reproduce the stable B-DNA structure, and that the predicted structure matches the known crystallographic structure of DNA to a few % in key structural parameters. The model produces persistence lengths in close agreement with experimental data, and the response of the persistence length to varying salt concentrations also agrees with the prediction for a linked-cylinder model. These tests suggest that the mechanical properties of the coarse-grained model show realistic trends when subject to the complex electrokinetic environment in the cell or in artificial systems such as the interior of a nanochannel during translocation experiments.

The tests on overstretched and on bubble formation show that the model may capture the qualitative features in such states, but also reveal some limitations of the model. To describe such states accurately, we may need a more realistic account of the solvent molecules and of the temperature dependence of the coarse-grained potentials. This will be the subject of future work. Other possible and more straightforward extensions of the current model will be to formulate the stacking interaction to work with ss-DNA, and to parametrize interactions for mismatched base-pairs and for uracil (to extend the model to RNA). Also, the current model does not account for inter-strand interactions except for the electrostatic repulsion, and so situations like long DNA strands under confinement may not be appropriate for this model.

The approach described in this work is attractive for describing ds-DNA using a minimal of empirically derived parameters and with a good compromise between accuracy and computational efficiency. We suggest that this model could be a useful tool for simulating the behavior of ds-DNA in a variety of biologically relevant or device-related scenarios with modest computational resources.

ACKNOWLEDGMENTS

The authors thank Sheng Meng and Wei Li Wang for a critical reading of the manuscript, and acknowledge G.N. Patey for helpful comments and discussions. C.W.H. thanks Wei Li Wang for his kind helps and fruitful discussions. M.F. acknowledges support by Harvard’s Nanoscale Science and Engineering Center, funded by the National Science Foundation, Award Number PHY-0117795. G.L. acknowledges support from the post-doctoral fellowship program of the Natural Science and Engineering Research Council of Canada.
