JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Sequence-Dependent Persistence Lengths of DNA

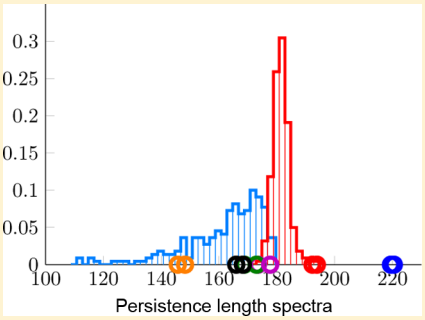
Jonathan S. Mitchell,[†,§] Jaroslaw Glowacki,[†] Alexandre E. Grandchamp,[†] Robert S. Manning,[‡] and John H. Maddocks*[,†]

[†]Ecole Polytechnique Fédérale de Lausanne, Lausanne CH 1273, Switzerland

[‡]Haverford College, Haverford, Pennsylvania 19041, United States

Supporting Information

ABSTRACT: A Monte Carlo code applied to the *cgDNA* coarse-grain rigid-base model of B-form double-stranded DNA is used to predict a sequence-averaged persistence length of l_F = 53.5 nm in the sense of Flory, and of l_p = 160 bp or 53.5 nm in the sense of apparent tangent−tangent correlation decay. These estimates are slightly higher than the consensus experimental values of 150 bp or 50 nm, but we believe the agreement to be good given that the *cgDNA* model is itself parametrized from molecular dynamics simulations of short fragments of length 10−20 bp, with no explicit fit to persistence length. Our Monte Carlo simulations further predict that there can be substantial dependence of persistence lengths on the specific sequence $\mathbb{S}$ of a fragment. We propose, and confirm the numerical accuracy of, a simple factorization that separates the part of the apparent tangent−tangent correlation decay $l_p(\mathbb{S})$ attributable to intrinsic shape, from a part $l_d(\mathbb{S})$ attributable purely to stiffness, i.e., a sequence-dependent version of what has been called sequence-averaged dynamic persistence length $\overline{l}_d$ (=58.8 nm within the *cgDNA* model). For ensembles of both random and λ-phage fragments, the apparent persistence length $l_p(\mathbb{S})$ has a standard deviation of 4 nm over sequence, whereas our dynamic persistence length $l_d(\mathbb{S})$ has a standard deviation of only 1 nm. However, there are notable dynamic persistence length outliers, including poly(A) (exceptionally straight and stiff), poly(TA) (tightly coiled and exceptionally soft), and phased A-tract sequence motifs (exceptionally bent and stiff). The results of our numerical simulations agree reasonably well with both molecular dynamics simulation and diverse experimental data including minicircle cyclization rates and stereo cryo-electron microscopy images.



1. INTRODUCTION

It is widely believed that understanding the sequence-dependent mechanical properties of DNA is an important step toward understanding many important biological processes, for example, nucleosome positioning,[1−3] and other protein−DNA interactions involving the phenomenon called indirect readout.[4] DNA "rigidity" is often expressed as a sequence-averaged, single parameter, namely, the persistence length, which is sometimes informally described as a measure of the length scale over which correlation between the tangents along a polymer centerline is lost.[5,6] Frequently, the persistence length is extracted by interpreting experimental data with the Kratky−Porod wormlike chain (or WLC) model[7,8] where the persistence length is one of only two free parameters. There is a consensus in the literature that the persistence length of DNA is approximately 150 bp, or 50 nm. This value is estimated by using diverse experimental techniques, each with their own assumptions necessary to interpret the data, and often at quite different length scales.[5,6] Consequently, the estimate can be regarded as robust, but not necessarily very precise, in part because it is understood that a single sequence-averaged persistence length combines (at least) two distinct physical effects on the correlations along the DNA, namely, both stiffness and intrinsic shape, which led to the notions of sequence-averaged dynamic and static persistence lengths.[9] And when the sequence dependence of a DNA fragment is of interest, then a description solely in terms of sequence-averaged persistence lengths is too imprecise.[10−15]

Consequently, although the WLC has proven extremely successful in interpreting diverse experimental results for DNA, its application to biological problems that depend significantly on sequence is precluded by its simplicity. Accordingly, there have been many efforts at developing more detailed, but still coarse-grained, models (see, e.g., the recent review by Dans et al.[16]). One class of such models involves the rigid base-pair approximation,[8,17] usually with the assumption of nearest-neighbor interactions and model parameters depending on the ten distinct dinucleotide steps. And one of the most successful parametrizations of such a model is by Olson et al.,[18] where the sequence-dependent variation in the parameter set was fit to protein−DNA crystal structure data. However, as is described with admirable clarity in the survey by Olson et al.,[19] the data require an overall scaling, which is determined by fitting to a sequence-averaged persistence length of 50 nm, the available sample size is not large, and it has many outliers, whose

Received: September 14, 2016
Published: December 28, 2016

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

treatment can strongly affect the fit. Other coarse-grain descriptions that incorporate an overall fit to a sequence-averaged persistence length include the oxDNA[20] and 3SPN[21] models, who use experimental data to estimate the sequence-dependent part of the parameter fit, and the work by Morriss-Andrews et al.,[22] Naome et al.,[23] and Uusitalo et al.,[24] who use molecular dynamics (or MD) simulations to estimate the sequence dependence. There are also sequence-dependent, coarse-grain models that *predict* sequence-averaged persistence lengths, for example, 15.2 nm,[25] 20 nm,[26] 96 bp,[27] and 75 nm.[28] Similarly, estimates of persistence length have been made directly from atomistic MD simulations of (necessarily) relatively short fragments at the scale 20−50 bp, e.g., 80 nm for poly(TA) and poly(GC),[29] 43 nm for a mixed sequence fragment,[30] and 40−57 nm also for mixed sequence fragments.[31]

We here assess the ability of the sequence-dependent, rigid *base*, coarse-grain *cgDNA* model[32] to reproduce the sequence-dependent statistical mechanics properties of B-form double helical DNA, by developing appropriate Monte Carlo (or MC) sampling methods to generate associated ensembles of configurations. The *cgDNA* model was itself parametrized from atomistic MD simulations of a library of 10−20 bp DNA fragments in explicit solvent with no explicit fit to persistence length. At the scale of tens of bp, the *cgDNA* free energy minimizers (or ground states) have already been shown to well approximate sequence-dependent intrinsic shapes when compared both to other MD simulations (with a sequence resolution of a single nucleotide permutation) and to NMR and X-ray crystallographic experimental structures.[33] The MC code developed here allows sampling of *cgDNA* Boltzmann distributions at the scales of tens to thousands of bp. Simulations of sequence-averaged persistence length yield the estimates of 53.5 nm in the sense of Flory (from simulations at the scale of 1 Kbp), and, independently, precisely the same value 53.5 nm in the sense of apparent tangent−tangent correlation decay (from simulations at the scale of 200 bp). Both of these estimates have a standard error of ±0.1 nm in the sense of multiple estimates from multiple MC simulations. We also find that the tangent−tangent persistence length has a mild dependence on the precise coarse-graining choices of tangents and arc length. Error associated with underlying imprecision in *cgDNA* parameters is harder to assess but is potentially significantly larger. However, comparison between expectations evaluated on MC and MD ensembles suggests that, at least for some short fragments, the coarse-grain *cgDNAmc* predictions of persistence length are just as accurate as those taken directly from MD, and with significantly less computational effort. The *cgDNAmc* code therefore offers a method of bridging the scales to allow sampling ensembles for much longer fragments, and of a much larger variety of sequences.

For a given DNA fragment, the MC simulations predict significant dependence of various ensemble expectations on the sequence $\mathbb{S}$ and led us to propose a sequence-dependent notion of dynamic persistence length $l_d(\mathbb{S})$ (detailed in eq (9) below) based on a factorization of the effects of stiffness and shape (at the scale of 200 bp) in the apparent tangent−tangent correlation persistence length $l_p(\mathbb{S})$ (detailed in eq (6)). A sequence-ensemble average of $l_d(\mathbb{S})$ is 58.8 ± 0.1 nm, with a standard deviation over sequence of 1 nm compared to the standard deviation in $l_p(\mathbb{S})$ of 4 nm, suggesting that most of the sequence-dependent variation in apparent persistence length is due to differences in intrinsic shape rather than differences in stiffness. However, our simulations also revealed some extreme outliers in $l_d(\mathbb{S})$, including poly(A) which has l_d = 73.1 nm, and poly(TA) which has l_d = 47.2 nm. As poly(A) is exceptionally straight, it also has the single largest apparent persistence length $l_p(\mathbb{S})$ of any sequence that we have observed, including an exhaustive search of all 151 distinct di-, tri-, tetra-, and penta-nucleotide repeating sequences. However, the *cgDNA* model is sufficiently detailed as to also be able to predict that phased A-tract motifs, i.e., certain periodic sequences with a short run of A followed by a short spacer region, are both exceptionally bent and exceptionally stiff.

The presentation is structured with first a Theory section, which describes the necessary statistical mechanics background in section 2.1 and then proposes our sequence-dependent dynamic persistence length $l_d(\mathbb{S})$ in section 2.2. We then turn to Simulation Methods in section 3 including choices in coarse graining, descriptions of the *cgDNA* free energy model, and the MC sampling methods that we employ. In section 4 we present the main results of our numerical simulations, which to some extent can be read independently of the other sections, and which show that DNA persistence lengths can have strong sequence dependence. Then in section 5 we present some more technical simulation data that serves to verify that the conclusions of section 4 are computationally robust, including in section 5.4 a direct comparison between MD and MC ensemble expectations. Finally, we make various comparisons between simulation and experiment, including in section 6.1 sequence-dependent 2D electron microscopy data[1,12] and minicircle cyclization data,[34] in section 6.2 a discussion of the special case of A-tracts, and in section 6.3 the 3D stereo cryo-electron microscopy (cryo-EM) approach of Bednar et al.[35] Possible sources of error in our modeling are discussed in section 6.4. We close with a summary in section 7.

2. THEORY

In section 2.1 we review the necessary background material regarding different notions of persistence lengths for general polymers, and then in section 2.2 we propose a *sequence-dependent dynamic persistence length* for DNA fragments.

2.1. Statistical Mechanics of Persistence Lengths. We will consider polymers modeled as a linear chain of rigid bodies whose configuration is described by a sequence of frames $(\mathbf{r}_n, \mathbf{R}_n)$ where $\mathbf{r}_n$ are the absolute coordinates of a reference point of each rigid body whose orientation is encoded in the direction-cosine (or proper rotation) matrix $\mathbf{R}_n$. Two of the classic expectations of polymer physics defined on such configurations are[7,14,36−38]

$$\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle, \qquad \langle \mathbf{R}_0^{\mathrm{T}} (\mathbf{r}_i - \mathbf{r}_0) \rangle \tag{1}$$

where $\langle \cdot \rangle$ denotes the ensemble average, i.e., the expectation of the argument with respect to an underlying equilibrium measure, $\mathbf{t}_0$ is a unit vector associated with a specific base pair labeled with index 0 (usually taken to be away from the physical end of the polymer to avoid any possible end effect), and $\mathbf{t}_i$ is the analogous unit vector at the ith base pair along the polymer. Usually $\mathbf{t}_i$ is to be interpreted as some approximation to a unit tangent to the polymer, so that $(1)_1$ is often described as a tangent−tangent correlation function. Similarly, $\mathbf{R}_0^{\mathrm{T}}(\mathbf{r}_i - \mathbf{r}_0)$ are the components of the chord vector between the zeroth and ith base-pair origins expressed in the chosen reference frame $\mathbf{R}_0$. We will call the expectations $(1)_2$ Flory persistence vectors, as

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

they were apparently first introduced by Flory[36] (for further discussion see Maroun and Olson[11] and Schellman and Harvey[13]). Accordingly, for each choice of reference frame $\mathbf{R}_0$, the expectations (1) are respectively scalar and vector functions of the index $i \geq 1$.

One of the simplest model ensembles in which to compute the expectations (1) is a discrete version of the Kratky–Porod WLC.[7,37] In this model the polymer is assumed to be a chain of rigid links all of length b, so that any configuration is described by unit chord vectors $\mathbf{t}_i := (\mathbf{r}_{i+1} - \mathbf{r}_i)/b$, and the equilibrium measure is assumed to be Boltzmann with inverse temperature scale $\beta = 1/k_B T$ and free energy (or Hamiltonian)

$$E = \frac{B}{b}\sum_{i=1}^{n}(1 - \mathrm{t}_i\cdot\mathrm{t}_{i+1}) \tag{2}$$

with B a (constant) bending rigidity parameter. In particular, the minimum energy, or ground, state of the WLC is intrinsically straight with all tangent vectors parallel. Then, provided that the nondimensional parameter $l_p := \beta B/b$ is large (compared to 1), it can be calculated analytically that the correlations $(1)_1$ are well approximated by the formula

$$\langle \mathbf{t}_i\cdot\mathbf{t}_0\rangle_{\mathrm{WLC}} = \mathrm{e}^{-\mathrm{i}/l_p} \tag{3}$$

The exponential decay scale l_p is the persistence length expressed in bp, whereas $bl_p = \beta B$ is the dimensional persistence length expressed in the (arc-)length units of b. Similarly, within the discrete WLC model, the expectations $(1)_2$ can be computed to be

$$\langle \mathbf{R}_0^{\mathrm{T}}(\mathbf{r}_i - \mathbf{r}_0)\rangle_{\mathrm{WLC}} = bl_p[0,\ 0,\ (1 - \mathrm{e}^{-\mathrm{i}/l_p})]^{\mathrm{T}} \tag{4}$$

provided only that the third column of $\mathbf{R}_0$ is chosen to coincide with $\mathbf{t}_0$. In fact, the specific functional forms of expressions (3) and (4) are only exact in the limit of the continuous WLC, in which the dimensional persistence length $\beta B = bl_p$ stays constant, while $b \to 0$, $N \to \infty$, $Nb \to L$, and $ib \to s \in [0, L]$. Explicit versions of more accurate analogous formulas for the discrete WLC are also available, e.g., in Schellman,[37] but the simpler approximations (3) and (4) suffice for our purposes.

For a DNA fragment with sequence $\mathbb{S}$, and motivated by the WLC formula (4), we introduce a Flory persistence length $l_F(\mathbb{S})$ as the limiting value of the magnitude (in the usual Euclidean distance $\|\cdot\|$) of the Flory persistence vector $(1)_2$ as $i \to \infty$, along with its sequence-averaged version $\overline{l}_F$:

$$l_F(\mathbb{S}) = \lim_{i\to\infty}\|\langle \mathbf{R}_0^{\mathrm{T}}(\mathbf{r}_i - \mathbf{r}_0)\rangle\|, \qquad \overline{l}_F = \lim_{i\to\infty}\|\{\langle \mathbf{R}_0^{\mathrm{T}}(\mathbf{r}_i - \mathbf{r}_0)\rangle\}\| \tag{5}$$

Here the brackets $\{\cdot\}$ in the second expression denote an additional average over an ensemble of sequences $\mathbb{S}_j$ of the average $\langle\cdot\rangle$ over an ensemble of configurations of a fragment with fixed sequence. The persistence lengths defined in (5) indeed have the dimension of length, which we will report in nanometers (or nm). For the (sequence-independent) WLC, the limiting value of the Flory persistence vector is simply computed from (4) to be $[0,\ 0,\ bl_p]^{\mathrm{T}}$, where the first two components of the vector vanish because the WLC model is isotropic with no distinguished direction for bending. Therefore, for the WLC the Flory persistence length coincides with the limiting value of the only nonvanishing, i.e., the third, component of the Flory persistence vector, namely $l_F = bl_p$. Thus, for the WLC, or any other isotropic model, the Flory persistence length (5) coincides with the original notion of Kratky–Porod persistence length, namely, an expected distance that an infinite polymer extends along an axis of isotropic symmetry. However, for anisotropic models, the limits of all three components of the Flory persistence vector can be nonzero, reflecting, for example, the effects of anisotropic intrinsic bends, so that the Flory persistence vector and associated Flory persistence length are a true generalization of the Kratky–Porod persistence length. Examples within the *cgDNA* model are provided in Figure 4 below.

Similarly, the WLC chain formula (3) motivates the definition of a sequence-dependent tangent–tangent correlation length $l_p(\mathbb{S})$ and its sequence-averaged version $\overline{l}_p$:

$$\mathrm{e}^{-\mathrm{i}/l_p(\mathbb{S})} \approx \langle \mathbf{t}_i\cdot\mathbf{t}_0\rangle, \qquad \mathrm{e}^{-\mathrm{i}/\overline{l}_p} \approx \{\langle \mathbf{t}_i\cdot\mathbf{t}_0\rangle\} \tag{6}$$

where the symbol $\approx$ signifies that, for a given sequence $\mathbb{S}$, $l_p(\mathbb{S})$ is computed as the number of base pairs equal to the (negative reciprocal) of the slope of the straight line through the origin that is the least-squares fit to the plot of $\ln\langle \mathbf{t}_i\cdot\mathbf{t}_0\rangle$ vs $i \in \mathbb{D}$, where $\mathbb{D}$ is a set of base-pair indices at which the fit is made (typically, but not necessarily, a range $i = 1, \ldots, N$); cf. Figure 5. Similarly, $\overline{l}_p$ is computed via the analogous semilog plot of the sequence-averaged data $\{\langle \mathbf{t}_i\cdot\mathbf{t}_0\rangle\}$ vs $i \in \mathbb{D}$.

We note that both the Flory l_F and tangent–tangent l_p persistence lengths as introduced above are equally valid scalar measures of the "persistence" of a polymer. We here reserve the widespread notation l_p for persistence in the tangent–tangent sense, rather than introducing an additional symbol such as l_T, merely because the tangent–tangent notion is perhaps the most commonly adopted meaning in the contemporary literature. Less trivially, it is important to note that for more realistic DNA free energies than the WLC, there is no *a priori* reason to believe that the dimensionless tangent–tangent persistence lengths $l_p(\mathbb{S})$ can be simply related to the Flory persistence lengths $l_F(\mathbb{S})$ via the introduction of a single length scale. Furthermore, it is well understood that there are some sequences with high intrinsic curvature, for example, those containing phased A-tracts,[11,13,39,40] for which the exponential fit in (6) to obtain $l_p(\mathbb{S})$ is an extremely poor approximation at scales of one or two persistence lengths or shorter (indeed, for some exceptional sequences of moderate length, $\langle \mathbf{t}_i\cdot\mathbf{t}_0\rangle$ can even become negative so that the semilog plot fit yielding $l_p(\mathbb{S})$ has no sense, in contrast to the more robust definition of $l_F(\mathbb{S})$). However, for "reasonable" (i.e., nonexceptional) sequence ensembles $\{\cdot\}$ it is believed that the sequence-averaged exponential fit to obtain $\overline{l}_p$ is a rather good approximation. Our numerical simulations will confirm these behaviors within the *cgDNA* coarse-grain model.

2.2. Sequence-Dependent Dynamic Persistence Length. To decompose the distinct effects on expectations due to the intrinsic shape of DNA and due to thermal fluctuation, Trifonov–Tan–Harvey[9] proposed the sequence-averaged relation

$$\frac{1}{\overline{l}_p} = \frac{1}{\overline{l}_s} + \frac{1}{\overline{l}_d} \tag{7}$$

Here the sequence-averaged tangent–tangent persistence length $\overline{l}_p$ is defined as before in eq (6), but hereafter, and following Trifonov–Tan–Harvey,[9] we introduce the additional adjective *apparent* sequence-averaged persistence length $\overline{l}_p$ when we wish to emphasize the decomposition (7). The

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539–1555

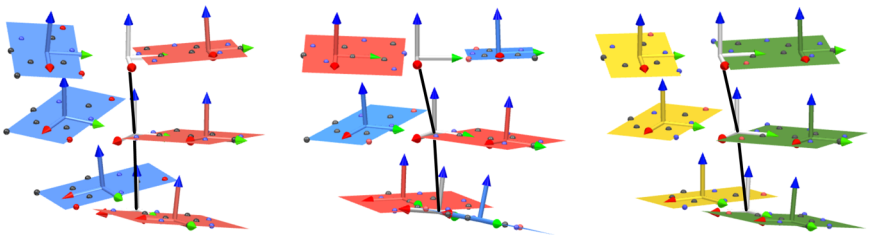


Figure 1. Visualization of three central base pairs in the *cgDNA* ground state configuration of three 20-mers: left, poly(A); middle, poly(TA); right poly(G). Each nucleotide is represented as a rigid body fit to base atoms that is visualized as a colored plate (A, red; T, blue; G, green; C, yellow) along with a base normal. The position and orientation of each base-pair frame (light gray) is an appropriate average of the two associated base frames (for visual clarity each base frame is offset by 0.35 nm toward its backbone from the standard Curves+ definition). The junction chords between the origins of adjacent base-pair frames are shown in black. Note that the poly(A) sequence has exceptionally high (propeller) intra base-pair rotations, and the junction chords are closely aligned with the base-pair normal, whereas for both poly(TA) and poly(G) there is a significant angle between the junction chords and associated base-pair normals.

sequence-averaged *static* persistence length $\overline{l}_s$ is defined via the fit

$$e^{-i/\overline{l}_s} \approx \{\hat{\mathbf{t}}_i \cdot \hat{\mathbf{t}}_0\} \tag{8}$$

where for each i, $\hat{\mathbf{t}}_i$ is the tangent evaluated on the ground state configuration for each sequence, and the average $\{\cdot\}$ is over all of the sequences in the ensemble. And $\overline{l}_d$ is a sequence-averaged *dynamic* persistence length that is to be estimated from (7). We note that in their original treatment Trifonov et al.[9] interpreted all three persistence lengths appearing in (7) within the context of a third classic expectation of polymer physics (in addition to the two introduced in (1)), namely, the mean square end-to-end distance function. However, we will consistently continue to use the tangent−tangent expectation $(1)_1$ to fit $\overline{l}_p$ as in (6) and the analogous (8) to fit $\overline{l}_s$. We make this choice because it allows the Trifonov−Tan−Harvey notion of sequence-averaged dynamic persistence length to be simply generalized to a *sequence-dependent* dynamic persistence length $l_d(\mathbb{S})$ (along with its sequence-averaged version $\overline{l}_d$) via the fits

$$\hat{\mathbf{t}}_i \cdot \hat{\mathbf{t}}_0 e^{-i/l_d(\mathbb{S})} \approx \langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle, \qquad e^{-i/\overline{l}_d} \approx \left\{ \frac{\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle}{\hat{\mathbf{t}}_i \cdot \hat{\mathbf{t}}_0} \right\} \tag{9}$$

where the prefactor $\hat{\mathbf{t}}_i \cdot \hat{\mathbf{t}}_0$ in the first expression is again evaluated on the ground state configuration, but now with no averaging because the sequence is prescribed. And to be explicit, $l_d(\mathbb{S})$ is obtained from the linear fit of a plot of $[\ln\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle - \ln(\hat{\mathbf{t}}_i \cdot \hat{\mathbf{t}}_0)]$ vs the index $i \in \mathbb{D}$. The simulations presented below indicate that both for the small number of MD simulations that we consider and for a wide range of *cgDNA* model MC simulations the quality of the exponential fit (9) is always better than the analogous fit in (6) and is remarkably good for nearly all sequences $\mathbb{S}$, with the only exceptions being relatively long fragments with highly bent ground states; cf. Figure 5. A simplified version of the three-dimensional formula for $l_d(\mathbb{S})$ in (9) appears in Schellman and Harvey,[13] with other more rudimentary planar versions having being used earlier by Théveny et al.[12] and later by Rivetti et al.[15] to interpret respectively classic EM and AFM data of DNA on a substrate. From our perspective, the definitions (6), (8), and (9) allow independent fits of all three of the sequence-averaged quantities $\overline{l}_p$, $\overline{l}_s$, and $\overline{l}_d$, from which we can evaluate the accuracy to which the Trifonov−Tan−Harvey relation (7) is satisfied. From this point of view, whenever $\{\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle / \hat{\mathbf{t}}_i \cdot \hat{\mathbf{t}}_0\} = \{\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle\} / \{\hat{\mathbf{t}}_i \cdot \hat{\mathbf{t}}_0\}$ $\forall$ i, then the Trifonov−Tan−Harvey relation simply reduces to an exact identity from properties of semilog linear fits.

Table 1 below summarizes the three sequence-dependent and four sequence-averaged notions of persistence length that have been introduced thus far. As we will see, they are all related, but distinct, quantities for realistic DNA models at the scale of hundreds of base pairs.

3. SIMULATION METHODS

In this section we make precise the coarse-grain variables that we will observe (section 3.1), describe the *cgDNA* model that predicts a free energy for a DNA fragment of arbitrary length and sequence (section 3.2), and explain the direct and Metropolis MC algorithms that we adopt to sample the equilibrium distribution implied by the *cgDNA* free energy (section 3.3).

3.1. Coarse-Graining and the Choice of Observables.

We will use Monte Carlo simulations applied to the *cgDNA* model of the free energies of a variety of different sequences to generate ensembles that yield numerical estimates of the expectation functions (1). For Flory vectors there is no reason to take the observables $\mathbf{r}_i$ and $\mathbf{R}_i$ as anything other than the *cgDNA* base-pair location and frame, after which the simulations are completely specified, and by evaluating the limits (5) over an ensemble they deliver sequence-dependent and sequence-averaged Flory persistence lengths directly in the dimensions of length with our units chosen to be nanometers. However, for the various tangent−tangent persistence lengths it remains to make precise the choice adopted for the unit vectors $\mathbf{t}_i$ in semilog fits of the forms (6) and (9), and also to introduce an appropriate arc-length parameter s_i to compare nondimensional persistence lengths given in units of numbers of bp with both simulated dimensional Flory persistence lengths and experimental data that reports other dimensional persistence lengths. In contrast to the WLC, because the *cgDNA* model encompasses fluctuations in the junction translations of shift, slide, and rise, there are several natural choices for both $\mathbf{t}_i$ and s_i. Moreover, though the *cgDNA* model can readily yield sequence-dependent expectations at the resolution of individual base pairs, some experimental data, for example, stereo cryo-EM, do not have this resolution, so that it is also of interest to compute ensemble expectations of a hierarchy of coarse-grained approximations of both $\mathbf{t}_i$ and s_i.

Concerning natural choices for the unit vectors $\mathbf{t}_i$, one possibility is the base-pair normal, i.e., the third column of each $\mathbf{R}_i$, or equivalently the frame vector most closely aligned with the helical axis (cf. Figure 1), which, as a matter of convention, will be denoted $\mathbf{t}_i^{[0]}$. Other natural possibilities are the unit tangents $\mathbf{t}_i^{[k]}$ (where for simplicity we only consider k odd) to

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

the straight lines that are the best least-squares linear approximation to a consecutive run of $(k + 1)$ base-pair locations $\mathbf{r}_i$, ..., $\mathbf{r}_{i+k}$ (and we associate $\mathbf{t}_i^{[k]}$ with the central junction in the run) where k is fixed and comparatively small; e.g., the cases $k = 1$, 9, 11 are of particular interest (a standard method to compute $\mathbf{t}_i^{[k]}$ is detailed in the Supporting Information). The case $k = 1$ reduces to the unit tangent to the junction chord between two consecutive base-pair origins $\mathbf{t}_i^{[1]} = (\mathbf{r}_{i+1} - \mathbf{r}_i)/\|\mathbf{r}_{i+1} - \mathbf{r}_i\|$ (shown in black in Figure 1). A comparison between expectations based on the base-pair normals $\mathbf{t}_i^{[0]}$ and the junction unit vectors $\mathbf{t}_i^{[1]}$ has previously been considered in Fathizadeh et al.[41] We will also consider nonlocal coarse-grain choices $k > 1$.

Just as there are different levels of coarse-grain approximations to the unit tangents $\mathbf{t}_i^{[k]}$, there are different coarse-grain approximations to the arc length. In particular, for each integer k, a coarse-grain arc length from base-pair 0 to base-pair i in any configuration can be computed as $s_i^{[k]} = \sum_{q=1}^{i/k} \delta s_{kq}^{[k]}$ where the coarse-grain arc length increment is taken as a step over k indices $\delta s_j^{[k]} := \|\mathbf{r}_{j+k} - \mathbf{r}_j\|$ (corrections for when the final index i is not an integer multiple of k are easily handled by a simple linear interpolation at one end). For large k the expectation $\langle \delta s_i^{[k]} \rangle$ is related to the expectation of the Flory vectors $(1)_2$, and any interpretation as an arc-length is nebulous at best. But for small k the different $\delta s_i^{[k]}$ give different coarse-grain approximations to different physically sensible notions of arc-length that may have significantly different values in highly coiled structures. The arc length $s_i^{[1]}$ is that of the piece-wise linear path traced by all base-pair origins $\mathbf{r}_j$, $j < i$, which, depending on the DNA sequence, may be significantly locally coiled, in which case $\delta s_j^{[10]}$ and $\delta s_j^{[11]}$ can be expected to be approximations to the pitch of the local helical structure close to base-pair j. And for experimental data coming for example from microscopy, it seems likely that the available observations will be of relatively coarse-grain notions, e.g., $k = 9$, 10, or 11, for both arc length and unit tangents.

Given an ensemble of configurations (either sequence averaged or not) it is then numerically straightforward to evaluate expectations such as $\langle s_i^{[k]} \rangle$ for each choice of i and k. In particular, if k indicates a chosen level of coarse graining of arc length $s_i^{[k]}$ at base-pair i, and j indicates a chosen level of coarse graining in the unit tangent $\mathbf{t}_i^{[j]}$, we may make a parametric plot of points $(\langle s_i^{[k]} \rangle, \ln\langle \mathbf{t}_i^{[j]} \cdot \mathbf{t}_0^{[j]} \rangle)$ indexed by i. A linear fit (through the origin) to this data then delivers a persistence length $l_{\mathrm{p}}^{[j,k]}(\mathbb{S})$ where the double superscript indicates the two (in principle independent) coarse-graining choices. As a matter of convention, $l_{\mathrm{p}}^{[j,0]}(\mathbb{S})$ will be taken to mean a persistence length for the coarse-grain tangent $\mathbf{t}_i^{[j]}$ expressed in bp, whereas for $k \geq 1$, $l_{\mathrm{p}}^{[j,k]}(\mathbb{S})$ has units of length, so that it can, for example, be immediately compared with the Flory persistence length $l_{\mathrm{F}}(\mathbb{S})$ for the same sequence. The same precise notions and notation carry over directly to all other tangent−tangent persistence lengths, e.g., a dynamic persistence length $l_{\mathrm{d}}^{[j,k]}(\mathbb{S})$.

To obtain reasonable sequence-averaged statistics from the limited number of images that are typically available in microscopy data (cf. section 6.3 below), and following Bednar et al.,[35] we will introduce one further additional variant of a sequence-averaged tangent−tangent persistence length defined via a sliding window approach. The sliding window persistence length $\overline{l}_{\mathrm{w}}^{[j,k]}$ is obtained from a fit to the ansatz

$$\mathrm{e}^{-\Delta_l / \overline{l}_{\mathrm{w}}^{[j,k]}} \approx \langle \mathbf{t}^{[j]}(s_i^{[k]} + \Delta_l) \cdot \mathbf{t}^{[j]}(s_i^{[k]}) \rangle_\delta, \qquad s_{(i+1)}^{[k]} = s_i^{[k]} + \delta \tag{10}$$

where $\{\Delta_l : l = 1, 2, ..., L\}$ is a range of sizes of windows, δ is a window shift, and (as before) $[j, k]$ denotes the coarse-grain choices for $\mathbf{t}^{[j]}$ and $s^{[k]}$. Here for each window size Δ_l, the ensemble average $\langle \cdot \rangle_\delta$ is over the index i of a sequence of window locations along a (possibly comparatively small) set of configurations of DNA oligomers, with data only taken appropriately far from either end, and with each window location $s_i^{[k]}$ separated from the next by the chosen increment δ in the coarse-grain arc length. We remark that as the scalar product is symmetric, each configuration can be read in either direction without changing the windowing ensemble average. As before, the symbol ≈ denotes a linear fit of a line through the origin to the logarithm of the data on the right-hand side of $(10)_1$, but now the fit is over a set of window sizes Δ_l; cf. Figure S7.

The definitions and notation for all of the sequence-dependent and sequence-averaged persistence lengths that we will evaluate are summarized in Table 1.

Table 1. Eight Distinct Notions of DNA Persistence Length[a]

	Flory (5)	apparent (6)	dynamic (9)	static (8)	window (10)
sequence dependent	$l_{\mathrm{F}}(\mathbb{S})$	$l_{\mathrm{p}}^{[j,k]}(\mathbb{S})$	$l_{\mathrm{d}}^{[j,k]}(\mathbb{S})$		
sequence averaged	$\overline{l}_{\mathrm{F}}$	$\overline{l}_{\mathrm{p}}^{[j,k]}$	$\overline{l}_{\mathrm{d}}^{[j,k]}$	$\overline{l}_{\mathrm{s}}^{[j,k]}$	$\overline{l}_{\mathrm{w}}^{[j,k]}$

[a]Notation for the different DNA persistence lengths that we compute, along with the numbers of each defining equation. Where appropriate, the superscript $[j, k]$ indicates the level of coarse graining assumed in the choice of tangents and arc-length (see text in section 3.1). Examples of sequence-dependent Flory persistence lengths are illustrated in Figure 4, and sequence-dependent apparent and dynamic persistence lengths are illustrated in Figure 5. Sequence-averaged persistence lengths are discussed in section 4.3, and window averaging is used in section 6.3.

3.2. *cgDNA* Rigid-Base Coarse-Grain Model. We will evaluate the diverse persistence lengths summarized in Table 1 on ensembles generated by Monte Carlo simulations applied to the sequence-dependent *cgDNA* coarse-grain model of DNA developed by Gonzalez et al.[32] and discussed in full detail in Petkevičiūtė et al.[33] (especially their Supporting Information). In the *cgDNA* model, each of the two DNA bases in the nth base pair are approximated (in a standard way respecting the Tsukuba convention[42]) as rigid bodies with location and orientation given by reference points $\mathbf{r}_n^{\pm}$ and orthonormal frames $\mathbf{R}_n^{\pm}$ fixed in each base (cf. Figure 1). The location and orientation of each base pair are prescribed by points $\mathbf{r}_n$ and orthonormal frames $\mathbf{R}_n$ that are appropriate averages of the rigid base quantities $\mathbf{r}_n^{\pm}$ and $\mathbf{R}_n^{\pm}$. Up to an overall rigid body motion, any coarse-grain configuration of the DNA is then described using standard internal helical coordinates.[42] Specifically, the relative configuration of bases within a base pair is determined by the *intra* base-pair parameters, comprising three translations (shear, stretch, and stagger) and three rotations (buckle, propeller, and opening), whereas the relative configuration of adjacent base pairs is determined by the *inter* base-pair parameters, comprising three translations (shift, slide, and rise) and three rotations (tilt, roll, and twist). These internal coordinates are then assembled into a vector $\mathbf{w} \in$

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

$\mathbb{R}^{(12N-6)}$ (in the alternating order intra-inter-intra) where N is the number of base pairs; given $\mathbf{w}$ along with the absolute position and orientation of a single frame, the coarse-grain DNA configuration is completely determined. It is, however, important to note that the reconstruction from the internal coordinates $\mathbf{w}$ to the observables $\mathbf{r}_n$ and $\mathbf{R}_n$ is highly nonlinear.

The *cgDNA* model prediction of the free energy of any configuration of a molecule of given sequence $\mathbb{S}$ with N base pairs is a shifted quadratic form in the internal helical coordinates $\mathbf{w}$

$$E(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^{\mathrm{T}}\mathbf{K}(\mathbf{w} - \hat{\mathbf{w}}) \tag{11}$$

where $\hat{\mathbf{w}}(\mathbb{S}) \in \mathbb{R}^{(12N-6)}$ describes the ground state of the molecule and $\mathbf{K}(\mathbb{S})$ is a $(12N - 6) \times (12N - 6)$ (positive definite, symmetric) stiffness matrix. The sequence dependence of both $\hat{\mathbf{w}}$ and $\mathbf{K}$ can be rather significant. For example, Figure 1 illustrates the central three base pairs in the *cgDNA* three-dimensional reconstructions of the ground states corresponding to the associated ground state vectors $\hat{\mathbf{w}}(\mathbb{S})$ for each of three sequence fragments poly(A), poly(TA), and poly(G), and the ground state structures can be seen to differ markedly. In general, the ground state vector $\hat{\mathbf{w}}(\mathbb{S})$ has nonlocal dependence on sequence due to the phenomenon of *frustration*.[32] For any sequence $\mathbb{S}$, the *cgDNA* stiffness matrix $\mathbf{K}(\mathbb{S})$ has the particular banded, sparsity structure illustrated in Figure 2, comprising overlapping 18×18 blocks that have 6×6 intersections. This sparsity pattern of $\mathbf{K}(\mathbb{S})$ corresponds to the model assumption that each base interacts directly with only its five nearest neighbors (its base-pair partner and those in the preceding and subsequent base pairs). Effects such as twist-bend coupling are captured by the model (because each 18×18 block is dense). The blocks of $\mathbf{K}(\mathbb{S})$ have local sequence dependence, specifically trinucleotide sequence dependence in the 6×6 overlaps, and dinucleotide dependence elsewhere.

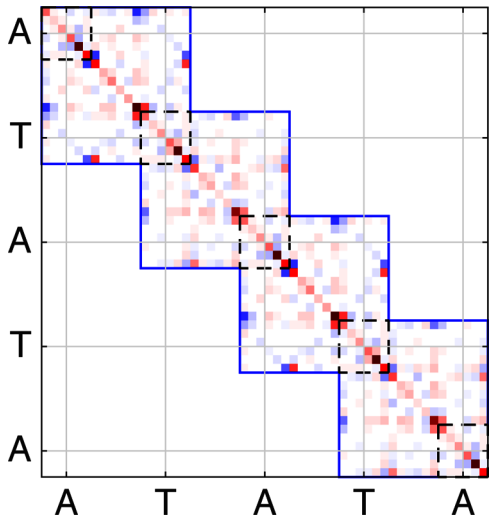


Figure 2. Visualization of the entries in a sub-block of the *cgDNA* model stiffness matrix K corresponding to five central base pairs for the specific sequence poly(TA). The banded structure with 18×18 overlapping blocks is a model assumption. The standard helical coordinates are ordered in alternating groups of six, intra-inter-intra, with the 6×6 overlaps corresponding to each set of intra base-pair variables. Individual entries in the stiffness matrix are represented in a color scale, with blue being large and negative, white representing vanishing or close to zero entries, and red and brown large positive entries. The largest entries are close to the diagonal, but the 18×18 blocks are dense with some large entries far from the diagonal indicating the significant couplings present in the model. Each 18×18 block corresponds to a dinucleotide junction or step, and the entries in an AT junction block can be seen to be slightly different to the entries in a TA block. Other sequences have stronger variations in the stiffness coefficients.

Both the *cgDNA* stiffness matrix and minimum energy shape for a DNA fragment of any specified base-pair sequence $\mathbb{S}$ can be explicitly constructed using freely available Matlab (or Octave) scripts that are downloadable from http://lcvmwww.epfl.ch/cgDNA, where the associated C++ Monte Carlo code *cgDNAmc*, which is described in the next section, is also freely available. Any *cgDNA* reconstruction and associated *cgDNAmc* simulations depend on a specific choice of a *cgDNA* model parameter set. All of the computations presented here use *cgDNAparamset2*, which is described in more detail in sections 5.4 and 6.4.

3.3. Monte Carlo Sampling. The ensemble expectation $\langle f \rangle$ of any function $f(\mathbf{w})$ of the *cgDNA* internal variables can be approximated as the simple average $\frac{1}{M}\sum_{j=1}^{M} f(\mathbf{w}_j)$ over a sequence of configurations $\mathbf{w}_j$ that is generated by a Monte Carlo method that appropriately samples the associated equilibrium distribution $p(\mathbf{w})\,d\mathbf{w}$. We will consider two specific cases of the probability density function, a pure Gaussian, or multivariate normal, and a perturbed Gaussian

$$p(\mathbf{w}) = \frac{1}{Z}e^{-\beta E(\mathbf{w})}, \qquad \tilde{p}(\mathbf{w}) = \frac{1}{\tilde{Z}}J(\mathbf{w})e^{-\beta E(\mathbf{w})} \tag{12}$$

where $E(\mathbf{w})$ is the shifted quadratic *cgDNA* energy (11), $\beta = 1/(k_{\mathrm{B}}T)$ is the inverse temperature scale, Z is the (explicitly known) normalization constant (or partition function), and $J(\mathbf{w}) > 0$ is an explicitly known function of $\mathbf{w}$, but now the value of the associated normalizing constant $\tilde{Z}$ is in general not known. There are several possible motivations for the generalization $(12)_2$, for example, modeling contributions to the *cgDNA* free energy from end-loading terms as in single molecule tweezer experiments, or modeling multiwell DNA backbone states as described in Pasi et al.[43] However, we focus here on a third motivation in which $J(\mathbf{w})$ is a Jacobian factor required[44−47] by the non-Cartesian nature of any rotational coordinates for the relative rotations between the frames $\mathbf{R}_i$ (and $\mathbf{R}_i^{\pm}$). In the scaled Cayley vector rotational coordinates adopted within the *cgDNA* model it can be computed explicitly that the appropriate configuration space equilibrium distribution is of the form $(12)_2$ with the explicit correction term

$$J(\mathbf{w}) = \prod_{i=1}^{2N-1}\left(1 + \frac{\zeta_i^2}{100}\right)^{-2} \tag{13}$$

where the ζ_i are norms of the intra- and inter-rotation parts of $\mathbf{w}$. Essentially, we here wish to be able to assess when the differences between the two pdfs in (12) (with the same sequence-dependent free energy $E(\mathbf{w})$) are sufficiently small that attention can be restricted to the simpler case $(12)_1$.

One approach to Monte Carlo simulation of multivariate normals such as $(12)_1$ involves the Cholesky decomposition of the covariance matrix.[48] We adapt this approach to take advantage of the sparsity structure of the stiffness matrix $\mathbf{K}$, performing the Cholesky decomposition on $\mathbf{K}$ itself:

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

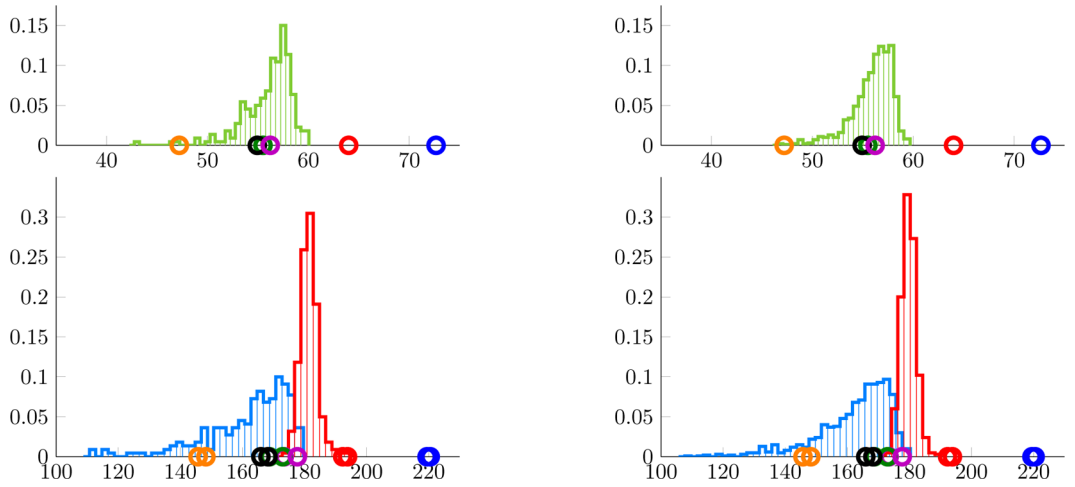


Figure 3. Normalized histograms of persistence lengths, $l_F(\mathbb{S}_j)$ in green (nm), $l_p(\mathbb{S}_j)$ in blue (bp), and $l_d(\mathbb{S}_j)$ in red (bp), for 220 bp fragments from λ-phage (left) and with random sequence (right). In addition, in each panel the associated persistence lengths for the six distinct poly(dinucleotide) sequences are marked with colored circles. The harmonic means of $l_F(\mathbb{S}_j)$ for the λ and random ensembles are respectively 55.7 and 55.6 nm, of $l_d(\mathbb{S}_j)$ are 181 and 179 bp, and of $l_p(\mathbb{S}_j)$ are 159 and 160 bp.

$$\mathbf{K} = \mathbf{L}\mathbf{L}^{\mathrm{T}} \tag{14}$$

where $\mathbf{L}$ is a lower triangular matrix. We can then use the Cholesky factorization to rewrite the *cgDNA* energy (11) as $E(\mathbf{y}) = \mathbf{y}^{\mathrm{T}}\mathbf{y}/2$, where $\mathbf{y} = \mathbf{L}^{\mathrm{T}}(\mathbf{w} - \hat{\mathbf{w}})$, and this distribution can be sampled directly as the product of uncoupled univariate normal distributions. For efficient Monte Carlo sampling in this way, the key property of the Cholesky factorization is that because our $\mathbf{K}$ matrix is highly banded (i.e., all nonzero entries are close to the diagonal) so is the Cholesky factor $\mathbf{L}$ (see, e.g., p. 154 of Golub and Van Loan[49]). For each sample, the configuration in the original variables $\mathbf{w}$ must first be reconstructed from the configuration $\mathbf{y}$. Then the observables $\mathbf{r}_n$ and $\mathbf{R}_n$ must also be computed from $\mathbf{w}$. As both of these computations occur at every draw, they should be done efficiently, and as described more fully in the Supporting Information, we make full use of the sparsity in the problem, as well as of quaternion multiplication in the many rotation matrix products. We believe the resulting code to be rather efficient; for example, for a 300 bp molecule 10^6 samples can be obtained in approximately 3 min on a contemporary laptop, and for a 1.5 Kbp fragment 10^6 samples can be computed in approximately 20 min, i.e., approximately linear scaling with oligomer length.

To sample the perturbed Gaussian distribution $(12)_2$, we adopt the following simple Metropolis algorithm.[50] Given a prior configuration $\mathbf{w}$, we draw a new $\mathbf{w}^*$ following the direct Monte Carlo procedure described in the previous paragraph applied to the Gaussian part of the pdf $(12)_2$. We then accept or reject $\mathbf{w}^*$ purely on the basis of the values of the perturbation J: if $J(\mathbf{w}^*) \geq J(\mathbf{w})$, we accept $\mathbf{w}^*$, whereas if $J(\mathbf{w}^*) < J(\mathbf{w})$ we accept $\mathbf{w}^*$ with probability $J(\mathbf{w}^*)/J(\mathbf{w})$ and otherwise reject it (in which case we append a new copy of $\mathbf{w}$ to our ensemble). As discussed in the Supporting Information, this move set satisfies the crucial property of *detailed balance* and does, in fact, sample the pdf $(12)_2$. The Metropolis procedure is computationally much more intensive than the direct sampling possible in the pure Gaussian case but is still reasonably efficient. For the specific perturbation (13) and a 300 bp fragment 10^6 accepted moves (with a 40% acceptance rate) can be generated in 11 min. For a 1.5 Kbp fragment the performance is 10^6 accepted (with a 4.5% acceptance rate) in 6 h. Because the acceptance criterion involves only the internal coordinates $\mathbf{w}$, the rejected moves absorb comparatively little computational time because the corresponding $(\mathbf{r}_n, \mathbf{R}_n)$ configurations need not be reconstructed.

4. SIMULATION RESULTS

In this section we present simulation data for the different persistence lengths summarized in Table 1. Section 4.1 demonstrates the strong sequence dependence that arises, whereas section 4.2 examines some specific sequences in more detail, including the quality of fits giving rise to the tangent–tangent persistence lengths $l_p(\mathbb{S})$ and $l_d(\mathbb{S})$. Sequence-averaged persistence lengths are discussed in section 4.3.

4.1. Sequence Is Significant: Persistence Length Spectra. The four panels of Figure 3 provide normalized histograms, or spectra, of the values of the individual Flory $l_F(\mathbb{S})$ (5), apparent $l_p(\mathbb{S})$ (6), and dynamic $l_d(\mathbb{S})$ (9) persistence lengths obtained from direct MC simulations of the Gaussian distribution $(12)_1$ for each of two ensembles of sequences, one being 1K random sequences of length 220 bp with equal probabilities for each of the four possible bases at each index i, and the other being 220 fragments of 220 bp of the λ-phage sequence. (We adopt the convention that $l_p(\mathbb{S})$ and $l_d(\mathbb{S})$ without superscripts denote $l_p^{[0,0]}(\mathbb{S})$ and $l_d^{[0,0]}(\mathbb{S})$; cf. Table 1.) For each of the selected sequences, the origin base-pair index 0 was chosen to be the 11th actual base pair from one end to avoid any initial end effects, and similarly, statistics were not taken from within 10 bp of the distal end. For the simulations of the Flory persistence length $l_F(\mathbb{S})$ to obtain good $i \to \infty$ convergence in the definition (5), sequence fragments of approximately 1.5 Kbp are needed, so each sequence was repeated seven times. The histograms indicate that there is strong sequence dependence of both $l_F(\mathbb{S})$ and $l_p(\mathbb{S})$ with at most small differences between the random and λ-phage ensembles, with perhaps a somewhat more prominent left tail (with many fewer samples) for λ. Both Flory distributions are quite broad and asymmetric, as are both $l_p(\mathbb{S})$ histograms, which have a notable and abrupt effective maximum close to

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

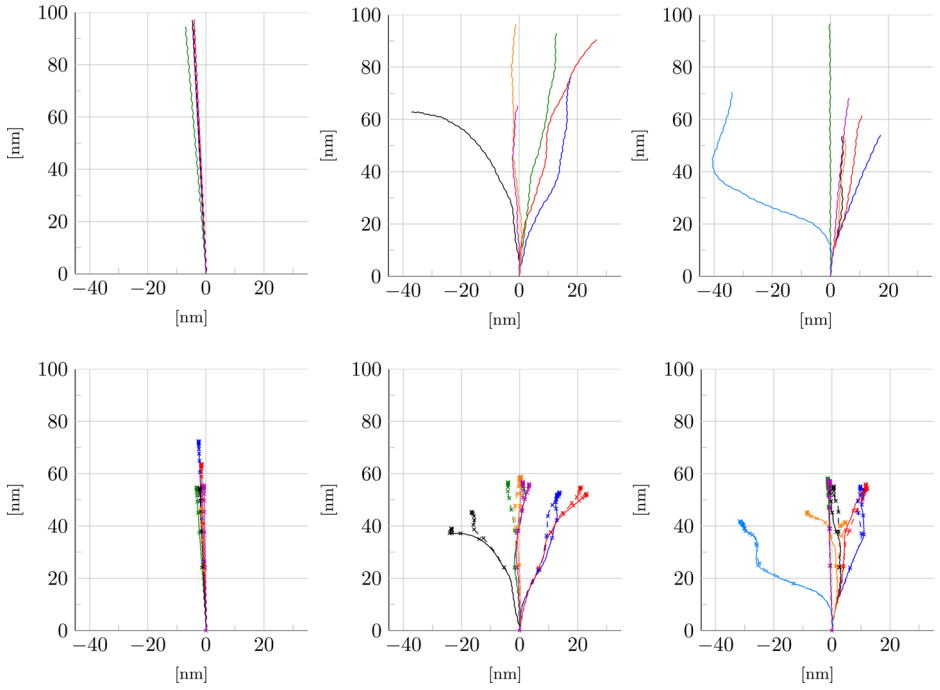


Figure 4. Ground state configurations and Flory persistence vectors for various DNA sequences. The columns show: (left) the six distinct poly dinucleotide sequences, (middle) the six selected λ-phage fragments λ_j, and (right) the seven sequences γ_j. The first row of panels shows visualizations of the shapes of *cgDNA* ground state configurations, whereas the second row shows plots of Flory persistence vectors $(1)_2$ for the Gaussian $(12)_1$ (solid) and perturbed $(12)_2$ (dashed) ensembles. (Interactive U3D versions of all six panels are available in Figure S1 of the Supporting Information.)

180 bp. In contrast, the histograms for the dynamic persistence length $l_d(\mathbb{S})$ are more sharply peaked and symmetric around 180 bp, which suggests at most a weak sequence dependence.

In each panel of Figure 3 the values of the associated persistence lengths for the six distinct poly dinucleotide sequences are also shown (as circles), and it is evident that for these particular sequences there is particularly strong sequence dependence of all three persistence lengths, with poly(AT) (or equivalently poly(TA)) being a low outlier with $l_F(AT)$ = 47.2 nm, $l_p(AT)$ = 146 bp, and $l_d(AT)$ = 148 bp, and poly(A) an exceptionally high outlier with $l_F(A)$ = 72.7 nm, $l_p(A)$ = 219 bp, and $l_d(A)$ = 221 bp.

4.2. Sequence Is Significant: Some Specific Cases. To better understand the behaviors manifested in the spectra of section 4.1, we now turn to a more detailed examination of a small number of selected sequences. The three columns of panels in Figures 4 and 5 provide data for three different groups of sequences: in the first column the six distinct poly dinucleotide sequences; in the second, six fragments λ_j from the λ-phage genome selected to span a wide range of observed $l_p(\lambda_j)$; and in the third, seven sequences γ_j with persistence lengths reported in the literature (full sequences are provided in the Supporting Information).

The panels in the first row of Figure 4 visualize the *cgDNA* ground states of each fragment with the reference base-pair frames $\mathbf{R}_0$ all aligned. For the polydinucleotides 300 bp are shown, and the shapes indicate that all six sequences have ground states that are tightly coiled helical structures with slightly different radii and pitches, and with straight center lines that are slightly differently aligned with respect to $\mathbf{R}_0$. All of the λ_j ground states (again 300 bp, except λ_6, 205 bp) are significantly bent, some more than others, as are sequences γ_{1-4}. In contrast, the ground states of γ_5, γ_7 are by design very straight, whereas the ground state of γ_6 has two large intrinsic bends each corresponding to a number of phased A-tracts.

The second row of panels in Figure 4 shows the Flory persistence vectors $(1)_2$ for the same sequences. The scales (in nm) on the axes of the panels in rows one and two are identical, but the Flory plots are for fragments of length 1.3−1.6 Kbp made of repeats of the basic sequence. In fact, the superimposed crosses on each curve indicate constant increments of 100 bp along each fragment, but the physical locations can be observed to accumulate, and the limiting distance from the origin to the accumulation point is the Flory persistence length of that sequence, which is considerably shorter than the end to end distance in the ground states shown in the first row of panels, and more importantly are quite different one from another. We emphasize that the Flory persistence vector curves are unrelated to any particularly likely configuration of the centerline of the corresponding DNA fragment, rather they "... mimic[s] the features of the static chain but with ever diminishing scale, until, when the last (976th) unit is added the change in position of the end point is imperceptible."[13]

Moving to Figure 5, the panels in the first row are plots of the ensemble average data $\ln\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle$ leading to the fit (6) for the persistence lengths $l_p(\mathbb{S}_j)$. The panels in the second row are plots of $\ln(\hat{\mathbf{t}}_i \cdot \hat{\mathbf{t}}_0)$ on the ground states (whose three-dimensional shapes are shown in the first row of Figure 4), and the panels in the third row are plots of the ensemble average data $[\ln\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle - \ln(\hat{\mathbf{t}}_i \cdot \hat{\mathbf{t}}_0)]$ leading to the fit (9) for the dynamic persistence lengths $l_d(\mathbb{S}_j)$.

For the polydinucleotide sequences in the first column, the data for the fits of both $l_p(\mathbb{S}_j)$ (row one) and $l_d(\mathbb{S}_j)$ (row three) are very close to linear, and the two persistence lengths are, in fact, very close for each sequence, but with strong variation

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

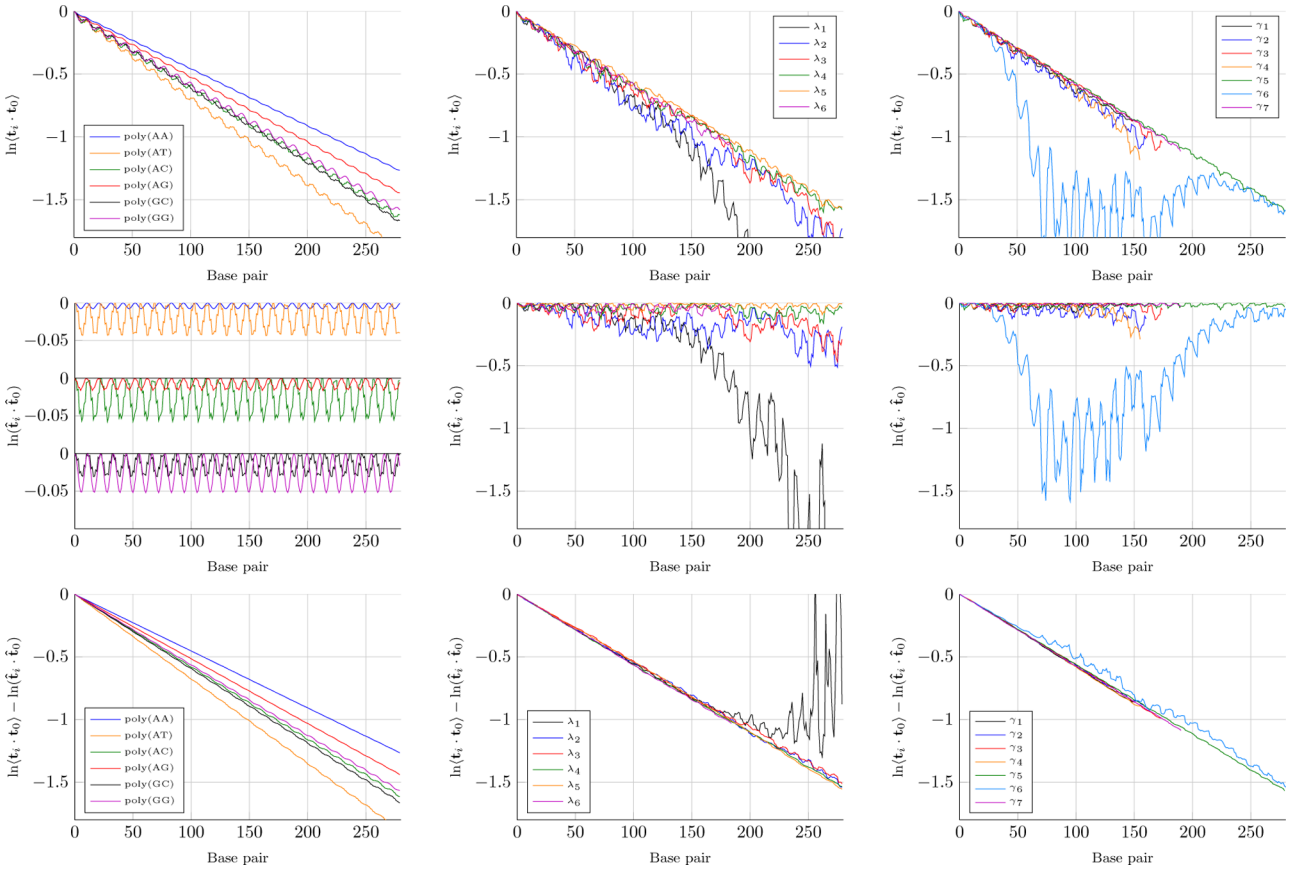


Figure 5. Plots related to apparent and dynamic persistence length for various DNA sequences. As in Figure 4, the columns show (left) the six distinct poly dinucleotide sequences, (middle) the six selected λ-phage fragments λ_j, and (right) the seven sequences γ_j. All nine panels are semilog plots of tangent−tangent data versus bp index for (row one) the apparent persistence length fit (6), (row two) the ground state, and (row three) their difference, for the dynamic persistence length fit (9). (In row two, column one, six small amplitude overlapping plots have been separated into three pairs for clarity.) Numerical values of the associated fits for persistence lengths are provided in Table S1 of the Supporting Information.

between sequences (cf. Figure 3). Nevertheless, the data in row one do exhibit significant short scale oscillations of different amplitude for the different sequences, which are almost entirely removed in row three via the implementation of our intrinsic-shape factorization; i.e., we subtract the data shown in row two.

For the selected λ_j sequences shown in column two, the plots in the first row are in some cases far from linear, leading to rather low estimates for $l_p(\mathbb{S}_j)$, but when the intrinsic shape is factored out, the data for the fit to $l_d(\lambda_j)$ are again close to linear, and now with rather little variation with sequence. The one exception is the sequence λ_1 shown in black, where the data for the fit for $l_d(\lambda_1)$ remain far from linear beyond bp 200. However, the sequence λ_1 was chosen as being the outlier with the lowest of all estimates for $l_p(\mathbb{S}_j)$ over all 161 300mer fragments drawn from the λ-phage sequence. And even for this extreme case, the data for the fit of $l_d(\lambda_1)$ are very close to linear if only the first 200 bp are used. In fact, we changed from the 220 bp fragments considered in the histograms of Figure 3 to be able to give this single example where the data for the fit (9) failed to be close to linear. This example suggests that the ansatz (9) for the factorization of the effects of shape is too simple for relatively long sequences with highly bent ground states.

For the sequences γ_j shown in the third column, the observed behavior is analogous to that of λ_j fragments, but now even the extremely bent sequence γ_6 (which contains phased A-tracts that are discussed further in section 6.2) has the acceptably linear data shown in row three, and the estimate for $l_d(\gamma_6)$ is quite close to that for the other γ_j and the λ_j, suggesting that its unusual tangent−tangent correlation plot in row one is due largely to its intrinsic shape shown in row two and not from any particularly unusual stiffness.

4.3. Sequence-Averaged Persistence Lengths. Table 2 provides estimates for $\overline{l}_F$, $\overline{l}_p^{[j,k]}$, $\overline{l}_s^{[j,k]}$, and $\overline{l}_d^{[j,k]}$ from evaluation of the sequence ensemble formulas in (5), (6), (8), and (9) for both of our examples (namely, random 220 bp fragments, and the collection of 220 fragments of 220 bp λ-phage fragments described above), and in each of the two cases $[j, k] = [0, 0]$ and $[11, 11]$; cf. Figure S6. For the random ensemble, we sample sufficient sequences and MC configurations for each sequence, to produce standard errors below 1 bp/0.1 nm for $\overline{l}_F$, $\overline{l}_p$, and $\overline{l}_d$ (by sampling 10^5 configurations for each of 1000 random sequences), and below 10 bp/1 nm for $\overline{l}_s$ (by considering intrinsic shapes of 10^5 random sequences). In

Table 2. Sequence-Averaged Persistence Lengths[a]

$\overline{l}_F$ = 53.4/**53.5** nm	$\overline{l}_p$	$\overline{l}_d$	$\overline{l}_s$
$[j, k] = [0, 0]$, bp	160/**160**	180/**178**	1442/**1610**
$[j, k] = [11, 11]$, nm	53.4/**53.5**	59.4/**58.8**	535/**609**

[a]For both λ and random sequence ensembles, with random sequence data in bold type.

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

contrast, λ-phage is a fixed sequence, which we have chosen to divide into certain consecutive 220 bp fragments; we draw sufficient MC samples to produce the same small standard error for the average over that particular set of fragments, but this does not guarantee the same small variation over different choices of λ-fragments.

These single ensemble estimates of $\overline{l}_F$, $\overline{l}_p^{[j,k]}$, and $\overline{l}_d^{[j,k]}$ are close to the appropriate averages of the histograms of the sequence-dependent quantities illustrated in Figure 3 for $[j, k] = [0, 0]$, and the analogous $[j, k] = [11, 11]$ histograms shown in Figure S5. For both choices of the coarse graining, the set of three independent estimates of the random-ensemble sequence-averaged persistence lengths $\overline{l}_{p,s,d}$ in Table 2, satisfy the Trifonov−Tan−Harvey relation (7) to a relative error of 3×10^{-3}, i.e., $|1 - \overline{l}_p/\overline{l}_s - \overline{l}_p/\overline{l}_d| < 3 \times 10^{-3}$.

5. SIMULATION VERIFICATION

We now present various additional numerical experiments as evidence to suggest that the conclusions about the sequence-dependent mechanics of DNA drawn from the numerical simulations described in section 4 are, in fact, robust.

5.1. Sensitivity to the Jacobian Perturbation. Ensembles for all of the sequences included in Figures 4 and 5 were generated with both direct and Metropolis MC simulations to assess differences between the Gaussian ensemble $(12)_1$ and the perturbed distribution $(12)_2$ in the case of the appropriate Jacobian factor (13) for our coordinates on the rotation group. The solid and dashed curves in row two of Figure 4 show that there are perceptible, but relatively small, differences between the two curves of Flory-vector expectations, with the differences increasing with the base-pair index i. The corresponding differences in the two estimates for the Flory persistence lengths with $i \approx 1.5K$ are nevertheless usually, although not always, close to the MC sampling error. For estimates of the tangent−tangent correlations $l_p(\mathbb{S}_j)$ and $l_d(\mathbb{S}_j)$ for sequences up to 200−300 bp we found the corresponding differences always to be negligible. (Figure S2 provides two examples showing differences between tangent−tangent correlation data for the two ensembles that are small, but perceptible, and accumulating with base-pair index.) We accordingly conclude that although the effect of the Jacobian perturbation to the equilibrium ensemble deserves further investigation in the case of long segments of DNA, it has a negligible influence on the computation of tangent−tangent persistence lengths at the scale of 200 bp, which is why all further calculations did not include it.

5.2. Convergence of MC Simulations. Irrespective of sequence, estimates of the Flory persistence vectors appeared to be converged to a standard error of less than 0.5 nm for fragments of 1.5 Kbp for multiple estimates each with 10^5 MC samples. Similarly, in the computation of tangent−tangent correlations of 300 bp fragments, 10^5 MC samples give a standard error of less than 1 bp or 0.5 nm in l_p (cf. Figure S3). For Metropolis MC simulations, longer runs are required for the same level of accuracy: 10^6 accepted samples for l_p and 3×10^6 for l_F. All our reported values for single-molecule l_p, l_F, or l_d come from samples that meet or exceed these requirements.

5.3. Sensitivity to Coarse-Graining Choices. For the Flory persistence length data reported in Figure 3 there is no choice of coarse graining to be made. The tangent−tangent data illustrated in Figures 3 and 5, and reported in Table S1 of the Supporting Information, have been for $l_p^{[0,0]}$, i.e., for base-pair normals $\mathbf{t}_i^{[0]}$ with persistence lengths reported in units of bp. As reported in Table S2, and independent of sequence, we found quite consistent estimates for both $l_p^{[j,0]}$ and $l_d^{[j,0]}$ for the differing coarse-grain choices $j = 0$, 11, 21 of the approximation to the unit tangent vector $\mathbf{t}^{[j]}$, with the larger j (unsurprisingly) yielding a decrease in the amplitude of the short scale oscillations in the correlation data. The other natural choice $j = 1$, corresponding to the junction chord vector, yields much larger amplitude short-scale oscillations and a tendency for the apparent best-fit line to pass significantly below the origin (cf. Figure S4).

In contrast, in different sequences, the choice of coarse-grain arc length $s^{[k]}$ has different consequences for the estimation of the dimensional tangent−tangent persistence lengths $l_p^{[0,k]}(\mathbb{S})$. Specifically for highly coiled sequences, the differences between $k = 1$ and $k = 11$ can be quite significant. Table 3 provides data for the highly coiled poly(TA) sequence, and the very straight poly(A). For poly(TA) the decrease between $l_p^{[0,1]}$ and $l_p^{[0,11]}$ is much bigger than for poly(A). Subsequent differences for $k > 11$ appear to always be negligible (cf. Table S2 and Figure S4). Similarly, as the sequence poly(A) with the highest of all l_d is very straight, whereas the sequence poly(TA) with the lowest l_d is highly coiled, the data of Table 3 show that the ratio 1.49 of the $l_d^{[0,0]}$ for the two sequences increases to the ratio 1.54 between the two $l_d^{[0,11]}$ for the coarse-grained arc length.

Table 3. Effects of Coarse Graining of Arc Length[a]

	$l_p^{[0,1]}$, nm	$l_p^{[0,11]}$, nm	$l_d^{[0,0]}$, bp	$l_d^{[0,11]}$, nm
poly(TA)	50.7	45.9	148	46.7
poly(A)	74.6	71.6	221	72.0

[a]Persistence lengths for the highly coiled poly(TA) are more sensitive to the choice of arc length coarse graining than for the very straight poly(A).

Further comparisons between coarse-graining choices are provided in Table S2, but we now concentrate on the two choices $[j, k] = [0, 0]$ of base-pair normals with persistence lengths reported in bp, and $[j, k] = [11, 11]$, with both tangent and arc length coarse grainings matched to a single turn of the double helix, and with persistence lengths reported in nanometers, which seems to be most appropriate for comparison with microscopy data.

5.4. Comparison between MC and MD Ensemble Expectations. The utility of a predictive, computationally efficient, coarse-grain model is that it allows large ensembles of configurations to be generated with comparatively little computational effort, both for long DNA fragments and for a large variety of sequences. However, the accuracy of the predictions made from these ensembles can of course be no better than the model parameter set. Each parameter set for the *cgDNA* coarse-grain model is constructed by fitting to estimates for expectations and covariances obtained from MD time-series simulations of a library of oligomers at the length scale of 10−20 bp, as described in detail in Gonzalez et al.,[32] and so depends upon the specific underlying MD simulation through both the choice of the physical conditions and the accuracy of the MD simulation package. Each *cgDNA* parameter set contains 1592 independent numbers, which is of course many more than the two parameters of the classic WLC, but also many fewer than in the underlying MD potentials. The estimation of a full parameter set is in and of itself a nontrivial process. In particular, because *cgDNA* stiffness matrices are

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

banded, with the form illustrated in Figure 2, their inverses are dense. Consequently, there is no simple and direct connection between specific entries in the stiffness matrix and specific entries in an associated covariance matrix. In the estimation of a *cgDNA* parameter set, all of the 1592 unknowns are coupled one to another.

cgDNAparamset2, which is used for all the simulations presented here, was derived in a systematic fitting procedure from fifty-three 50−100 ns duration MD simulations at 300K with K^+ and Cl^- ions at approximately physiological concentration, run using the Amber simulation package[51] with the specific ABC bsc0 protocol as described in detail in Lavery et al.[52] As discussed in two articles by Gonzalez et al.,[32,53] once a specific training set of MD time series has been generated, the resulting *cgDNA* parameter set further depends upon mathematical choices that have to be made in the detailed coarse-grain parameter extraction procedure. A comparison between various *cgDNA* predictions of ground states and experimental data has already been made.[33] The *cgDNA* parameter estimation procedures contain no explicit fit to the persistence length of DNA at any stage. Nevertheless, it is reasonable to pose the question of whether the parameter set estimation process might introduce a systematic bias in the persistence lengths predicted from *cgDNAmc* simulations. In this section we present data to suggest that any such bias in *cgDNAparamset2* is rather small, and that the *cgDNAmc* predictions of persistence length are of similar accuracy to predictions that can be made directly from comparable, but computationally much more intensive, MD simulations.

To justify this claim, we compared ensemble average data for eq (6) (leading to estimates for the sequence-dependent apparent persistence length $l_p(\mathbb{S})$) and data for eq (9) (leading to estimates for the dynamic persistence length $l_d(\mathbb{S})$) derived from both coarse-grain *cgDNA* MC simulations and from 10^6 snapshots generated in microsecond duration fine-grain MD simulations[43] of 18 bp oligomers with the sequences GC-$(XY)_7$GC, i.e., six independent oligomers with seven central repeats of each of six distinct dimer steps. The MD simulations were run with precisely the same protocol as the shorter duration ones used to train *cgDNAparamset2*. The MD data estimates for $\hat{\mathbf{t}}_i$ were obtained by extracting Curves+[42] base-pair coordinates for each MD snapshot, averaging the coordinates and reconstructing the configuration corresponding to the averaged coordinates. After dropping three bp from each end (both to minimize end effects and to suppress the different sequence effect of the terminal GC steps, while still leaving a minimal number of data points), we obtain 11 nontrivial values for $\ln\langle \mathbf{t}_i\cdot\mathbf{t}_0\rangle$ and $[\ln\langle \mathbf{t}_i\cdot\mathbf{t}_0\rangle - \ln(\hat{\mathbf{t}}_i\cdot\hat{\mathbf{t}}_0)]$ for each of the six sequences, and for each of the MC and MD ensembles. The data for the two cases XY = AA, TA are plotted in Figure 6 (which is a short length scale version of Figure 5 for the two specific polydinucleotide sequences). The numerical data for all six cases are provided in Tables S3 and S4 of the Supporting Information. It can be seen that the difference between MC and MD ensemble data is rather small compared to the differences between the two sequences. It can also be seen that the $\ln\langle \mathbf{t}_i\cdot\mathbf{t}_0\rangle$ data deviate strongly from linear to the extent that making a linear best fit to estimate $l_p(\mathbb{S})$ is rather questionable, particularly for the TA sequence. This strong deviation from linearity of $\ln\langle \mathbf{t}_i\cdot\mathbf{t}_0\rangle$ for short scale MD data has certainly been observed before, for example, by Noy and Golestanian,[30] and the sense of apparent persistence length $l_p(\mathbb{S})$ at these scales

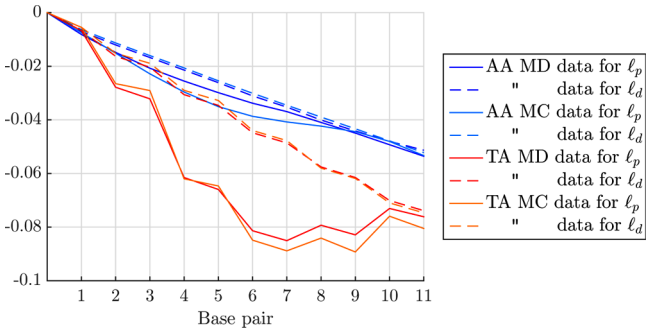


Figure 6. In solid lines the data for $\ln\langle \mathbf{t}_i\cdot\mathbf{t}_0\rangle$ vs base-pair index i = 1, ..., 11 for the TA sequence in orange and red, respectively, for coarse-grain MC and fine-grain MD ensemble averages. And for the AA sequence in light and dark blue, again, respectively, for MC and MD ensemble averages. Dashed lines with the same color scheme are analogous plots of $[\ln\langle \mathbf{t}_i\cdot\mathbf{t}_0\rangle - \ln(\hat{\mathbf{t}}_i\cdot\hat{\mathbf{t}}_0)]$.

has been called into question. In contrast, it has not been previously observed that the data $[\ln\langle \mathbf{t}_i\cdot\mathbf{t}_0\rangle - \ln(\hat{\mathbf{t}}_i\cdot\hat{\mathbf{t}}_0)]$ leading to the estimate of the sequence-dependent dynamic persistence length $l_d(\mathbb{S})$ are still very close to linear (at least for all six polydinucleotide sequences) even at these short length scales, suggesting that at short lengths the observed values of $\ln\langle \mathbf{t}_i\cdot\mathbf{t}_0\rangle$ are dominated by the intrinsic shape of the DNA and not by stiffness and associated fluctuations. The values of the available estimates of $l_p(\mathbb{S})$ and $l_d(\mathbb{S})$ for all six polydinucleotide fragments at both long and short length scales are provided in Table 4. The comparison between MD and MC ensemble

Table 4. Apparent and Dynamic Persistence Lengths for Six poly(XY) Fragments[a]

	MD	MC	
$l_p^{[0,0]}/l_d^{[0,0]}$	12 bp	12 bp	280 bp
AA	193/203	186/205	219/221
AG	132/190	153/184	192/194
GG	104/209	110/167	173/178
TG	113/165	106/163	169/173
CG	143/189	122/160	166/168
TA	106/143	102/144	146/148

[a]Estimates of the apparent $l_p^{[0,0]}(\mathbb{S})$ (6) and dynamic $l_d^{[0,0]}(\mathbb{S})$ (9) persistence lengths made at the scale of 12 bp from both MD and MC ensembles, and for comparison from MC ensemble data for 280 bp fragments as illustrated in Figure 5a. The estimates for dynamic persistence length are quite consistent between MC and MD ensembles and at different length scales, whereas the estimates for apparent persistence length between short and long lengths are inconsistent because a linear fit is being made to highly nonlinear data.

estimates is better for some sequences than for others, but we see in particular that the coarse-grain *cgDNAmc* prediction that poly(A) has approximately 50% greater dynamic persistence length than poly(TA) is already contained in the fine-grain, short length scale MD data ensemble.

6. COMPARISON WITH EXPERIMENT

We next compare *cgDNAmc* predictions against 2D microscopy and cyclization data for a range of sequences (section 6.1), experimental and simulation data for A-tracts (section 6.2), and stereo cryo-microscopy images (section 6.3). Possible explan-

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

ations for discrepancies between simulation and experiment are discussed in section 6.4.

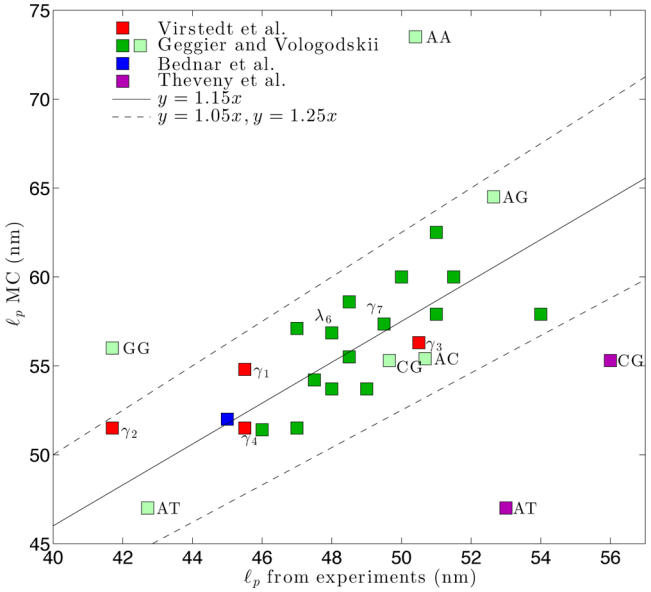


Figure 7. Scatter plot comparison of simulated and experimentally estimated persistence lengths, with differing experimental data grouped by color. See main text for details, including a description of three data points that are off scale in this plot. The three straight lines pass through the origin with slopes 1.05, 1.15, and 1.25.

6.1. 2D Microscopy and Cyclization Data. Figure 7 is a scatter plot of *cgDNAmc* predictions versus experimentally observed apparent persistence lengths reported in the literature, colored according to different groupings of sequence and experiment. More specifically, each point is for a specific sequence that has as ordinate its simulated value of $l_p^{[11,11]}$ (which we believe to be the appropriate level of coarse graining for these data), and as its abscissa an experimental value reported (a) from AFM for sequences γ_1–γ_4[1] (red), (b) from classic EM (indigo) for poly(TA) and poly(CG) with two additional data points (135, 73.5) for poly(A) and (120, 56.0) for poly(G) off scale to the right[12], (c) from cyclization data for 16 sequences detailed in the Supporting Information (one of which lies directly under the red box for sequence γ_4), and the six distinct polydinucleotides[34] (dark and light green, respectively), and (d) from a 3D cryo-EM and window averaging approach for average λ-phage DNA, and with one additional point (82, 58.3) out of figure to the right for the sequence γ_5[35] (blue).

Virstedt et al.[1] chose the sequences γ_1, γ_2, and γ_4 for their believed enhanced flexibility, with γ_3 serving as a control sequence with no distinguishing feature, and indeed its persistence length is found to be the highest of the four for both simulation and experiment; the simulated and experimental data correlate quite well, with Pearson correlation coefficient $r = 0.82$, although MC estimates are on average 17% higher. Similarly, the four persistence lengths reported by Théveny et al.[12] have correlation coefficient $r = 0.80$ with our simulated data, now with the MC estimates on average 28% lower. Both Virstedt et al.[1] and Théveny et al.[12] extract an estimate for l_p from correlations of the 2D angle between tangents to the planar curves observed in their data, but we note that they fit to different formulas. The issue (as fully discussed in these two papers as well as in Bednar et al.[35]) is that the exponential decay scale l_p in the three-dimensional WLC formula (3) changes to the exponential decay scale $2l_p$ in dimension two. And depending upon the precise sample preparation technique some authors prefer to fit their 2D microscopy data to the 2D decay rate $2l_p$, e.g., Virstedt et al., and some to the 3D decay rate, e.g., Théveny et al. The Pearson correlation coefficient is insensitive to this difference, whereas the average certainly is not. An additional issue in comparing 2D EM techniques to 3D simulations is the potentially complicated interaction between a sequence-dependent 3D ground state configuration, e.g., the highly coiled poly(AT) sequence, and the planar substrate.

Geggier and Vologodskii[34] report persistence lengths using experimentally measured minicircle cyclization *j*-factors[54] for a training set of oligomers that were then fit to an analytical formula for *j* within a certain helical WLC model.[55] The fit is over three free parameters for helical repeat, torsional stiffness, and bending stiffness, which they identify with l_p. In particular, the Geggier and Vologodskii sense of persistence length is as in the WLC free energy, i.e., a scalar, bending stiffness at each junction, but now dependent on the overall sequence of the fragment. For these 16 sequences, the Pearson correlation coefficient between *cgDNAmc* and Geggier and Vologodskii estimates is $r = 0.73$, with our Monte Carlo estimates on average 15% higher. With an additional fitting step to their own data, Geggier and Vologodskii construct a parameter set that can predict a persistence length for any sequence, which is how we obtained the "experimental" data for the six distinct polydinucleotide sequences (none of which were actually directly measured experimentally by Geggier and Vologodskii). We note that although their model reports a stiffness for poly(A), Geggier and Vologodskii explicitly exclude the poly(A) sequence from their actual cyclization experiments due to the exceptional properties of A-tracts. The Pearson correlation coefficient for these six sequences is $r = 0.60$, with *cgDNAmc* estimates on average 22% higher, and the poly(A) estimate notably larger.

6.2. Exceptional Behaviors of poly(A) and of A-Tracts. The values of persistence lengths for the two homopolymers poly(A) and poly(G) computed by Olson et al.[19] (reported in their Table 14.1 for both "Refined" and "Complete" samples) via a MC simulation within a rigid base-pair model parametrized from crystal-structure data,[18] yield four further comparison points in Figure 7, but now of simulation vs simulation, which would lie on the left axis for the "Refined" and off scale to the left for "Complete" data, specifically for poly(G) (40.5, 56.0) or (29.8, 56.0), and for poly(A) (39.5, 73.5) or (15.0, 73.5). In particular, the Olson et al. model predicts poly(A) to be one of the softest sequences, whereas *cgDNAmc* predicts it to be the single stiffest sequence. Of data presented here we have for supporting poly(A) being unusually stiff, the ratio of persistence lengths poly(A)/poly(AT) being reported as 135/53 = 2.5 by Théveny et al.[12] and 50.4/42.7 = 1.18 by Geggier and Vologodskii,[34] as compared to the prediction 73.5/47.0 = 1.56 from *cgDNAmc* simulations, as well as recent NMR data[56] reporting A tracts to be stiffer than average. On the contrary, on the basis of looping data, Johnson et al.[57] suggest that A-rich sequences are softer than average.

Resolving the statistical mechanics properties of DNA fragments with runs of consecutive A in the sequence is of significance as the properties of A-tracts continue to be of interest,[58] and in particular they are believed to play a central

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

role in nucleosome positioning.[3,59] An A-tract is generally taken to mean a short run of consecutive A_n, usually with $n = 5$ or $n = 6$, followed by another short spacer sequence, so for example, the sequence γ_6,[39] which is one of the sequences considered in section 4.2, contains two 63 bp regions with triple repeats of the 21 bp fragment A_6CGGCA_6CGGGC, which would be described as being two regions each with six phased A-tracts. DNA fragments including multiple repeats of precisely the same 21mer have also been examined experimentally more recently.[40] We shall therefore briefly analyze such sequences in more detail using the *cgDNA* model.

We first remark that although the 18mer $GCA_{14}GC$ (as described in section 5.4) is an approximation to poly(A) that does appear as one of the 53 sequences in the MD simulation library used to train *cgDNAparamset2*; there is no A-tract appearing anywhere in that library. Nevertheless, the *cgDNA* rigid base model is sufficiently rich to predict that whereas a long poly(A) is both exceptionally straight and stiff, an A-tract can induce a strong local bend in the ground state. This strong distinction can be seen by comparing the strikingly different data for poly(A) compared to the A-tract sequence γ_6, both of whose ground states and Flory vectors are shown in Figure 4, along with their tangent−tangent correlations in Figure 5. The only other sequence among those described there that is approaching being as bent as γ_6 is λ_1 which, as remarked earlier, was chosen as being the outlier with the lowest of all estimates for $l_p(\mathbb{S}_j)$ over all 161 300mer fragments drawn from the λ-phage sequence. As it happens, the 300 bp λ_1 fragment contains three instances of A_6, although that was not explicitly used as part of our selection criterion.

The fact that A-tracts have exceptional properties is experimentally indisputable, but whether it is the run of A_n itself, or the adjacent spacer region, or the juxtaposition of the two that generates these exceptional properties has been the subject of a long-running and ongoing debate.[60] The estimation procedure for the *cgDNA* parameter sets is in no way explicitly biased to model A_n runs differently from any other sequence, but *cgDNAparamset2* predicts that they give rise to the localized structure visualized in Figure 1 with exceptionally high values of propeller. One could therefore imagine significant disruption of the DNA double helical structure at a transition to other sequences. The *cgDNA* prediction of ground states for various A-tract junctions are described in more detail in Petkevičiūtė et al.[33] The properties of phased A-tracts are further described in Figures 8 and 9. Figure 8 provides the data for estimating apparent and dynamic persistence lengths of the central 63 bp $(A_6CGGCA_6CGGGC)_3$ within the full sequence $(A_6CGGCA_6CGGGC)_5$. The data for fitting the apparent and dynamic persistence lengths are each shown for three levels of coarse graining, $k = 0, 3, 11$, for the tangent vectors $\mathbf{t}_i^{[k]}$. The coarse-graining choice $k = 3$ is included because it fits a tangent vector to a run of four consecutive base-pair origins, so that for the appropriate i it fits a tangent to the center of each A_6 run, and to the center of each spacer sequence. In each case the data for $l_p^{[k,0]}(\mathbb{S})$ are far from linear with large oscillations associated with each A-tract, whereas in each case the data for $l_d^{[k,0]}(\mathbb{S})$ are still rather close to linear. The associated values are reported in Table 5. For each level of coarse graining the apparent persistence lengths $l_p^{[k,0]}$ (to the extent that they have any sense) are extremely low and are off scale to the left in the histograms of Figure 3 due to the large intrinsic bends, whereas the dynamic persistence lengths $l_d^{[k,0]}$ are exceptionally high and approach that of poly(A).

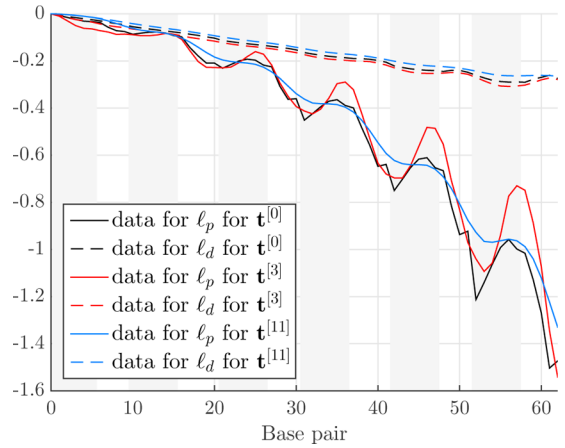


Figure 8. Tangent−tangent correlation data with three levels of tangent coarse graining, $k = 0, 3, 11$, for the fragment $(A_6CGGCA_6CGGGC)_3$, with the A_6 runs highlighted by centered shaded strips of width six. The highly nonlinear solid lines are plots of $\ln\langle \mathbf{t}_i^{[k]} \cdot \mathbf{t}_0^{[k]} \rangle$ leading to the estimates of the apparent persistence lengths $l_p^{[k,0]}(\mathbb{S})$ reported in Table 5, whereas the almost linear dashed lines are plots of $[\ln\langle \mathbf{t}_i^{[k]} \cdot \mathbf{t}_0^{[k]} \rangle - \ln(\hat{\mathbf{t}}_i^{[k]} \cdot \hat{\mathbf{t}}_0^{[k]})]$ leading to the estimates of the dynamic persistence lengths $l_d^{[k,0]}(\mathbb{S})$, also reported in Table 5.

As extensively reviewed by Brunet et al.,[40] there are many numbers reported in the literature for estimates of the bend of an A-tract. A full analysis of A-tract bends within the *cgDNA* model is beyond the scope of the current presentation, but we remark that, analogously to the plots in Figure 8, *cgDNAmc* can easily simulate estimates of expected angles in the sense of $\arccos\langle \mathbf{t}_i^{[k]} \cdot \mathbf{t}_0^{[k]} \rangle$ or more generally $\arccos\langle \mathbf{t}_i^{[k]} \cdot \mathbf{t}_j^{[k]} \rangle$, or true expected angles in the form $\langle \arccos(\mathbf{t}_i^{[k]} \cdot \mathbf{t}_j^{[k]}) \rangle$. We instead limit ourselves to providing plots in Figure 9 which further demonstrate that although the ground state of poly(A) is very close to straight (as already indicated in Figures 4 and 5), the ground state of our phased A-tract sequence is highly bent. However, quantifying the amount of that bend is quite sensitive to how you wish to measure it, both to the scale at which tangents $\mathbf{t}_i^{[k]}$ are fit to base pairs and to the number of base pairs separating the two tangents between which the angle is measured, so that the value of an intrinsic (or ground state) bend angle could reasonably be taken within a wide range of values.

6.3. 3D Cryo-electron Microscopy. In an effort to distinguish between the effects of intrinsic curvature and flexibility on tangent−tangent correlation, Bednar et al.[35] used stereo cryo-EM data to estimate the persistence lengths of the fragment γ_5 mentioned above. They designed this sequence to be intrinsically straight by a construction involving repeated pentanucleotide sequences so that any local intrinsic bend will to a good approximation be canceled by an oppositely phased bend 5 bp, or approximately one-half-turn, later. (In fact, the γ_5 sequence is made of repeats of a 20mer that is itself made up of double repeats of two different pentamers.) The accuracy within the *cgDNA* model of this clever sequence design is confirmed by the visualization of the very straight ground states shown in Figure 4 of γ_5 and γ_7 (which was introduced in Geggier and Vologodskii[34] using the same design principle). Bednar et al. also analyzed segments of the λ-phage genome as a control.

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

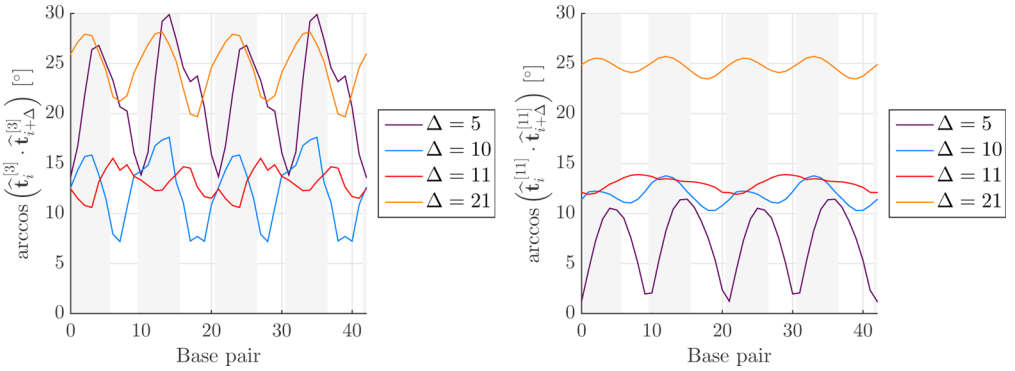


Figure 9. For the ground state of the A-tract fragment described in Figure 8. Left panel, plots of $\hat{\Theta}_{i,\Delta}^{[3]} = \arccos(\hat{\mathbf{t}}_i^{[3]} \cdot \hat{\mathbf{t}}_{i+\Delta}^{[3]})$ (in degrees) for Δ = 5, 10, 11, 21 against base-pair index i. The ground state tangents $\hat{\mathbf{t}}_i^{[3]}$ approximate a linear fit to four base-pair origins, so that for appropriate i they fit a unit tangent localized to an A_6 run or to a spacer; e.g., when i is at the middle of an A_6, Δ = 5 accesses an angle between the A_6 and an adjacent spacer. Similarly, Δ = 10, 11 is the repeat length of an entire A-tract, so for certain i, $\hat{\Theta}_{i,10/11}^{[3]}$ is an angle between adjacent A_6, or between adjacent spacers. And $\hat{\Theta}_{i,21}^{[3]}$ is an overall bend of the basic sequence repeat, i.e., two A-tracts. Right panel: same plots, but now for the coarser-grain choice of tangent $\hat{\mathbf{t}}_i^{[11]}$ and associated $\hat{\Theta}_{i,\Delta}^{[11]}$. The coarse-grain tangents $\hat{\mathbf{t}}_i^{[11]}$ approximate a linear fit to 12 base-pair origins, so that for appropriate i they fit a unit tangent localized to each A-tract; i.e., an A_6 run plus its spacer sequence. Here the Δ = 10, 11, 21 data suggest that a reasonable coarse-grain estimate for the static angle between adjacent A-tracts is 12−13°. The very low values at the minima for Δ = 5 indicate that the coarse-grain $\hat{\mathbf{t}}_i^{[11]}$ tangents at either end of each A_6 run are close to parallel.

Table 5. Persistence Lengths (in bp) for the Phased A-Tract Sequence Fragment More Fully Described in Figure 8[a]

k	0	3	11
$l_p^{[k,0]}$	60	67	66
$l_d^{[k,0]}$	200	189	216

[a]The high sensitivity of the dynamic persistence lengths $l_d^{[k,0]}$ is due to the rather small slopes of the three, almost overlapping, almost linear dashed lines in Figure 8 to which they are fit.

Bednar et al. applied the sliding window averaging method (10) to analyze two different sets of images. One set contained fragments (of lengths 41 ± 15 nm) made up of repeats of the basic 20 bp subunit repeat in γ_5, with a total imaged length of approximately 2 μm. The shifted window fit (10) was applied with δ = 1 nm, and the single window size Δ = 40 nm to obtain the estimated persistence length 82 nm, with a reported uncertainty of 15 nm, and with the levels of coarse graining $[j, k]$ only being implicitly defined via their procedures for curve fitting in the raw images. As the γ_5 sequence is so straight, there can be little or no contribution to the apparent persistence length $l_p(\gamma_5)$ from the intrinsic shape, and thus Bednar et al. took $\overline{l}_d$ = 82 ± 15 nm as an estimate of the (sequence-averaged) dynamic persistence length of DNA.

For their second set of data, λ-phage was enzymatically cleaved into short fragments with lengths 110 ± 52 nm, and a total of 4 μm was imaged. The same sliding window method with (Δ, δ) = (40, 1) nm was used, and a persistence length of 45 nm was reported. Because of the sliding origin of the window average, it is reasonable to take this 45 nm as an estimate of the sequence-averaged apparent persistence length of λ-phage DNA. Bednar et al. then used their two estimates in the Trifonov−Tan−Harvey relation (7) to further estimate the static persistence length of DNA as 130 nm.

With the luxury of *in silico* experiment, we are not subject to the limited sampling possible *in vitro*, but to better compare our simulations with the actual experimental data, we can restrict ourselves to the Bednar et al. sampling. To that end we generated 33 samples of 180 bp (for a total of about 2 μm) of the γ_5 sequence and applied the window averaging fit (10) to compute $\overline{l}_w^{[11,11]}(\gamma_5)$. As the MC simulations deliver data at each base pair, we took the window shift to be δ = 3 bp (i.e., ≈1 nm). To implement a window size Δ_l, for each third base-pair i we found the index j with $s_j^{[11]} \le \Delta_l < s_{j+1}^{[11]}$, where $s_j^{[11]}$ is the coarse-grain arc length from base-pair i to base-pair j. Then the required data were evaluated by interpolating $\mathbf{t}_j^{[11]} \cdot \mathbf{t}_i^{[11]}$ and $\mathbf{t}_{j+1}^{[11]} \cdot \mathbf{t}_i^{[11]}$. No data were taken within 15 bp of an end. This process was then repeated 1K times to obtain an estimate and standard error $\overline{l}_w^{[11,11]}(\gamma_5)$ = 58.3 ± 11 nm, when the single window size Δ = 40 nm was used, as was done by Bednar et al., and 59.9 ± 10 nm, when 50 equally spaced window sizes from Δ = 0.8−40 nm were used. Thus, in agreement with Bednar et al., we find a notably larger than average apparent persistence length for the intrinsically straight sequence γ_5, but our increase is much less dramatic than theirs; in particular, the associated data point is off scale to the right in Figure 7.

We remark that for the intrinsically straight sequence γ_7 of Geggier and Volodoskii[34] there is the data point (49.5, 57.4), which lies close to the center of Figure 7. For both intrinsically straight sequences we can compute $l_p^{[11,11]}$ and $l_d^{[11,11]}$, and we find results that are quite close to each other (as we expect for intrinsically straight DNA) and to our window-averaged l_w results. Specifically, for γ_5, we find $l_p^{[11,11]}$ = 58.2 nm and $l_d^{[11,11]}$ = 58.5 nm, whereas for γ_7 we find $l_p^{[11,11]}$ = 57.4 nm and $l_d^{[11,11]}$ = 57.7 nm.

Similarly, we also divided the first 12 210 bp of λ-phage into 37 separate 330 bp (approximately 110 nm) fragments, generated a single configuration of each, for 4 μm of observed DNA snapshots, and proceeded as in the previous paragraph to obtain the estimate $\overline{l}_w^{[11,11]}(\lambda)$ = 52 ± 5 nm, in reasonable agreement with the 45 nm of Bednar et al. (blue square in Figure 7). Once again, this window-averaged persistence length l_w is fairly close to our λ-fragment-averaged $l_p^{[11,11]}$ value of 53.4 nm from Table 2 (but not close to the corresponding $l_d^{[11,11]}$ = 58.8 nm, because λ is not straight).

6.4. Is *cgDNAparamset2* Too Stiff? As discussed in sections 6.1 and 6.3, the comparison of our coarse-grain *cgDNAmc* simulations to experimental data addressing the sequence dependence of l_p, has overall good correlations (cf. Figure 7), albeit with some exceptions. In some cases the absolute values of our simulated apparent persistence lengths l_p

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

are substantially lower than reported experimental data, but generally they are some 15% higher than experiment. This trend poses the question of whether our coarse-grain MC simulations are being made with a *cgDNA* model free energy that is too stiff. We first note that Yamakawa[14] (p 16) observed "... but neither [the Kuhn length] nor [the persistence length] is a measure of chain stiffness except for the [WLC] chain." In other words, and in contrast to the specific case of the simple WLC, for comparatively detailed models of DNA mechanics it is a mistake to conflate statistical properties such as persistence lengths with stiffnesses, i.e., coefficients in an effective free energy. Nevertheless, an overall scaling of the *cgDNA* free energy (11), in effect changing the temperature of the solvent heat bath, would scale persistence lengths and could easily remove the observed 15% discrepancy with experiment. As already remarked, many coarse-grain models fit an overall average persistence length in similar ways,[18,20−24] whereas other coarse-grain models *predict* an average persistence length.[25−28] We adopt the predictive approach, and we believe that our sequence-averaged estimates for both apparent and Flory persistence lengths of $\overline{l}_p = \overline{l}_F = 53.5$ nm to be the closest to the experimental consensus value of 50 nm that have been obtained to date among coarse-grain models not containing an explicit fitting parameter.

We prefer the predictive approach in part because our intention is to significantly extend the length, time, and sequence multiplicity scales that are available in contemporary MD simulations of persistence length by building a much less computationally intensive MC coarse-grain code with comparable accuracy. The data presented in section 5.4 suggest that the statistics garnered from *cgDNAmc* and MD generated ensembles are rather close for sequence fragments where both are available, so that the 15% discrepancy with experiment cannot be attributed to an estimation error between our MD training sets and our coarse-grain parameter sets. We have also tested that there is no significant statistical error in the *cgDNAmc* simulations themselves.

In the MD simulations that were used to develop the *cgDNA* parameter set, we used the Amber force field bsc0 with potassium chloride salt at approximately physiological concentration.[43,52] MD protocols, including which potentials are used for the water, ions, and solute, are constantly evolving. For example, there are now two more recent Amber-family force field modifications available.[31,61] Estimation of the sensitivity of the *cgDNA* parameters to such refinements in MD potentials is underway, but it seems quite unlikely that a difference of 15% will arise. A more likely source of the discrepancy is that the MD training set data that we have used has all been run with monovalent salt at physiological concentrations, whereas there are experimental data suggesting that the DNA persistence length can change substantially with different salt concentration and species, particularly multivalent salts.[62,63] And the available experimental data to which we compare in this article have a wide variety of salt conditions, frequently involving magnesium. From the viewpoint of the *cgDNA* model, changing salt concentration or including multivalent salt is unproblematic, provided that accurate and sufficiently converged MD training library simulations are available. But the MD simulation of multivalent salts and associated polarizable force fields is still an active research field with high computational demands.[64]

Finally we note that the *cgDNA* model is Gaussian in the internal helical coordinates **w** and in particular assumes that the DNA structure is reasonably close to helical, so that these helical coordinates are a reasonable description. For example, though *cgDNA* predicts poly(A) to be an exceptionally straight and stiff helical structure, the model can say nothing about the robustness, or fragility, of that double helical structure to non-Gaussian deformations such as kinking, which is one possible resolution to apparently contradictory data regarding exceptional softness or stiffness of A-tracts. In fact, it has already been suggested[65,66] that in both 2D microscopy and cyclization experiments, and for any sequence, small denaturation bubbles in the double helix can be nucleated, which would effectively decrease the apparent DNA persistence length that is observed experimentally. Such effects could easily give rise to a 15% discrepancy, as well as raising the basic issue of whether a persistence length of DNA as a double helix or as a kinked double helix is the central object of interest.

Kinking of short, and therefore highly loaded, covalently closed DNA minicircles has been observed in MD simulations.[67] More recently, non-Gaussian behaviors have also been observed in long duration simulations of unloaded short linear double-stranded fragments, and their sequence dependence has started to be understood,[43] which may well be related to the variation of magnitude of error with sequence in the comparisons between MD and *cgDNAmc* ensembles presented in Table 4. All of these remarks indicate the desirability of extending the current Gaussian version of the *cgDNA* model to be able to predict nonquadratic sequence-dependent free energies. The Metropolis version of *cgDNAmc* is already available to generate ensembles with any such non-Gaussian equilibrium distributions. However, the current model already reveals the strong sequence dependence of DNA persistence lengths, and the necessity of considering the shape factorization introduced in the definition (9) of sequence-dependent dynamic persistence length to obtain close-to-linear decay in the semilog tangent−tangent correlations. We believe that such effects are likely to persist, and even to remain dominant, in any perturbed non-Gaussian model.

7. SUMMARY

A new notion of sequence-dependent dynamic persistence length $l_d(\mathbb{S})$ (9) has been introduced to deconvolve the separate effects of intrinsic shape and stiffness in the usual tangent−tangent correlation statistics along double-stranded DNA. Working within the context of a numerically efficient code *cgDNAmc*, which we have developed to sample sequence-dependent equilibrium distributions predicted by the rigid base *cgDNA* coarse-grain model, we have verified that the sequence-dependent dynamic persistence length $l_d(\mathbb{S})$ arises from semilog linear data fits that are of uniformly higher quality at the scale of 200 bp than those for the standard, or apparent, tangent−tangent persistence length $l_p(\mathbb{S})$ (6) where the effects of intrinsic shape are not accounted for, and consequently, the data to be fit are liable to deviate significantly from linearity. The same phenomenon has been demonstrated in a small number of simulations of much shorter fragments where the statistics are generated directly from MD simulations, and where comparison with *cgDNAmc* results are also rather good. However, the coarse-grain *cgDNAmc* code provides a bridge over the scales, allowing predictions of sequence-dependent DNA behaviors at lengths not currently accessible to MD, up to thousands of bp, and this for a large range of possible sequences, all with comparatively minor computational effort.

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

The *cgDNAmc* code allows spectra of persistence lengths over an ensemble of sequences to be generated, revealing that the resulting histograms for persistence lengths in the two standard senses of apparent tangent−tangent $l_p(\mathbb{S})$ and Flory $l_F(\mathbb{S})$ (5) are quite similar one to the other. Flory and apparent persistence lengths are known to coincide exactly for the standard sequence-independent WLC model, but the similarity of these two spectra in their variation over sequence is not implied by any currently known theory. Both spectra exhibit strong dependence of persistence length on the sequence $\mathbb{S}$, with in each case many sequences having quite low persistence lengths. These low persistence lengths (in both senses) can almost invariably be attributed to the sequence being significantly intrinsically bent, so that the semilog linear fit to compute $l_p(\mathbb{S})$ can be called into question. In comparison, the spectra for dynamic persistence lengths $l_d(\mathbb{S})$ are sharply focused with a peak at approximately l_d = 180 bp, with comparatively little variation with sequence. There are a few striking exceptions to this observation, including poly(AT) and poly(A), which are both very straight, and which have respectively the lowest and highest dynamic persistence lengths that we observed in any sequence simulated within *cgDNAmc*, with poly(A) at l_d = 221 bp being approximately 50% stiffer than poly(AT) at l_d = 148 bp. This large difference in stiffness between the two polydinucleotide sequences can also be observed directly from MD simulations of very short fragments. As discussed in section 6.2, the *cgDNA* model additionally predicts that the homopolymer sequence poly(A) behaves very differently from sequences containing phased A-tracts, which have an exceptionally high dynamic persistence length but strong localized intrinsic bends.

The *cgDNAmc* code allows us to easily compute sequence-averaged Flory $\overline{l}_F$ and apparent $\overline{l}_p$ persistence lengths (cf. Table 2), the values of which we find to be in good agreement with the generally accepted value of 50 nm. Indeed, among current coarse-grain models we believe these sequence-averaged predictions to be the best available. When sequence-averaged, our dynamic persistence length $\overline{l}_d$ is approximately 10% higher than the apparent persistence length $\overline{l}_p$, and the Trifonov−Tan−Harvey relation (7) between sequence-averaged apparent, dynamic, and static persistence lengths is satisfied to a good accuracy.

For specific sequences the *cgDNAmc* code allows us to easily predict persistence lengths, in various senses, that have already been estimated from experimental data. Section 6.1 presents comparisons of this type. Generally, we obtain quite reasonable correlation over sequences between our coarse-grain predictions and prior experimental data, but with a trend of overestimation by 15% in the absolute values of persistence lengths. As discussed in section 6.4, this discrepancy could be attributed to any of a number of different reasons. Our simulations do validate two experimental approaches proposed by Bednar et al.,[35] namely (1) the use of periodic sequences that are designed to be straight as a means to access dynamic persistence length (for both of our straight sequences γ_5 and γ_7 our computed l_d and l_p agree to within 0.5 nm) and (2) the use of sliding-window measurements for the interpretation of microscopy images (our computed l_w and l_p agree to within 0.1 nm for $\gamma_{5,7}$, and to within 2 nm for λ-phage). Such microscopy techniques thus seem promising for continuing to directly explore sequence-dependent shape and flexibility of double-stranded DNA.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.6b00904.

(1) Further description of the coarse-grain tangent vectors $\mathbf{t}_i^{[k]}$ and of the Monte Carlo code; (2) full DNA sequences for all of the fragments considered; (3) figures showing (a) interactive U3D versions of Figure 4, (b) the sensitivity of results to the Jacobian perturbation and choice of coarse-grain parameters, (c) Monte Carlo convergence, (d) histograms of persistence length for $j = k = 11$ coarse graining, and (e) tangent−tangent plots used to produce some of the key results in the article; (4) tables of persistence length values, from which some of the plots in the article were produced (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*J. Maddocks. E-mail: john.maddocks@epfl.ch.

ORCID

John H. Maddocks: 0000-0003-1127-8481

Present Address

§The Microsoft Research-University of Trento, Centre for Computational and Systems Biology (COSBI), Rovereto, Italy.

Notes

The authors declare no competing financial interest.

The codes used to generate the simulation data for the article are freely downloadable from http://lcvmwww.epfl.ch/cgDNA.

■ ACKNOWLEDGMENTS

J.S.M., J.G., A.E.G., and J.H.M. were supported in part by the Swiss National Science Foundation Award 200020 143613/1 to J.H.M. It is a pleasure for the authors to be able to thank A. Patelli for help in producing Figure 6 and Table 4 and the anonymous referees for their suggested revisions and additions to an earlier version of the manuscript.

■ REFERENCES

(1) Virstedt, J.; Berge, T.; Henderson, R. M.; Waring, M. J.; Travers, A. A. *J. Struct. Biol.* **2004**, *148*, 66−85.

(2) Garcia, H. G.; Grayson, P. A.; Han, L.; Inamdar, M.; Kondev, J.; Nelson, P. C.; Phillips, R.; Widom, J.; Wiggins, P. A. *Biopolymers* **2007**, *85*, 115−130.

(3) Segal, E.; Widom, J. *Curr. Opin. Struct. Biol.* **2009**, *19*, 65−71.

(4) Sarai, A.; Kono, H. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 379−398.

(5) Hagerman, P. J. *Annu. Rev. Biophys. Biophys. Chem.* **1988**, *17*, 265−286.

(6) Rittman, M.; Gilroy, E.; Koohya, H.; Rodger, A.; Richards, A. *Sci. Prog.* **2009**, *92*, 163−204.

(7) Kratky, O.; Porod, G. *Recueil des Travaux Chimiques des Pays-Bas* **1949**, *68*, 1106−1122.

(8) Peters, J. P.; Maher, L. J. Q. *Q. Rev. Biophys.* **2010**, *43*, 23−63.

(9) Trifonov, E. N.; Tan, R. K.; Harvey, S. C. In *Structure and Expression (Vol. 3): DNA Bending & Curvature*; Olson, W. K., Sarma, R. H., Sarma, M. H., Sundaralingam, M., Eds.; Adenine Press: New York, 1988.

(10) Flory, P. *Statistical Mechanics of Chain Molecules*; Interscience: New York, 1969.

(11) Maroun, R. C.; Olson, W. K. *Biopolymers* **1988**, *27*, 585−603.

(12) Théveny, B.; Coulaud, D.; Le Bret, M.; Révet, B. In *Structure and Expression (Vol.3): DNA Bending & Curvature*; Olson, W. K.,

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555

Sarma, R. H., Sarma, M. H., Sundaralingam, M., Eds.; Adenine Press: New York, 1988.
(13) Schellman, J. A.; Harvey, S. C. *Biophys. Chem.* **1995**, *55*, 95−114.
(14) Yamakawa, H. *Helical Wormlike Chains in Polymer Solutions*; Springer: Berlin, Heidelberg, 1997.
(15) Rivetti, C.; Walker, C.; Bustamante, C. *J. Mol. Biol.* **1998**, *280*, 41−59.
(16) Dans, P. D.; Walther, J.; Gómez, H.; Orozco, M. *Curr. Opin. Struct. Biol.* **2016**, *37*, 29−45.
(17) Vologodskii, A. *Biophysics of DNA*; Cambridge University Press: Cambridge, U.K., 2015.
(18) Olson, W. K.; Gorin, A. A.; Lu, X.-J.; Hock, L. M.; Zhurkin, V. B. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 11163−11168.
(19) Olson, W. K.; Colasanti, A. V.; Czapla, L.; Zheng, G. In *Coarse-graining of condensed phase and biomolecular systems*; Voth, G. A., Ed.; CRC Press: Boca Raton, FL, 2008.
(20) Sulc, P.; Romano, F.; Ouldridge, T. E.; Rovigatti, L.; Doye, J. P. K.; Louis, A. A. *J. Chem. Phys.* **2012**, *137*, 135101.
(21) Hinckley, D. M.; Freeman, G. S.; Whitmer, J. K.; de Pablo, J. J. *J. Chem. Phys.* **2013**, *139*, 144903.
(22) Morriss-Andrews, A.; Rottler, J.; Plotkin, S. S. *J. Chem. Phys.* **2010**, *132*, 035105.
(23) Naome, A.; Laaksonen, A.; Vercauteren, D. P. *J. Chem. Theory Comput.* **2014**, *10*, 3541−3549.
(24) Uusitalo, J. J.; Ingólfsson, H. I.; Akhshi, P.; Tieleman, D. P.; Marrink, S. J. *J. Chem. Theory Comput.* **2015**, *11*, 3932−3945.
(25) Maciejczyk, M.; Spasic, A.; Liwo, A.; Scheraga, H. A. *J. Chem. Theory Comput.* **2014**, *10*, 5020−5035.
(26) Knotts, T. A.; Rathore, N.; Schwartz, D. C.; de Pablo, J. J. *J. Chem. Phys.* **2007**, *126*, 084901.
(27) Sayar, M.; Avsaroglu, B.; Kabakcioglu, A. *Phys. Rev. E* **2010**, *81*, 041916.
(28) Savelyev, A.; Papoian, G. A. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 20340−20345.
(29) Mazur, A. K. *Biophys. J.* **2006**, *91*, 4507−4518.
(30) Noy, A.; Golestanian, R. *Phys. Rev. Lett.* **2012**, *109*, 228101.
(31) Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A.; Portella, G.; Battistini, F.; Gelpi, J. L.; Gonzalez, C.; Vendruscolo, M.; Laughton, C. A.; Harris, S. A.; Case, D. A.; Orozco, M. *Nat. Methods* **2016**, *13*, 55−58.
(32) Gonzalez, O.; Petkevičiūtė, D.; Maddocks, J. H. *J. Chem. Phys.* **2013**, *138*, 055102.
(33) Petkevičiūtė, D.; Pasi, M.; Gonzalez, O.; Maddocks, J. H. *Nucleic Acids Res.* **2014**, *42*, e153−e153.
(34) Geggier, S.; Vologodskii, A. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 15421−15426.
(35) Bednar, J.; Furrer, P.; Katritch, V.; Stasiak, A. Z.; Dubochet, J.; Stasiak, A. *J. Mol. Biol.* **1995**, *254*, 579−594.
(36) Flory, P. *Proc. Natl. Acad. Sci. U. S. A.* **1973**, *70*, 1819−1823.
(37) Schellman, J. A. *Biopolymers* **1974**, *13*, 217−226.
(38) Doi, M.; Edwards, S. F. *The Theory of Polymer Dynamics*; Clarendon Press: Oxford, U.K., 1988.
(39) Kahn, J. D.; Crothers, D. M. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 6343−6347.
(40) Brunet, A.; Chevalier, S.; Destainville, N.; Manghi, M.; Rousseau, P.; Salhi, M.; Salomé, L.; Tardin, C. *Nucleic Acids Res.* **2015**, *43*, e72.
(41) Fathizadeh, A.; Eslami-Mossallam, B.; Ejtehadi, M. R. *Phys. Rev. E* **2012**, *86*, 051907.
(42) Lavery, R.; Moakher, M.; Maddocks, J. H.; Petkevičiūtė, D.; Zakrzewska, K. *Nucleic Acids Res.* **2009**, *37*, 5917−29.
(43) Pasi, M.; Maddocks, J. H.; Beveridge, D.; Bishop, T. C.; Case, D. A.; Cheatham, T. E.; Dans, P. D.; Jayaram, B.; Lankas, F.; Laughton, C.; Mitchell, J.; Osman, R.; Orozco, M.; Perez, A.; Petkevičiūtė, D.; Spackova, N.; Sponer, J.; Zakrzewska, K.; Lavery, R. *Nucleic Acids Res.* **2014**, *42*, 12272−12283.
(44) Becker, N. B.; Everaers, R. *Phys. Rev. E* **2007**, *76*, 021923.
(45) Lankas, F.; Gonzalez, O.; Heffler, L. M.; Stoll, G.; Moakher, M.; Maddocks, J. H. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10565−10588.
(46) Walter, J.; Gonzalez, O.; Maddocks, J. H. *Multiscale Model. Simul.* **2010**, *8*, 1018−1053.
(47) Chirikjian, G. *Stochastic Models, Information Theory, and Lie Groups, Vol. 2: Analytic Methods and Modern Applications*; Birkhäuser: Boston, 2011.
(48) Gentle, J. E. *Random Number Generation and Monte Carlo Methods*. Springer: New York, 1998.
(49) Golub, G. H.; Van Loan, C. F. *Matrix Computations*; Johns Hopkins University Press: Baltimore, 1996.
(50) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087−1092.
(51) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. *Comput. Phys. Commun.* **1995**, *91*, 1−41.
(52) Lavery, R.; Zakrzewska, K.; Beveridge, D.; Bishop, T. C.; Case, D. A.; Cheatham, T. E.; Dixit, S.; Jayaram, B.; Lankas, F.; Laughton, C. A.; Maddocks, J. H.; Michon, A.; Osman, R.; Orozco, M.; Perez, A.; Singh, T.; Spackova, N.; Sponer, J. *Nucleic Acids Res.* **2010**, *38*, 299−313.
(53) Gonzalez, O.; Pasi, M.; Petkevičiūtė, D.; Glowacki, J.; Maddocks, J. H. Preprint 2016.
(54) Shore, D.; Langowski, J.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U. S. A.* **1981**, *78*, 4833−4837.
(55) Shimada, J.; Yamakawa, H. *Macromolecules* **1984**, *17*, 689−698.
(56) Nikolova, E. N.; Bascom, G. D.; Andricioaei, I.; Al-Hashimi, H. M. *Biochemistry* **2012**, *51*, 8654−8664.
(57) Johnson, S.; Chen, Y.; Phillips, R. *PLoS One* **2013**, *8*, e75799.
(58) Haran, T. E.; Mohanty, U. Q. *Rev. Biophys.* **2009**, *42*, 41−81.
(59) de Boer, C. G.; Hughes, T. R. *PLoS One* **2014**, *9*, e110479.
(60) Zhurkin, V.; Tolstorukov, M.; Fei, X.; Colasanti, A.; Olson, W. In *DNA conformation and transcription*; Ohyama, T., Ed.; Springer: Berlin, 2005; pp 18−34.
(61) Zgarbova, M.; Luque, F. J.; Sponer, J.; Cheatham, T. E.; Otyepka, M.; Jurecka, P. *J. Chem. Theory Comput.* **2013**, *9*, 2339−2354.
(62) Brunet, A.; Tardin, C.; Salomé, L.; Rousseau, P.; Destainville, N.; Manghi, M. *Macromolecules* **2015**, *48*, 3641−3652.
(63) Savelyev, A. *Phys. Chem. Chem. Phys.* **2012**, *14*, 2250−2254.
(64) Savelyev, A.; MacKerell, A. D. *J. Comput. Chem.* **2014**, *35*, 1219−1239.
(65) Yan, J.; Marko, J. F. *Phys. Rev. Lett.* **2004**, *93*, 108108.
(66) Destainville, N.; Manghi, M.; Palmeri, J. *Biophys. J.* **2009**, *96*, 4464−4469.
(67) Lankaš, F.; Lavery, R.; Maddocks, J. H. *Structure* **2006**, *14*, 1527−1534.

DOI: 10.1021/acs.jctc.6b00904
J. Chem. Theory Comput. 2017, 13, 1539−1555