Analysis of a DNA simulation model through hairpin melting experiments

Margaret C. Linak and Kevin D. Dorfman^{a)}

*Department of Chemical Engineering and Materials Science, University of Minnesota-Twin Cities,
421 Washington Ave. SE, Minneapolis, Minnesota 55455, USA*

(Received 5 May 2010; accepted 29 July 2010; published online 22 September 2010)

We compare the predictions of a two-bead Brownian dynamics simulation model to melting experiments of DNA hairpins with complementary AT or GC stems and noninteracting loops in buffer A. This system emphasizes the role of stacking and hydrogen bonding energies, which are characteristics of DNA, rather than backbone bending, stiffness, and excluded volume interactions, which are generic characteristics of semiflexible polymers. By comparing high throughput data on the open-close transition of various DNA hairpins to the corresponding simulation data, we (1) establish a suitable metric to compare the simulations to experiments, (2) find a conversion between the simulation and experimental temperatures, and (3) point out several limitations of the model, including the lack of G-quartets and cross stacking effects. Our approach and experimental data can be used to validate similar coarse-grained simulation models. © 2010 American Institute of Physics. [doi:10.1063/1.3480685]

I. INTRODUCTION

Coarse-grained modeling is an increasingly widespread method to study the dynamics of single-stranded DNA. These approaches include Monte Carlo,^{1–6} Brownian dynamics (BD),^{7–10} molecular dynamics (MD),^{11–25} and lattice Boltzmann^{26–28} methods. For a given level of computational resources, there always exists a tradeoff between the available length and time scales. Coarse-grained models allow access to relatively long time scales, albeit at a correspondingly coarse length scale.^{22,27}

At the heart of any DNA simulation model are the potential energy functions used to quantify the interactions between different parts of the chain. These functions (and their parameters) need to be selected to capture the physical properties of single-stranded DNA. At a macroscopic level, we would expect a robust model to capture thermal denaturation. As a prototypical example, we consider here the case of a DNA hairpin. At the minimum, the model should at least lead to the correct melting-point temperature, i.e., the temperature at which there is a 50% probability of locating the hairpin in the open state. The melting-point temperature is a function of ionic strength, which provides an approach to tune the parameters for different experimental conditions.^{6,22} At the next level of complexity, we would like the model to mimic the entire dependence of the sigmoidal melting temperature curve from the 100% closed state at low temperatures to the 100% open state at high temperatures. In particular, capturing the “shoulders” of the melting temperature curve near the fully closed and fully open states is particularly challenging. A model that passes this test, especially in a biologically relevant buffer, could be used with confidence in more complex *in vivo* scenarios.

Here, we provide such a test of a two-bead BD model of single-stranded DNA (Ref. 10) by comparing high through-

put data on the open-close transition of various DNA hairpins to the corresponding simulation data. In the course of this test, we also establish a suitable metric to compare the simulations to experiments. Although a great deal of literature exists for the melting-point temperature of different DNA sequences, the data depend strongly on the particular experimental conditions and are prone to experimental artifacts. To provide a stringent test of the model, we obtained a large set of experimental data under well-controlled conditions. The two-bead BD model, like other coarse-grained representations, can be made to reproduce the melting-point temperature at a single point (50% melted), but does not exactly predict the entire melting behavior. The approach used here (and the concomitant experimental data) could be used to examine other models^{1–5,7–24,26–29} to determine their ability to capture a basic experimental system. Such work is important to the continued use of coarse-grained models of single-stranded DNA; given the simplicity of our experimental system, we would hope that more sophisticated models can capture the full range of the experimental data.

II. METHODS

A. Experimental protocol

We examined a base case of $A_5C_5T_5$ and select variants, which are grouped in the three classes listed in Table I. (1) In the first class we retain the nucleotide sequence $A_XC_YT_X$ while varying the values of X and Y ; (2) In the second class, the stem identity is conserved but the cytosine bases in the loop are replaced with guanines of different lengths. (3) In the third class, the stem bases were changed to guanine and cytosine with adenine and thymine as the loop bases. To simplify the subsequent discussion of the data analysis, we assigned each sequence the value $j=1:7$ appearing in Table I. In addition to the sequences in Table I, we also considered several other sequence variants that fall into these classes ($A_{10}G_{20}T_{10}$, $A_5G_{20}T_5$, $A_{10}C_{20}T_{10}$, $C_{10}T_{20}G_{10}$, $G_5T_{10}C_5$,

^{a)}Electronic mail: dorfman@umn.edu.

TABLE I. List of single-stranded DNA sequences. The Δ classes refer to the change in stem length (1), loop (2), or sequence (3) when compared to the base case $A_5C_5T_5$. The index j will be used throughout to denote the DNA hairpin type. The *mfold* predicted melting-point temperatures were found for 1 μ M monovalent sodium, similar to buffer A, and are given for the aligned stem bonding configuration (Ref. 30).

Class	Type	Sequence	j	T_{MP} (mfold) (K)
1	Base case	$A_5C_5T_5$	1	319.1
	Δ Length	$A_5C_{10}T_5$	2	308.5
		$A_7C_5T_7$	3	330.6
		$A_7C_{10}T_7$	4	323.1
2	Δ Loop	$A_5G_5T_5$	5	331.8
		$A_5G_{10}T_5$	6	324.3
3	Δ Identity	$G_5A_{10}C_5$	7	347.1

$A_{10}C_5T_{10}$, $A_{12}C_5T_{12}$, and $A_{12}C_{10}T_{12}$), but these were rejected due to a high melting-point (T_{MP}) temperature or other experimental difficulties (such as low synthesis yields, especially with poly-G sequences). Each single-stranded DNA sequence was obtained from Integrated DNA Technologies and HPLC purified by the manufacturer prior to use. The lyophilized powder was serially diluted with deionized (DI) water to a concentration of 1 μ M to make a stock solution.

A quantitative polymerase chain reaction (QPCR) machine (Mx3000P, Stratagene, La Jolla, California) was used to collect the temperature and fluorescence intensity data for the DNA hairpins in Table I. Although acquiring melting curve data is not the standard use of a QPCR system, the machine's accurate temperature control, ability to excite and detect fluorescent molecules, and its 96 sample capacity make it well suited for our experiments. However, this equipment choice led to some experimental restrictions. For example, the DNA hairpins must have relatively low melting-point temperatures ($T_{MP} < 353$ K) due to the fact that (i) the buffer experiences significant vaporization in the upper temperature range, which can lead to inaccurate fluorescence readings due to condensation on the cap of the sample tube, and (ii) the QPCR system is not designed to capture temperatures greater than 363 K. We only used sequences with *mfold* predicted melting-point temperatures $T_{MP} < 353$ K (1M monovalent sodium, similar to buffer A) to ensure a sufficient dynamical range to capture the upper plateau of the melting curve.³⁰ The *mfold* predicted melting-point temperatures are summarized in Table I. This restriction limits the possible sequences to be examined experimentally.

The experiments were conducted in a biologically relevant buffer, buffer A (Ref. 31) at 1X concentration (0.05M HEPES, 0.5M NaCl, and 0.5M KCl; ionic strength=1; pH =7.1 at 25 °C). This standard composition of buffer A is a reasonable model for *in vivo* conditions and has proven useful for *in vitro* applications as well; for example, buffer A at these specifications was used in the initial evolution of the 10–23 DNAzyme.^{31,32} While appropriate for studying basic physics, a monovalent salt poorly captures the effects of a complex biologically relevant buffer, like buffer A.

Three different types of wells (control buffer wells, control dye wells, and experimental signal wells), each containing 50 μ l of solution, were arranged randomly on each 96 well plate. The contents of each well are summarized in Table II. The control buffer wells each contained only a 1X buffer A solution and were used to measure the background fluorescence signal of the biological salt solution. The control dye wells contained a 1X buffer A solution and a 2X SYBR Green I dye solution. The SYBR Green I fluorescence when bound to double stranded DNA is 800- to 1000-fold greater than when bound to single-stranded DNA (Molecular Probes, Eugene, Oregon). This well type was used to measure the background fluorescence signal of the unbound SYBR Green I intercalating dye. In the experimental signal wells, the 1X buffer A and 2X SYBR Green I dye solutions were combined with 2.2 μ M of DNA hairpin stem base pairs. These optimal concentrations of stem base pairs and dye were determined by an independent set of measurements of the melting of the $A_{10}C_5T_{10}$ sequence over a range of SYBR I concentrations (0.02X–10X) and DNA concentrations (0.02 μ M to 0.2 μ M). The replicates of each well type are enumerated in Table II. Two identical plates were made and each plate was run twice to generate the experimental intensity data.

We tested temperature ramps of $\delta T=1$ K/5 min and $\delta T=1$ K/min and did not notice any significant change in the data when the fluorescence versus temperature curves obtained at each ramp rate were overlaid. Most of the results reported here were obtained with $\delta T=1$ K/min as the temperature transition rate. Fluorescence data were collected over the range of 293–363 K. The fluorescence signal at a single temperature was measured for each well 14 times before raising the temperature by one degree. The specifications of the QPCR system allowed for temperature plateaus to be programmed in increments of 1 K while the actual temperature has a precision of ± 0.1 K. The plate was allowed to equilibrate for 5 min once the new temperature was

TABLE II. List of well types randomly loaded on each 96 well plate. One of the seven DNA hairpins ($j=1-7$) under study was loaded into each experimental signal well. The number of replicates of each well type per plate is the number of well numbers, k , corresponding to that experimental condition.

Well type	Well contents			Well number, k
	Buffer A	SYBR Green I dye	DNA hairpin	
Control buffer well	1X	0	0	$k_{\text{buffer}}=1-4$
Control dye well	1X	2X	0	$k_{\text{dye}}=5-12$
Experimental signal well j	1X	2X	2.2 μ M stem base pairs	$k_{\text{DNA}}=1+12j$

FIG. 1. Example of a two-bead model of the single-stranded sequence $G_{10}A_{18}C_{10}$; this chain is used for illustration purposes and is not a sequence in the study. The chain is comprised of a series of contiguous backbone beads and a series of base beads. The chain is initialized in either a linear or square U configuration. The Watson-Crick base pairing stem and noninteracting loop sections are labeled.

reached before obtaining fluorescence data. This process was repeated for each temperature in the specified range, creating over 80 000 measurements per plate. The resultant raw data consist of the temperature and corresponding fluorescence intensity (in arbitrary units) of each sample.

Each 96 well plate was loaded in a randomized manner in order to control for the known edge effects in QPCR systems. The wells consist of a series of control buffer wells, control dye wells, and experimental signal wells. All data collected from areas of the plates deemed inconsistent were eliminated from further analysis if it satisfied either of the following criteria: (i) the signal strengths were less than 10% or greater than 500% of the average signal strength at the corresponding temperature, or (ii) the signal variance was greater than ten times the average signal variance at the corresponding temperature.

B. Simulation protocol

We utilized a standard Brownian dynamics algorithm with a base-backbone single-stranded DNA model.^{6–10,33} In this two-bead model, each phosphate-sugar group is modeled as a single backbone bead. The beads are linked together to form a contiguous backbone, as seen in Fig. 1. In addition, each DNA base is modeled as a second bead that is connected to a backbone bead. Thus, the single-stranded DNA molecule is comprised of multiple base-backbone units.

The particular two-bead model we use here and the basic parameters were described previously.^{10,33} It is important to note that all of the potentials describing the system are defined in terms of the base units for energy, ϵ , and length, σ , in the model. The complete parameter space that explicitly describes the model is comprised of 41 variables, some of which are discrete.³³ In this study, a small section of the parameter space was explored by varying the magnitude of the base stacking energy, ϵ_{stack} , and the hydrogen bonding energy, ϵ_{Hbond} , while keeping the ratio of these quantities fixed; data obtained for values $\epsilon_{\text{stack}}/\epsilon_{\text{Hbond}}=2.5/1$, $5/2$, and $10/4$ are presented here. We kept the ratio of the stacking and hydrogen bonding energies constant at the value approximated by Bloomfield *et al.*³⁴ from thermodynamic data, but

allowed the magnitude of these quantities to change in relation to the other energies of the system. The system is written in reduced units with a nondimensional temperature of T and a unit mass attributed to all beads.

At the start of the simulations, the molecule is initialized into one of the two conformations illustrated in Fig. 1: a linear chain (open configuration) or a square U chain (almost closed configuration). To ensure relaxation from this initial state, the simulations were carried out for 5×10^6 BD time steps to erase memory of the initial configuration. We then obtained configuration data for approximately 5×10^8 BD time steps with a time step of $\delta t=0.01$, where a single time step corresponds approximately to a nanosecond. This total simulation time leads to a sufficiently large number of binding and unbinding steps when $T \approx T_{\text{MP}}$, allowing reliable measurements of the phenomenon. Independent runs with different initial conditions were performed to ensure robustness with respect to the starting conformation of the molecules.

The simulations use a nondimensional temperature, but prior to this study we did not have an experimentally validated conversion to a real temperature. To limit the dimensionless temperature phase space that we needed to explore, we used a two step procedure. In the first step, we noted that the maximum characteristic hydrogen bonding potential in the model is approximately $0.123(\epsilon_{\text{Hbond}})$.¹⁰ This leads to a characteristic energy of $\epsilon=-6.1$ kcal/mol. With a choice of $u_{\text{Hbond}}=2$ we find a temperature conversion of $T=0.1 \rightarrow 310$ K. By first narrowing the temperature space through these estimates, we were able to make initial sweeps of simulations at nondimensional temperatures in the range of 0.1–0.6 in increments of 0.05. In the second step, we analyzed these data with the metrics described below and then fine tuned the temperature search near the melting-point, and thus the transition region of the hairpin. Having narrowed the temperature range to within ± 0.1 of the melting-point of the simulated sequence, the model was examined in more detail within this regime in temperature increments of 0.005. Additional simulations were added if the high temperature plateau was not sufficiently stable and clearly delineated. Three independent simulations were completed at every temperature point. The position of each of the simulated beads was saved every 1000 BD time steps (≈ 1 μs).

III. DATA ANALYSIS

A. Experimental data analysis

For each well, we binned the raw intensity data into temperature increments of 1 K and then averaged the 14 data replicates in each bin to produce an intensity $I_k(T)$ for each of the $k=1:96$ wells, where T is measured in $\Delta T=1$ K increments. For the control buffer wells, $k=k_{\text{buffer}}=1:4$ on a single plate, the intensity signals were averaged to create a plate specific background signal corresponding to the specific characteristics of buffer A. This produced a plate specific background signal,

FIG. 2. Postprocessing of the experimental data for the $A_5C_5T_5$ sequence. (a) Average background fluorescence signal, $B_{\text{plate}}(T)$. (b) Average background corrected unbound SYBR Green I temperature dependence, $\langle \text{ubSG}(T) \rangle$. (c) An example well's melting curve, $I_k^{\text{MC}}(T)$; the temperature dependence of the bound SYBR Green I signal is fit from the closed hairpin state, $b\text{SG}_{\text{fit},k}(T)$. (d) Normalized and averaged replicates for the $j=1$ sequence. The temperature variation range is reported at selected normalized intensity values.

$$B_{\text{plate}}(T) = \langle I_k \rangle_{1:4}, \quad (1)$$

such as the one in Fig. 2(a). In the latter and what follows, $\langle \dots \rangle_k$ represents an average of the k wells. The background fluorescence signal, $B_{\text{plate}}(T)$, is subtracted from the raw data for the control dye wells ($k=k_{\text{dye}}=5:12$) and the DNA containing wells ($k=k_{\text{DNA}}=13:96$) to create the corrected intensity

$$I'_k(T) = I_k(T) - B_{\text{plate}}(T). \quad (2)$$

After correcting for the background signal, the data from the control dye wells ($k=k_{\text{dye}}=5:12$) were similarly smoothed into an average curve to form the unbound SYBR Green I, background corrected, fluorescence corrected signal

$$\langle \text{ubSG}(T) \rangle = \langle I'_k \rangle_{5:12} \quad (3)$$

seen in Fig. 2(b). The $\text{ubSG}(T)$ curve was fit with a linear temperature relationship, $\text{ubSG}_{\text{fit}}(T)$,

$$\text{ubSG}_{\text{fit}}(T) = a_{\text{fit}}T + b_{\text{fit}}, \quad (4)$$

where a_{fit} and b_{fit} are the coefficients of the linear regression for the averaged unbound SYBR I signals. The linear SYBR I background fluorescence, $\text{ubSG}_{\text{fit}}(T)$, is subtracted from the raw data for the DNA containing wells ($k=k_{\text{DNA}}=1+12j$, where $j=1:7$ counts the seven DNA hairpin types,

$$I''_k(T) = I'_k(T) - \text{ubSG}_{\text{fit}}(T). \quad (5)$$

The value $I''_k(T)$ thus corrects for the background fluorescence signal of the buffer and the signal due to the excess amounts of intercalating dye.

The maximum of each $I''_k(T)$ curve, I''_k^{max} , was found for each of the k_{DNA} wells. Applying the melting curve (MC) algorithm,

$$I_k^{\text{MC}}(T) = I_k^{\text{max}} - I''_k(T), \quad (6)$$

transforms the data into the standard melting curve format prevalent in literature (though not yet normalized on $[0,1]$). This melting curve has the familiar sigmoidal shape, with high temperatures corresponding to high intensity values and low temperatures corresponding to low intensity data.

In addition to the unbound dye, we also needed to correct for the temperature dependence of the bound SYBR Green I. To make this correction, we assumed that (i) the stems are all fully closed between 298 and 303 K, (ii) the bound SYBR I fluorescence depends linearly on temperature, similar to the unbound dependence measured in Fig. 2(b), and (iii) the fluorescence intensity is proportional to the number of bonded base pairs. From the low temperature (and thus closed state) data, an extrapolation of the bound SYBR Green I background signal

$$b\text{SG}_k(T) = a_kT + b_k \quad (7)$$

was formed, as shown in Fig. 2(c), for each of the DNA containing wells $k=k_{\text{DNA}}$.

We then removed each bound SYBR I contribution from the corresponding fluorescence signal at every temperature using the iterative approach described in Eqs. (8)–(10). We first approximated the fraction of closed base pairs as

$$\text{cbp}_k(T) = 1 - I_k^{\text{MC}}(T)/I_k^{\text{MC max}}, \quad (8)$$

where $I_k^{\text{MC max}}$ is the maximum value for the adjusted intensity I_k^{MC} . We then subtracted the scaled extrapolated bound dye intensity, given by

$$I_k^{\text{scale}}(T) = \text{cbp}_k(T) \times b\text{SG}_k(T), \quad (9)$$

from every data point to arrive at a new intensity value for I_k ,

$$I_k^*(T) = I_k^{\text{MC}}(T) - I_k^{\text{scale}}(T). \quad (10)$$

In the algorithm, I_k^* is set equal to I_k^{MC} in Eq. (8) and the algorithm of Eqs. (8)–(10) is iterated until the difference between $I_k^*(T)$ and I_k^{MC} is less than 1×10^{-8} . Once convergence is reached for each $I_k^*(T)$ value, where $k=k_{\text{DNA}}=1+12j$ for each DNA sequence j , the 12 DNA replicates are averaged to form $\widehat{I_j^*}(T)$,

$$\widehat{I_j^*}(T) = \langle I_k^*(T) \rangle_{1+12j}. \quad (11)$$

The $\widehat{I_j^*}(T)$ is then normalized on $[0,1]$. Figure 2(d) shows the data $\widehat{I_1^*}(T)$ (for the $j=1$, $A_5C_5T_5$ sequence). Each 96 well plate was run twice through the QPCR equipment and similarly processed. Each plate was also prepared in duplicate and these data were similarly processed; the four $\widehat{I_j^*}(T)$ values thus amassed were finally averaged to create a single melting curve for each of the seven DNA hairpins under study.

B. Simulation bonding metrics and melting curve generation

We considered three different bonding metrics to determine the state of the single-stranded DNA hairpin in the simulation; that is, whether the hairpin was open or closed. All three metrics rely on the number of complementary in-

FIG. 3. The three metrics used to determine the “closed” state in the simulation data. Metrics 1 and 2 enumerate all stem possible bonding pairs (1) and aligned stem possible bonding pairs (2). Metric 3 only considers the hairpin system to be closed when all of the possible pairs in the aligned stems are bonded. The $A_5G_5T_5$ simulation data are presented for each of the three metrics with nondimensional temperature and normalized intensity. A sigmoidal curve is fit to the data from each metric.

interactions between the bases in the stem section of the nucleotide sequences. Complementary interactions were defined as follows: if two Watson–Crick complementary bases were found to be within a distance where the hydrogen bonding potential is effectively nonzero (since it decays quickly as a function of distance) then the two bases are considered bonded. This threshold distance is defined as $\sigma = 2^{1/6}$.

In metric 1 of Fig. 3, all possible bonding pairs are calculated every 1000 BD time steps. There is a problem with this metric, especially for the block copolymer-like sequences considered here. Since the model allows for multiple beads to bind to the same base bead via hydrogen bonding interactions, the number of bonding incidents described by this metric is often more than the number of bases in the stem section of the DNA hairpin. This artifact of two- and three-bead simulation models is prevalent throughout the literature.^{2,3,7,8,21,22,35} When we analyzed the instantaneous chain configurations, we found that the beads are often in a closed, zipperlike configuration with the base beads slightly out of alignment.³⁵ This simulation artifact is due to steric and hydrogen bonding stabilization of this closed state form. Although the model does allow multiple binding incidents to occur, the change in free energy in the misaligned zipper configuration is not simply equal to twice the energy in the aligned system because the distances between the bonded beads are different.

Due to the shortcomings of metric 1, we designed addi-

tional metrics that could characterize the bound state. Metric 2, depicted in Fig. 3, only considers bases to be bonded if they are paired “correctly” to lead to complete bonding in the stem. For example, for a sequence that is n bases long, the distance between the positions of the $i=1$ bead and $j=n$ bead is calculated (providing that the two bases in question are complementary) and, if this distance is less than σ , they are counted as bonded. This calculation continues for the $i=2$ and $j=n-1$ beads and so on up the stem of the hairpin. This metric thus eliminates the problem of double counting in the first metric.

However, this second metric may still not fully capture the experimental system due to the nature of an intercalating dye. If double stranded DNA is modeled as a ladder, then intercalating dyes like SYBR Green I fit between the rungs. Therefore, in order to bind to the DNA, at least two bases need to be bonded. Other simulation studies have used metrics that rely on contiguous bonded bases.^{7,8} As a result, we developed a third metric to capture this feature of the experiments. Metric 3, as depicted in Fig. 3, requires that all of the complementary aligned bases be bonded in order for the hairpin to be deemed closed; this final metric is effectively bimodal.

We computed the average metric value for each simulation temperature run by computing the time average of the number of bonds found at every saved simulation frame (1000 BD time steps $\approx 1 \mu\text{s}$). Each of the three simulation run replicates were then averaged together to find a single metric measurement for each nondimensional temperature. The MC algorithm transformed the data to produce a sigmoidal-shaped melting curve (with high temperature corresponding to high intensity) and normalized to a $[0,1]$ scale with 0 corresponding to a fully closed state and 1 corresponding to a fully open hairpin. Normalization of each of the metrics, as in Fig. 3, to the same $[0,1]$ scale not only allowed for the three metrics to be compared among one another, but also with the experimental data.

Figure 3 also shows that each bonding metric defines a slightly different sigmoidal-shaped melting curve. The effective melting-point temperature, T_{MP} , is defined where the chain has an equal probability of being open or closed. We obtain the melting temperature curve, T_M , by fitting the melting curve data with a sigmoidal distribution and finding the inflection point.

When we consider the sharpness of their transition regimes, we find that by construction metric 2 will form the sharpest melting curve, even though metric 3 is an “on-off” metric. (Metric 1, due to its multiple base bonding allowances, will have the broadest transition regime.) Imagine the base case sequence of $A_5C_5T_5$, which is comprised of stem length of five base pairs. In the completely closed state, all five pairs are bonded (to the aligned corresponding base, i.e., $A_1=T_{15}$) and the instantaneous value of both metrics 2 and 3 is 1. If one bond is lost along the chain, the instantaneous value of metric 2 will be equal to 0.8 while the instantaneous value of metric 3 will be 0. Assuming for the sake of argument, that the pair is bonded for half of the simulation time, the average value of metric 2 will be 0.9 and the average

FIG. 4. Plot of the normalized fluorescence intensity as a function of temperature for different DNA and dye concentrations with the sequence $A_{10}C_5T_{10}$. The (red) dashed line in the center plot was chosen as the experimental condition for all subsequent studies: 0.15 μM DNA solution and 2X SYBR Green I dye. This sets the ratio of dye molecules to stem base pairs at 2X: 4.5×10^{13} stem base pairs.

value of metric 3 will be 0.5. Therefore, with the bimodal metric 3, the transition regime will similarly become broader near the melting-point temperatures.

IV. RESULTS AND DISCUSSION

A. Optimal experimental conditions

We found that the concentrations of the DNA and the SYBR I dye solution can shift the measured curve by as much as 18 K over the range of SYBR I concentrations (0.02X–10X) and the DNA concentrations (0.02–0.2 μM) examined here, consistent with other studies.^{36,37} To determine the experimental conditions that maximize the amount of useful data, we first sorted the mfold predicted melting-point temperatures (which ranged from 319.1 to 347.1 K) for all of the DNA hairpins in Table I.³⁰ The parameters chosen for the mfold investigation were 1M monovalent sodium ions, which is similar to the 0.5M sodium and 0.5M potassium monovalent ions primarily comprising buffer A. The mfold melting-point temperatures are summarized in Table I. The DNA hairpin sequence $A_{10}C_5T_{10}$ has the median mfold predicted T_{MP} of 338.5 K. We then examined this sequence using various combinations of the DNA and dye concentrations described above. Figure 4 shows nine of the 30 results thus obtained; the optimal set of concentrations for this se-

quence was found to be 2X SYBR I and 0.15 μM of $A_{10}C_5T_{10}$ DNA solution. We found that the melting-point temperature $T_{\text{MP}}^{\text{exp}} = 337$ K, under these conditions, is closest to the predicted mfold melting-point temperature $T_{\text{MP}}^{\text{mfold}} = 338.5$ K.³⁰

By generating these data with control over the specific experimental conditions and concentrations, we are better able to interpret the raw data and process it in a manner that provides a good correspondence with the computational simulations. In addition, after understanding that the specific experimental conditions (such as reagent concentrations) can shift the melting curve by as much as 5%, we found it vital to collect our own experimental data. Although a literature search can produce melting-point temperature data for a variety of DNA sequences, a highly reliable, self-consistent data set over the entire temperature range is necessary to arrive at meaningful conclusions about the quality of the simulation data.

B. Optimal metric and bonding potential exploration

In order to compare the experimental and simulation data, we first needed to determine both the conversion between simulation and experimental temperatures, T_{scale} , and the best ratio of stacking to bonding energy, $\epsilon_{\text{stack}}/\epsilon_{\text{Hbond}}$. These two mutually dependent simulation parameters determine how well the simulation data matches the experimental data. We used the $A_5C_5T_5$ sequence to determine the best choices for these parameters because it is computationally efficient and the synthesis yield is high.

First, the three different metrics, as defined in Sec. III B, provide a spectrum of descriptions for the closed hairpin state. The simulation data for the test sequence $A_5C_5T_5$ were processed by metrics 1, 2, and 3 for each of the $\epsilon_{\text{stack}}/\epsilon_{\text{Hbond}}$ values. This produces a series of curves for the $A_5C_5T_5$ sequence in nondimensional temperature space, similar to that of Fig. 3. Recall that we kept the ratio of the stacking and hydrogen bonding energies constant at the thermodynamic value³⁴ of 2.5, but allowed the magnitude of these quantities to change. Due to the large number of simulation runs at every temperature that are needed to create a hairpin melting curve, only three values were investigated: $\epsilon_{\text{stack}}/\epsilon_{\text{Hbond}} = 2.5/1$, $5/2$, and $10/4$. This allowed the base specific potentials to vary with respect to the other bead potentials (such as spring stiffness and bending potentials) in the system.

Next, the simulated melting curves for the different metrics were shifted by some $T_{\text{scale}} = dT_{\text{exp}}/dT_{\text{sim}}$ which converts between the nondimensional simulation temperature and the dimensional experimental temperature schemes,

$$\frac{dI_{\text{sim}}}{dT_{\text{sim}}} = \left(\frac{dI_{\text{exp}}}{dT_{\text{exp}}} \right) \left(\frac{dT_{\text{exp}}}{dT_{\text{sim}}} \right) = \left(\frac{dI_{\text{exp}}}{dT_{\text{exp}}} \right) T_{\text{scale}}. \quad (12)$$

The T_{scale} value for each metric melting curve was chosen so that the simulated melting-point temperature for this metric, $T_{\text{MP}}^{\text{sim}}$, matches to the experimental melting-point temperature $T_{\text{MP}}^{\text{exp}} = T_{\text{MP}}^{\text{sim}}(\text{metric 1}) = T_{\text{MP}}^{\text{sim}}(\text{metric 2}) = T_{\text{MP}}^{\text{sim}}(\text{metric 3})$. The data will be presented in degrees Kelvin due to the conventions in the physics community, however the conversion factor was computed in degrees Celsius, as is common in biol-

FIG. 5. Plot of the three metrics with the experimental data overlaid for the $A_5C_5T_5$ base case sequence. Metric 2 best captures the slope in the transition regime of the experimental data.

ogy. As an example, we present in Fig. 5 the $A_5C_5T_5$ sequence using each of the three metrics for the $\epsilon_{\text{stack}}/\epsilon_{\text{Hbond}}=5/2$.

To quantify the fit of the three melting metrics depicted in Fig. 5 (with the bonding potential of $\epsilon_{\text{stack}}/\epsilon_{\text{Hbond}}=5/2$), we calculated the coefficient of multiple determination adjusted for the number of parameters in the sigmoidal model, R_a^2 , values. We repeated the above process with the $\epsilon_{\text{stack}}/\epsilon_{\text{Hbond}}=2.5/1$ and $10/4$ bonding potentials. It should be noted that for $\epsilon_{\text{stack}}/\epsilon_{\text{Hbond}}=2.5/1$, some error in very low temperature data is expected. In this regime, extremely long simulation equilibration times are needed due to the low thermal energy of the system. The resulting T_{scale} and R_a^2 values for each bonding potential and metric are summarized in Table III. From these data, we concluded that the $\epsilon_{\text{stack}}/\epsilon_{\text{Hbond}}=5/2$ and metric 2 are the optimal bonding potential and metric, respectively. This choice produces a $T_{\text{scale}}=220^\circ\text{C}$ and $R_a^2=0.998$ for the base case sequence. These parameters and metric definition will be used to evaluate the model.

C. Conversion from dimensionless temperature and evaluation of the model

With $T_{\text{scale}}=220^\circ\text{C}$, a simulation temperature of 0.3 corresponds to an experimental temperature of approxi-

TABLE III. Summary of the T_{scale} and R_a^2 values for each of the $\epsilon_{\text{stack}}/\epsilon_{\text{Hbond}}$ values and metrics examined in the study. Row 2 is depicted graphically in Fig. 5 and the center, bolded cell contains the optimal values utilized for the remainder of the investigation.

	Metric 1	Metric 2	Metric 3
$\frac{\epsilon_{\text{stack}}}{\epsilon_{\text{Hbond}}} = \frac{2.5}{1}$	$T_{\text{scale}}=400$ $R_a^2=0.399$	$T_{\text{scale}}=430$ $R_a^2=0.680$	$T_{\text{scale}}=470$ $R_a^2=0.635$
$\frac{\epsilon_{\text{stack}}}{\epsilon_{\text{Hbond}}} = \frac{5}{2}$	$T_{\text{scale}}=205$ $R_a^2=0.416$	$T_{\text{scale}}=\mathbf{220}$ $R_a^2=\mathbf{0.998}$	$T_{\text{scale}}=280$ $R_a^2=0.775$
$\frac{\epsilon_{\text{stack}}}{\epsilon_{\text{Hbond}}} = \frac{10}{4}$	$T_{\text{scale}}=170$ $R_a^2=0.451$	$T_{\text{scale}}=190$ $R_a^2=0.712$	$T_{\text{scale}}=200$ $R_a^2=0.603$

FIG. 6. Comparison of the metric 2 simulation data and experimental sigmoidal fit curve with a $T_{\text{scale}}=220^\circ\text{C}$ factor. Each of the seven investigated sequences is depicted and sorted into their original classes. The R_a^2 values of these fits are reported in Table IV.

mately 340 K (66°C). Although the thermodynamic estimate for the conversion factor in Sec. II B ($T_{\text{scale}}=370^\circ\text{C}$) is different, the value of T_{scale} found with the present method corresponds to our particular experimental system. Indeed, this experimental approach allows the model to be adjusted for any biological condition.

Let us now see if the value $T_{\text{scale}}=220^\circ\text{C}$ leads to similar agreement (using metric 2) for the other sequences. As seen in Fig. 6, there is qualitative agreement. Unfortunately, the R_a^2 values reported in Table IV indicate a problem with the model. With sequences similar to the base case of $A_5C_5T_5$, such as the $A_5C_{10}T_5$ sequence, we see high correspondence between the simulated curves and the experimental data. We find somewhat reduced matching with sequences containing polyguanine sequences such as $A_5G_5T_5$, though 94% of the simulation data for these sequences lie within one temperature standard deviation of the median experimental

TABLE IV. Summary of R_a^2 values for $T_{\text{scale}}=220^\circ\text{C}$ for each of the sequences examined.

Sequence	R_a^2
$A_5C_5T_5$	0.998
$A_5C_{10}T_5$	0.958
$A_7C_5T_7$	0.471
$A_7C_{10}T_7$	0.590
$A_5G_5T_5$	0.632
$A_5G_{10}T_5$	0.736
$G_5A_{10}C_5$	0.545

FIG. 7. Comparison of the metric 2 simulation data and experimental sigmoidal fit curve for the $A_7C_5T_7$ sequence with $T_{\text{scale}}=200$ °C, which is the conversion required to match the melting-point temperature. The R_a^2 value is 0.707.

data value. It is important to note that the “shoulder” regions of the melting curve, which is the initial transition from fully closed to 10% open and from fully open to 10% closed, show the greatest degree of mismatch to the experimental data. Finally, as the sequence stem gets longer, for example, in the $A_7C_5T_7$ sequence, the agreement between the simulation and experimental data is again reduced. If we take the temperature range of the experimental data, depicted in Fig. 2(d) by the horizontal bars, we find that 91% of the averaged metric 2 simulation data fall within the measured spread.

Perhaps the low R_a^2 values in Table IV indicate that we chose the wrong base case. To test this possibility, we determined the optimal T_{scale} values for each sequence independently. However, this does not lead to all sequences having a value $R_a^2 > 0.9$, which we would consider to be a good fit. Indeed, simply changing the choice of T_{scale} to get the right melting-point temperature, which is the usual approach in matching simulation to experiment, does not imply that the simulation will then mimic the full behavior of T_M . This is exemplified by the data for the $A_7C_5T_7$ sequence in Fig. 7. Longer stem lengths are characteristic of this (and the other poor performing class I molecule $A_7C_{10}T_7$) and may explain the broadening of the transition regime.

Qualitatively poor fits point toward parameters and features that may need to be adjusted or added to the two-bead DNA model. Although the DNA model includes potentials describing both hydrogen bonding and nucleotide base stacking, it does not include the effects of cross stacking interactions between bases.¹⁰ Due to the fact that we see diminished R_a^2 values for sequences with longer stems, and thus more cross stacking interactions, we may need to add this feature to the model. From additional examination of Fig. 6, we see that sequences with polyguanines also have reduced fits. While the two-bead model includes pairwise interactions, it does not contain more complex (i.e., four base coordination) interactions that would be needed to describe the features of G-quartets.³⁸ The mismatch in the simulation and experimental systems with the $A_5G_5T_5$, $G_5A_{10}C_5$, and $A_5G_{10}T_5$ may be due to this lacking G-quartet coordination. Further study with incorporation of both cross stacking and multibase bead interactions features could be conducted to determine if the experimental data can be better captured.

V. CONCLUSION

We provided a comparison between systematic melting experiments of DNA hairpins in a common biological buffer and the predictions of a coarse-grained single-stranded DNA Brownian dynamics model. By employing the method described here, we were able to tune and extract important parameters of our system, convert between the nondimensional world of simulations and the dimensional experimental realm, and evaluate several definitions of melting. The second metric of Fig. 3 best captured the experimental system. This metric represents the state of the hairpin by the fraction of correctly bonded base pairs in the stem.

Instead of seeking only to mimic the melting-point temperature of the DNA sequence, we instead hoped to capture the entire dependence of the melting temperature curve from the fully closed state at low temperatures to the fully open state at high temperatures. Our analysis showed that the simplistic two-bead DNA model fails. However, the approach used here (and the concomitant experimental data) could be used to examine other models. Such comparison and model validation are important if coarse-grained DNA models are to be used in increasingly more complex *in vitro* and *in vivo* environments.

ACKNOWLEDGMENTS

We acknowledge numerous discussions on modeling single-stranded DNA with Martin Kenward. This work was supported by the International Human Frontiers Science Program Organization and a Biotechnology Training Grant from the NIH (Grant No. 5T32GM008347-20).

- ¹E. Carlon, E. Orlandini, and A. L. Stella, *Phys. Rev. Lett.* **88**, 198101 (2002).
- ²B. Mergell, M. R. Ejtehadi, and R. Everaers, *Phys. Rev. E* **68**, 021911 (2003).
- ³M. Sales-Pardo, R. Guimera, A. A. Moreira, J. Widom, and L. A. N. Amaral, *Phys. Rev. E* **71**, 051902 (2005).
- ⁴A. Jayaraman, C. K. Hall, and J. Genzer, *Biophys. J.* **91**, 2227 (2006).
- ⁵G. Zheng, L. Czapla, A. R. Srinivasan, and W. K. Olson, *Phys. Chem. Chem. Phys.* **12**, 1399 (2010).
- ⁶T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, *Phys. Rev. Lett.* **104**, 178101 (2010).
- ⁷K. Drukker and G. C. Schatz, *J. Phys. Chem. B* **104**, 6108 (2000).
- ⁸K. Drukker, G. Wu, and G. C. Schatz, *J. Chem. Phys.* **114**, 579 (2001).
- ⁹S. P. Mielke, N. Grønbech-Jensen, V. V. Krishnan, W. H. Fink, and C. J. Benham, *J. Chem. Phys.* **123**, 124911 (2005).
- ¹⁰M. Kenward and K. D. Dorfman, *J. Chem. Phys.* **130**, 095101 (2009).
- ¹¹M. Levitt, in *Cold Spring Harbor Symposium on Quantitative Biology* (Cold Spring Harbor Press, Cold Spring Harbor, 1983), Vol. 47, pp. 251–262.
- ¹²B. Tidor, K. K. Irikura, B. R. Brooks, and M. Karplus, *J. Biomol. Struct. Dyn.* **1**, 231 (1983).
- ¹³D. L. Beveridge and G. Ravishanker, *Curr. Opin. Struct. Biol.* **4**, 246 (1994).
- ¹⁴F. Zhang and M. A. Collins, *Phys. Rev. E* **52**, 4217 (1995).
- ¹⁵N. Bruant, D. Flatters, R. Lavery, and D. Genest, *Biophys. J.* **77**, 2366 (1999).
- ¹⁶S. Sen and L. Nilsson, *J. Am. Chem. Soc.* **123**, 7414 (2001).
- ¹⁷T. E. Cheatham III and M. A. Young, *Biopolymers* **56**, 232 (2000).
- ¹⁸J. Norberg and L. Nilsson, *Acc. Chem. Res.* **35**, 465 (2002).
- ¹⁹S. Y. Ponomarev, K. M. Thayer, and D. L. Beveridge, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 14771 (2004).
- ²⁰H. L. Tepper and G. A. Voth, *J. Chem. Phys.* **122**, 124906 (2005).
- ²¹S. Buyukdagli, M. Sanrey, and M. Joyeux, *Chem. Phys. Lett.* **419**, 434 (2006).

- ²²T. A. Knotts IV, N. Rathore, D. C. Schwartz, and J. J. de Pablo, *J. Chem. Phys.* **126**, 084901 (2007).
- ²³M. Orozco, A. Noy, and A. Pérez, *Curr. Opin. Struct. Biol.* **18**, 185 (2008).
- ²⁴A. Noy, I. Soteras, F. J. Luque, and M. Orozco, *Phys. Chem. Chem. Phys.* **11**, 10596 (2009).
- ²⁵A. Morriss-Andrews, J. Rottler, and S. S. Plotkin, *J. Chem. Phys.* **132**, 035105 (2010).
- ²⁶M. S. Causo, B. Coluzzi, and P. Grassberger, *Phys. Rev. E* **62**, 3958 (2000).
- ²⁷Y. L. Chen, H. Ma, M. D. Graham, and J. J. de Pablo, *Macromolecules* **40**, 5978 (2007).
- ²⁸A. Izmitli, D. C. Schwartz, M. D. Graham, and J. J. de Pablo, *J. Chem. Phys.* **128**, 085102 (2008).
- ²⁹D. Marenduzzo, S. M. Bhattacharjee, A. Maritan, E. Orlandini, and F. Seno, *Phys. Rev. Lett.* **88**, 028102 (2001).
- ³⁰M. Zuker, *Nucleic Acids Res.* **31**, 3406 (2003).
- ³¹N. Carmi, S. R. Balkhi, and R. R. Breaker, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 2233 (1998).
- ³²S. W. Santoro and G. F. Joyce, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 4262 (1997).
- ³³M. Kenward and K. D. Dorfman, *Biophys. J.* **97**, 2785 (2009).
- ³⁴V. A. Bloomfield, D. M. Crothers, and I. Tinoco, *Nucleic Acids: Structures, Properties, and Functions* (Univ. Science, Sausalito, CA, 2000).
- ³⁵E. J. Sambriski, D. C. Schwartz, and J. J. De Pablo, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 18125 (2009).
- ³⁶K. M. Ririe, R. P. Rasmussen, and C. T. Wittwer, *Anal. Biochem.* **245**, 154 (1997).
- ³⁷R. H. Lipsky, C. M. Mazzanti, J. G. Rudolph, K. Xu, G. Vyas, D. Bozak, M. Q. Radcliff, and D. Goldman, *Clin. Chem.* **47**, 635 (2001).
- ³⁸A. T. Phan and J. L. Mergny, *Nucleic Acids Res.* **30**, 4618 (2002).