AIP | The Journal of Chemical Physics

Coarse-grained modeling of DNA oligomer hybridization: Length, sequence, and salt effects

Daniel M. Hinckley, Joshua P. Lequieu, and Juan J. de Pablo

Citation: The Journal of Chemical Physics **141**, 035102 (2014); doi: 10.1063/1.4886336
View online: http://dx.doi.org/10.1063/1.4886336
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/141/3?ver=pdfcov
Published by the AIP Publishing

Coarse-grained modeling of DNA oligomer hybridization: Length, sequence, and salt effects

Daniel M. Hinckley,[1] Joshua P. Lequieu,[2] and Juan J. de Pablo[2,3,a)]
[1]*Department of Chemical and Biological Engineering, University of Wisconsin–Madison, Madison, Wisconsin 53706, USA*
[2]*Institute for Molecular Engineering, University of Chicago, Chicago, Illinois 60637, USA*
[3]*Materials Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA*

(Received 30 January 2014; accepted 17 June 2014; published online 15 July 2014)

A recently published coarse-grained DNA model [D. M. Hinckley, G. S. Freeman, J. K. Whitmer, and J. J. de Pablo, J. Chem. Phys. **139**, 144903 (2013)] is used to study the hybridization mechanism of DNA oligomers. Forward flux sampling is used to construct ensembles of reactive trajectories from which the effects of sequence, length, and ionic strength are revealed. Heterogeneous sequences are observed to hybridize via the canonical zippering mechanism. In contrast, homogeneous sequences hybridize through a slithering mechanism, while more complex base pair displacement processes are observed for repetitive sequences. In all cases, the formation of non-native base pairs leads to an increase in the observed hybridization rate constants beyond those observed in sequences where only native base pairs are permitted. The scaling of rate constants with length is captured by extending existing hybridization theories to account for the formation of non-native base pairs. Furthermore, that scaling is found to be similar for oligomeric and polymeric systems, suggesting that similar physics is involved. © *2014 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4886336]

I. INTRODUCTION

At the proper solution conditions, complementary strands of single-stranded DNA (ssDNA) self-assemble into double-stranded DNA (dsDNA). This process, DNA hybridization, is enabled by the formation of Watson-Crick (W-C) base pairs between the two complementary strands. DNA hybridization is central to biology, and has gradually become an integral part of nanotechnology, where the reversibility and specificity of this reaction can be harnessed to assemble intricate structures.[1–4] Examples include the use of DNA oligomers, 10–40 base pairs (bps) in length, as linkers,[5] staples,[6] or fuel.[7] As important as DNA hybridization has become in nanotechnology, relatively little is known about the mechanism and the actual pathways of hybridization.

Early pioneering experiments examined the mechanism of DNA hybridization via changes in hyperchromicity during temperature jump experiments.[8–11] While these experiments were non-disruptive, they provided macroscopic information from which reaction mechanisms were gradually inferred. More recent measurements have relied on fluorescence resonance energy transfer[12,13] (FRET) to probe hybridization. FRET has the ability to probe the molecular processes leading to hybridization, but it is important to ensure that the inclusion of fluorescent labels does not influence the mechanism of hybridization. Perhaps more importantly, as informative as FRET has been for the study of DNA, it does not provide the picosecond temporal and nanometer spatial resolution required to ascertain the full details of DNA hybridization.

Over the years, a number of mechanisms have been proposed to account for the experimental observations of DNA hybridization. An "all-or-none" hybridization hypothesis was originally proposed based on observed second-order kinetics[8,9] for DNA polymers. Multiple data sets have corroborated the general elements of that hypothesis,[14] but the molecular origins of the all-or-none process have not been demonstrated. More detailed mechanisms have been proposed to account for the possibility of differing rate-limiting steps[10,11] or out-of-register, staggered intermediate states[10,15] in oligomeric systems. However, such theories cannot explain the scaling of rate constants as a function of sequence and ionic strength.[10,12,15–17] The shortcomings of these theories could potentially be addressed if the fast processes that occur in hybridization were not masked by the slower processes that are sampled by kinetic relaxation measurements. Additional microscopic measurements are needed to develop a complete picture of DNA oligomer hybridization.

In this work, we resort to molecular simulations to explain DNA hybridization. While several levels of description are available and could be adopted, atomistic models[18,19] cannot currently provide a statistically meaningful description of the ssDNA–dsDNA transition because of the prohibitive cost associated with modeling the solvent. Implicit solvent methods[20] could be adopted to decrease the computational effort considerably, but such models have not been calibrated to provide a reliable description of DNA thermodynamics and dynamics.

Coarse-grained (CG) models, which represent DNA using a reduced number of interaction sites, have gradually gained acceptance because they enable simulation of statistically relevant ensembles of trajectories while preserving a

a)Author to whom correspondence should be addressed. Electronic mail: depablo@uchicago.edu

molecular-level description of the process. These ensembles can then be used to examine mechanisms of hybridization. Furthermore, recent CG models have been parameterized to capture macroscopic observables in a top-down manner, and can be compared readily to experiment.

Several CG models have been proposed;[21–32] in this work, we adopt the 3SPN.2 model.[32] This CG model includes explicit electrostatic interactions along the backbone, and has been parameterized to be consistent with experimentally-measured structural, mechanical, thermodynamic, and kinetic properties. In particular, rate constants were calculated for the association of 25 bp sequences with various degrees of complementarity, thus demonstrating that 3SPN.2 can be used to study DNA hybridization. It is used here to investigate the hybridization behavior of a wide array of sequences, and to examine the scaling of rate constants as a function of oligomer length and sequence identity. The effect of ionic strength on the mechanisms of hybridization and scaling of rate constants is also addressed.

The paper begins with a description of the sequences considered, the model, and the methods used to simulate DNA hybridization. Results are then presented to establish consistency between various sampling methods. A second section considers the effect of oligomer sequence homogeneity, length, and ionic strength on rate constants, followed by a discussion of results in the context of available experimental data and the modeling work of others.

II. METHODS

A. Sequences and solution conditions

Three types of sequences are simulated in this work: H, poly(A), and poly(AC); e.g., three types of sequences are simulated in this work: poly(A), poly(AC), and H. The 30 base sequences are listed in Table I. The sequences were selected to test for distinct mechanisms of hybridization while simultaneously demonstrating a low propensity to form secondary structures such as hairpins. It has been observed by others that CG content has a negligible effect on rate constants;[14] however, in that study the range of CG contents examined was relatively small (34%–50%). Here, we use a homogeneous sequence with 0% CG content, as well as two sequences with 50% CG content, one consisting of a repetitive motif (AC) and the other displaying complete heterogeneity.

B. 3SPN.2 DNA model

The 3SPN coarse-grained model was originally developed to study equilibrium properties of DNA.[21,22] It uses three sites to separately represent the sugar, phosphate, and base as shown in Fig. 1. The latest version of the model accounts for the structural, mechanical, thermodynamic, and kinetic properties of single and double-stranded DNA, including reaction rate constants.[32] It has been implemented in LAMMPS[33] and is available for download.[34]

There are three aspects of the 3SPN.2 model that are worth emphasizing in the context of the current work and our own past simulations of DNA hybridization. First, anisotropic non-bonded potentials are used to account for base stacking, base pairing, and cross stacking. These angle-dependent interactions preserve distinct major and minor grooves throughout our simulations, a feature that is absent from other models. A local Gaussian well along the backbone enforces an equilibrium dihedral angle while simultaneously permitting rotation around the backbone to capture the flexibility of ssDNA. The DNA molecule is charged but, motivated by Oosawa-Manning theory,[35,36] the effective charge of the phosphate sites is modified from -1 to -0.6. The inclusion of charge on the phosphate allows the model to examine the subtle effects of electrostatics on hybridization. Note that the electrostatic interactions are treated at the level of Debye-Hückel (DH) theory. While this approximation can be justified at ionic strengths below 100 mM, it is not valid at high ionic strength. Here, we assume that at high ionic strengths electrostatic interactions are approximately correct and play a secondary role in determining the pathway of hybridization. This assumption is examined in greater detail in Secs. III A and III D.

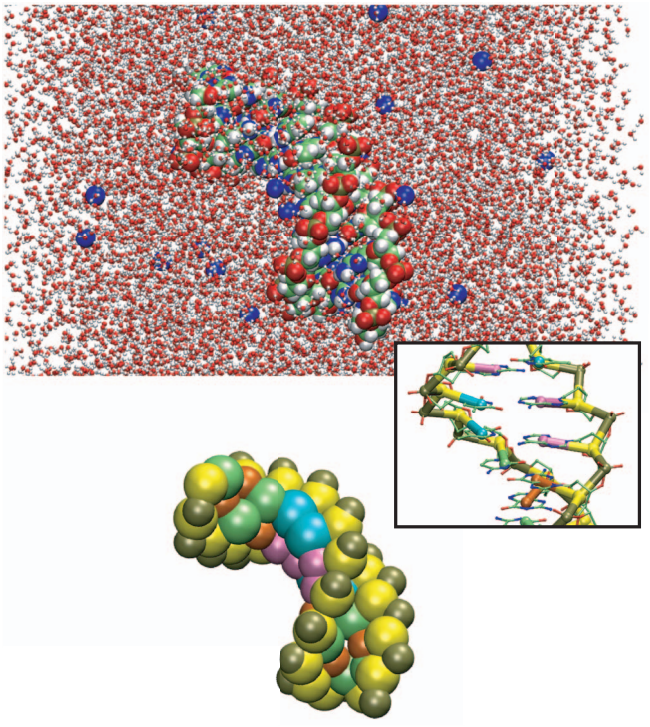

FIG. 1. (Top) All-atom system of Drew-Dickerson Dodecamer (5′-CGCGAATTCGCG-3′) containing DNA, sodium counter-ions, and water. (Bottom) The all-atom system using 3SPN.2 methodology. Note that the water and counter-ions are not explicitly modeled. (Inset) Schematic representation of the mapping between the all-atom and coarse-grained representations of DNA.

TABLE I. Sense sequences used for studying the physics of DNA hybridization. For sequences with $N < 30$, the first N bases from the 5′ end are used for simulations. The antisense sequences were perfect complements of the sense strands.

Name	Sense sequence
poly(A)	5′-AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA-3′
poly(AC)	5′-ACA CAC ACA CAC ACA CAC ACA CAC ACA CAC-3′
H	5′-TCG GAG AAA TCA CTG AGC TGC CTG AGA AGA-3′

DNA simulations are performed in the *NVT* ensemble with a Langevin dynamics thermostat.[37] In Langevin dynamics, a friction coefficient is assigned to each site to capture drag on the moiety as it moves through the solvent. The friction coefficients of each site are equal, and are assigned using a free-solution diffusion constant from ssDNA.[38] The diffusivity is scaled with temperature according to the Einstein relation, and sequence length according to $N^{-0.68}$.[38,39] This approach allows implicit treatment of the solvent and permits simulation of the large systems needed for analysis of rate constants. Importantly, the random nature of the integrator makes it possible to generate distinct trajectories from the same starting configuration.

C. Forward flux sampling

DNA hybridization is a rare event; large free energetic barriers separate the hybridized and dehybridized states. Even at high ionic strengths, the probability that an association event leads to hybridization is small. As such, DNA hybridization is not amenable to direct simulation techniques, and enhanced sampling methods that preferentially sample transitions between initial and final states are needed.

Past studies of DNA hybridization have resorted to umbrella sampling (US),[40,41] transition path sampling (TPS),[42–44] and forward flux sampling (FFS).[32,45] In this work, we use FFS, which calculates the flux of a system moving forward through order parameter space. By doing so, it is possible to calculate rate constants that can be directly compared to those measured in experiment. In FFS, a series of non-intersecting interfaces λ_i are introduced; these interfaces separate the initial and final states. Given an initial reacting flux Φ_0 through the first interface λ_0, the reaction probability, or the probability of advancing to the last interface λ_n, can be calculated as the product of a set of conditional probabilities

$$k = \Phi_0 P\left(\lambda_n \mid \lambda_0\right) = \Phi_0 \prod_{i=0}^{n-1} P\left(\lambda_{i+1} \mid \lambda_i\right). \quad (1)$$

Several variants to FFS have been proposed; in this work, we use one in which the contribution of each FFS trajectory at each interface is governed by its "Rosenbluth" weight. Rosenbluth FFS (or RFFS) generates reaction probabilities using an ensemble of independent, unbranched reactive trajectories. In all of the results presented here, at least 200 independent simulations were performed, with 20 trials at each interface. "Direct" FFS, another commonly used approach, generates reactive trajectories that can also overcome large energy barriers but the reactive paths arise from branched trajectories, leading to reduced sampling of the ensemble. For a discussion of the differences between variants of FFS, readers are referred to the original literature on FFS.[46,47]

In this work, the reacting flux is calculated using bimolecular reaction theory,[48] as explained in Ref. 32. The interfaces are defined as a linear function of center-of-mass (COM) separation δ_{COM}, the number of bases N in each hybridizing oligomer, and the number of inter-strand base pairs i,

$$\lambda_i(N, \delta_{\mathrm{COM}}) = \frac{N}{\Delta(N)}\delta_{\mathrm{COM}} + \frac{2N}{N+1}i - N, \quad (2)$$

with the length-dependent distance $\Delta(N)$ defined as

$$\Delta(N) = (5.47N^{0.794} + 49.5)\ \text{Å}. \quad (3)$$

These interfaces are depicted graphically in Fig. 2. We note that the number of interfaces is arbitrarily set equal to the number of bases in each sequence. The conditional probabilities calculated using such interfaces produce "probability curves" that can reveal differences in reaction mechanisms. A representative curve is depicted in the inset of Fig. 2. The significance of the features in such curves is discussed in greater detail in Secs. III A and III B.

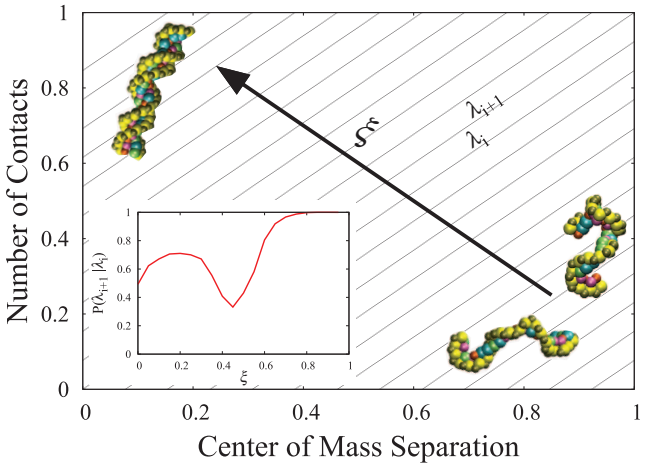

FIG. 2. Schematic representation of interfaces λ_i used in the Forward Flux Sampling calculations. ξ represents the normalized interface number. (Inset) A representative probability curve depicting the variation of conditional probabilities $P(\lambda_{i+1}|\lambda_i)$ through interface space.

III. RESULTS

A. Validation

We begin by demonstrating the consistency between the various methods implemented here, and by showing that 3SPN.2 does indeed capture hybridization trends measured experimentally. First, we seek consistency between Rosenbluth and Direct FFS. Then we examine the dependence of hybridization rate constants on ionic strength.

Figure 3(a) shows that both FFS techniques are in good agreement. Direct FFS exhibits a larger statistical uncertainty than Rosenbluth FFS, due to a reduced number of independent trials at each interface. The probability curves calculated at the initial interfaces are in good agreement with the probabilities calculated using first passage probabilities.[50] These first passage probabilities represent the random, diffusive behavior of non-interacting point particles. While ssDNA is not a point particle, we observe that the probabilities of visiting smaller values of δ_{COM} at early FFS interfaces are in excellent agreement with analytical probabilities from first passage theory. Therefore, the curve labeled "random diffusion," calculated using a theory for point particles, represents the results

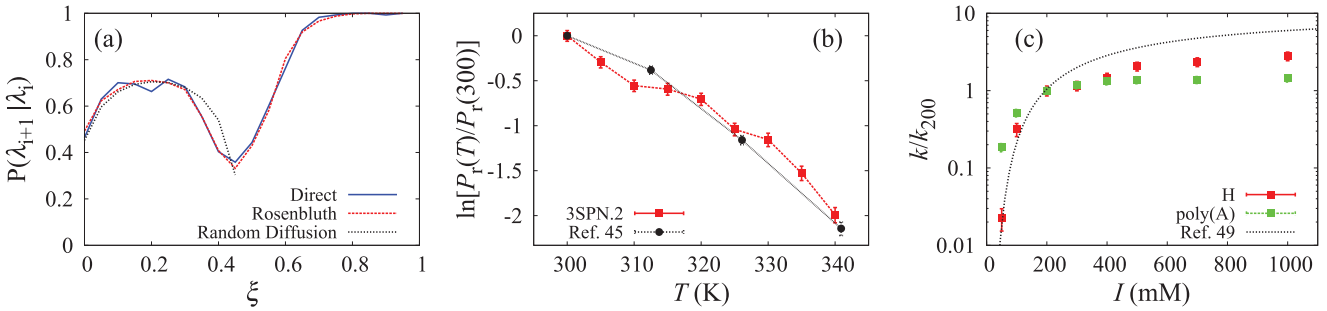

FIG. 3. (a) Consistency between probability curves extracted from Direct and Rosenbluth Forward Flux Sampling. The results are also compared with the first passage probabilities (random diffusion). ξ is the normalized interface number. ($N = 20$, $I = 1$ M $T = 321.3$ K). (b) Consistency between reaction probabilities $P_r = P(\lambda_n|\lambda_0)$ from 3SPN.2 and oxDNA[45] as a function of temperature for the same 14 bp sequence (5′-TATCTGGCTTGTAC-3′). Reaction probabilities from 3SPN.2 calculated using an ionic strength of 1 M. (c) Agreement between simulated relative rate constants for 20 bp poly(A) and H sequences and data from Ref. 49 as a function of ionic strength. Rates are normalized with respect to calculated rate constants at $I = 200$ mM. poly(A): $T = 294.4$ K, H: $T = 310.7$ K. Error bars represent the standard error.

obtained for randomly diffusing, non-interacting polymers. Note that the interfacial probability naturally curves downwards as a consequence of the interfaces being a function of δ_{COM} and the number of inter-strand base pairs formed. A positive deviation from the first passage probability is indicative of net attractive interactions arising from base pairing; negative deviations represent the effect of excluded volume or electrostatic repulsions.

Relative rate constants have been reported as a function of ionic strength by Britten and co-workers.[49] These data serve to validate the accuracy of the present treatment of electrostatics. In 3SPN.2, the solvent and ions are modeled implicitly at the level of DH theory; at an ionic strength of 1 M, DH theory is not expected to be quantitative, but it is consistent with the level of description adopted by most available CG DNA models. We chose to normalize the rate constants near physiological conditions ($I = 200$ mM) because the model is more likely to be used under such conditions. Figure 3(c) suggests that the model performs in a satisfactory manner. For the H sequence, the agreement with experiment is quantitative at low ionic strengths. However, at high ionic strengths the rate constant underpredicts the experimental data by a factor of two. This observation suggests that at high ionic strength DH overpredicts the repulsion between single-strands. In order to better understand these results, we are currently repeating these simulations with explicit ions, and the results will be presented in a forthcoming publication.

For the poly(A) sequence, the rate constants are less sensitive to ionic strength. This finding is consistent with experiments on RNA homo-oligomers,[16] which show an increase in rate constants by a factor of 5–10 over the range of ionic strengths shown in Fig. 3(c). On the basis of these results, we conclude that for the model considered here, our approximate treatment of electrostatic interactions is not quantitative at high ionic strengths, but does not lead to qualitative deviations from experiment.

B. Effect of sequence

Assisted by the probability curves introduced in Sec. II C, we begin by examining the effect of sequence on the mechanism of hybridization. We consider 20 bp segments, with the sequences given in Sec. II A. Longer sequences were not selected for this analysis due to the increased computational cost. Simulations were performed at 1 M ionic strength, consistent with early experimental work.[9]

First, as a control, we performed RFFS with two entirely non-complementary strands, i.e., poly(A)+poly(A). Such a system could never advance to the hybridized state (λ_n) because of its inability to form any base pairs. Next, we examined a system constrained in a manner that only allows native base pairs. By "native," we mean base pairs that form when the sense and antisense strands are perfectly aligned. Regardless of sequence, the resulting probability curves were almost identical. The minor differences that arise are due to the differing strength of the A–T and G–C base pairs and the intrastrand stacking interactions (see Ref. 32 for additional details regarding these interactions). Finally, the 20 bp segments of the poly(A), poly(AC), and H sequences were simulated while allowing non-native base pair formation. The resulting probability curves are shown in Fig. 4.

Several features of these curves are worth discussing. First, we see that at early stages all systems follow the diffusive probability curve. This is expected, given the shielding of electrostatic interactions at high ionic strength. We also find that the probability curves for the system with forbidden base pairs and only native base pairs follow each other closely for $\xi < 0.4$, suggesting that the ensembles of configurations at each

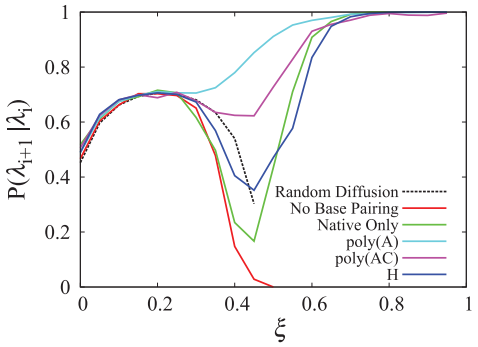

FIG. 4. Comparison of the probability curves calculated analytically (i.e., random diffusion) or using Forward Flux Sampling. N=20, $I = 1$ M, and $T = 321.3$ K. Error bars are omitted for clarity and can be found in the supplementary material[51] (see Fig. S2). $N = 20$, $I = 1$ M, and $T = 321.3$ K.

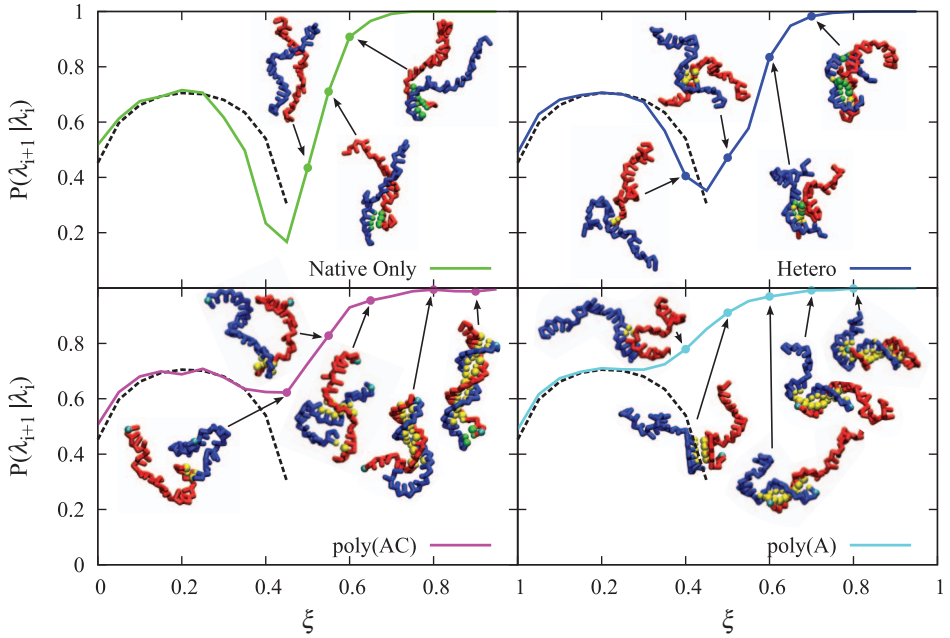

FIG. 5. Representative configurations and probability curves taken from Forward Flux Sampling calculations. The cyan sites represent the 5′ ends of the sense and antisense strands, green sites represent native base pairs, and yellow sites represent non-native base pairs. (Top left) The H sequence is constrained such that only native base pairs can be formed; (top right) the H sequence, (bottom left) the poly(AC) sequence, and (bottom right) the poly(A) sequence. $N = 20$, $I = 1$ M, and $T = 321.3$ K.

interface are similar. Most interestingly, we find that systems where non-native base pairs are permitted deviate positively from the systems where only native base pairs are permitted. This suggests that non-native base pairs play an important role in assisting oligomer hybridization.

Representative configurations are plotted in Fig. 5. These configurations provide several insights. First, non-native base pairs facilitate associative events by providing attractive interactions at larger strand-to-strand separations. These base pairs serve as anchor points that offset electrostatic repulsions while the complementary strands re-arrange to achieve more favorable alignments. As the base pairing interactions reduce the effective repulsion between the strands, the rate constant for hybridization increases. Second, the mechanisms of hybridization for DNA oligomers are highly sequence-specific. Heterogeneous, non-repetitive sequences follow a mechanism consistent with zippering, though non-native base pairs appear to stabilize intermediates along the reaction pathway. The repetitive poly(AC) system follows a rather complex mechanism, due to the ease with which it forms misaligned strands. We observe "inchworm" displacements, as well as a "pseudoknot" mechanism, consistent with the findings of a recent study.[45] These mechanisms involve the breaking of multiple base pairs and result in finite probabilities for the strands to disassociate; this is demonstrated by the non-unity conditional probabilities at values of ξ close to 1.0. We note that out-of-register conformations occasionally form between longer oligomers, resulting in configurations that do not fully anneal during the FFS calculations. The homopolymer poly(A) also forms misaligned strands with great ease. However, the mechanism of hybridization is not the same as that of poly(AC). In contrast to poly(AC), the complementary strands exhibit behavior similar to what was previously posited in the literature to be a "fast chain sliding" mechanism.[15] This mechanism is not exactly "slithering," as discussed in the work of Sambriski *et al.*;[42,43] rather, it most resembles the 1D diffusion of a defect that is formed by simultaneous breaking of several contiguous base pairs. This disruptive event originates from one of the dangling ends and propagates to the opposite end. Repeated propagation of defects eventually permits the system to arrive at the energy minimum.

Given that only limited information can be obtained from the probability curves and by inspection of individual trajectories, we now characterize the ensembles of suitably weighted configurations that populate each interface of the FFS calculations. The results are presented as contact maps, similar to those reported in previous calculations using 3SPN.1.[42] The contact maps are constructed for each sequence where non-native contacts were permitted, and are shown in Fig. 6.

It is clear that non-native base pairs, characterized by base pairs off the diagonal with non-zero probabilities, play an important role in the hybridization of each sequence. At small values of ξ, non-native base pairs tend to be located along the periphery of the contact map. This behavior has two origins: first, the relatively slow translational diffusion of the strands and the rapid conformational changes within each strand, particularly at the ends, lead to occasional base pair formation between strand ends. Second, the dependence of the order parameter ξ on δ_{COM} makes base pairs at the center of the strand less likely to form than those at the ends for small values of ξ. As ξ increases, base pairs begin to form between DNA, and as ξ approaches unity, the strands begin to become more fully aligned.

For each sequence, the contact maps suggest a "hybridization funnel," akin to that encountered in protein folding,[52] which leads to the final hybridization state. The H sequence appears to be characterized by a funnel that is

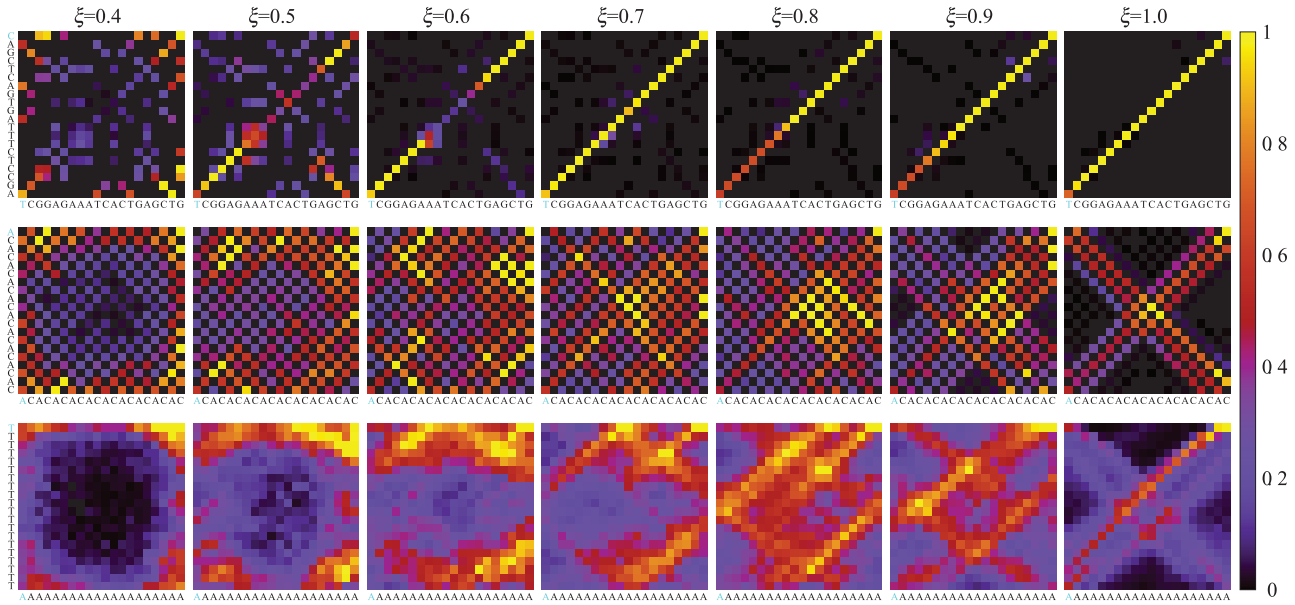

FIG. 6. Normalized probabilities of contact formation as a function of normalized interface number ξ for the H (top), poly(AC) (middle), and poly(A) (bottom) sequences. The sense sequence and antisense sequences are located on the x and y axes, respectively. The cyan letters represent the 5′ end of each strand. $N = 20$, $T = 321.3$ K, $I = 1$ M.

sharp and deep, as demonstrated by the relatively limited distribution of states as hybridization proceeds. In contrast, the poly(A) and poly(AC) sequences are characterized by a broader hybridization funnel, resulting in great diversity in the intermediate states that lead to hybridization. We emphasize that these intermediate states are not metastable; rather, they merely represent configurations along the pathway to the fully hybridized state.

It is also clear that 3SPN.2 predicts that some strands will associate in a parallel manner, instead of the traditional antiparallel helix. Parallel stranded DNA (psDNA) states are represented by contacts that form diagonal lines from the upper left to bottom right, while antiparallel stranded DNA (apsDNA) displays the opposite trend. This is not inconsistent with experiment, as psDNA has been observed to have stability similar to that of apsDNA.[53] However, such psDNA states populate free energetic minima that are not as deep as those for apsDNA, as observed using metadynamics simulations (see Fig. S3 in the supplementary material[51]). Palindromic sequences are key to the stabilization of such psDNA configurations. Indeed, the existence of a few palindromic tracts stabilizes a few psDNA configurations for the H sequence. The poly(AC) and poly(A) sequences are almost and entirely palindromic, respectively, and this leads to approximately equal weights of psDNA and apsDNA in the contact maps at high values of ξ. This point reveals a limitation of our order parameter: it samples trajectories that are entering a minimum characterized by small δ_{COM} and the formation of many base pairs. However, it cannot discern global and local minima.

C. Effect of length

The length-dependence of hybridization rate constants was originally examined by Davidson and Wetmur.[9] Based on their observations, they suggested that DNA hybridization is initiated by a nucleation event between complementary strands, in which a few (2–3) in-register base pairs are formed. If this nucleus is stable (i.e., not disrupted by thermal fluctuations or excluded volume repulsions), the remainder of the base pairs form in a rapid process known as zippering. Reduction of their original theoretical formalism (see the supplementary material[51] for additional details) leads to the following expression of the rate constant k, directly applicable to the computational experiments performed in the present work:

$$k \propto k_{\mathrm{N}} N. \tag{4}$$

In this equation, k_{N} is the nucleation rate constant and N is the length of the hybridizing oligomer. Experiments indicate that $k \propto N^{0.5}$ (Refs. 9 and 14). The inferred $N^{-0.5}$ scaling of k_N was proposed to be due to excluded volume interactions that impede complementary in-register base pairs from finding each other. However, the unavailability of sites due to steric hindrance violates one of the assumptions of the theory. Wetmur[54] also proposed that for short sequences ($L < 100$), steric constraints should not arise and k_{N} should be independent of sequence length. This hypothesis, however, has not been verified. A more nuanced theory[10] has been developed for oligomers, which hinges on the relative stability of the initial base pairs that form between complementary strands. As shown in the supplementary material,[51] that theory can be re-arranged to give a form analogous to Eq. (4).

Both of the aforementioned treatments assume that all bases are fully aligned when forming base pairs, i.e., they form "native" contacts. However, as demonstrated above, non-native base pairs play an important role in facilitating oligomer hybridization. It therefore becomes necessary to extend the aforementioned theories to predict the scaling of DNA hybridization rate constants as a function of sequence.

Like the previous theories, the scaling arguments developed here assume that the rate limiting step in the hybridization process is the formation of a nucleus of three contiguous W-C base pairs. This nucleus can form regardless of alignment of the hybridizing strands. We also assume that the DNA hybridization consists of two complementary ssDNA randomly diffusing until they are within a defined capture radius. Within that radius, there is a finite probability of proceeding to the dsDNA state. The probability of hybridization is affected by several length-dependent phenomena. The first is the translational diffusivity of each ssDNA molecule. As explained previously, the translational diffusivity of each ssDNA scales as $N^{-\kappa}$ where $\kappa = 0.68$.[38] The theory of Northrup *et al.*[48] used here assumes that hybridization is a diffusion-controlled reaction by incorporating the Smoluchowski result. The validity of this assumption is discussed below.

The scaling of hybridization rate constants with length is also influenced by the combinatorics associated with the alignment of the two hybridizing strands. This contribution, N^{Ω}, captures the possibility of nuclei forming between strands that are in- and out-of-register. In a random sequence, three contiguous base pairs will likely only form when the W-C base pairs are in-register, leading to $\Omega \approx 1$. In repetitive sequences, a stable nucleus can form with the strands both in- and out-of-register. This leads to many more possible nuclei and results in different scaling with $\Omega \approx 2$; the precise value and accompanying prefactors depend on the degree of repetitiveness and are explained in the supplementary material.[51] These combinatorics are analogous to the "complexity" introduced by Ref. 54, which is defined as the total number of base pairs in non-repeating sequences of an organism.

Finally, the structure of the polymer plays an important role in determining the accessibility of a nucleation site.[55,56] The contact or collision probability between different segments of a polymer can be expressed as a function of the Flory exponent ν and a contact probability θ. This leads to a $N^{-\nu\theta}$ contribution to the length-dependence of the rate constant.

The combination of the diffusion, alignment, and structural effects yields the scaling relation

$$k \propto N^{\Omega-(\kappa+\nu\theta)}. \tag{5}$$

3SPN.2 calculations of the radius of gyration of ssDNA predict a value of $\nu = 0.65$ (see Fig. S5 of the supplementary material[51]). Encouragingly, this value is consistent with the scaling exponent inferred from the experimental observations of Ref. 38. Values of θ have been derived[55] for end-to-end contacts, middle-to-end contacts, and middle-to-middle contacts. We use only the value for the last scenario ($\theta = 0.71$), as middle-to-middle nucleation events represent more than 80% of all those possible for the shorter oligomers examined here. As N increases, the fraction of middle-to-middle nucleation events approaches unity.

Hybridization rate constants were calculated for fragments of increasing length taken from the sequences listed in Sec. II A. The simulations were performed at high (1 M) ionic strength to be consistent with experiment[9] and at a constant temperature for each sequence type. In the past, rate constant measurements have often been carried out at a temperature 25 K below T_m of each sequence,[8] which is considered to be optimal for hybridization. As T_m is not a strong function of N in polymeric DNA, Hinnebusch *et al.*[57] performed hybridization experiments at a constant temperature when studying sequences 250–4100 bps in length. In contrast, for the small to intermediate length molecules considered here, T_m changes appreciably with N. Additionally, as shown in Fig. 3(b) and Fig. 7(b), 3SPN.2 does not display an "optimal" hybridization temperature; instead, DNA oligomer rate constants decrease monotonically with temperature below T_m. In order to facilitate comparison with data for longer DNA sequences, and to eliminate the effect of temperature on the scaling of rate constants, the simulation temperature for all sequences was fixed at 25 K below T_m of the longest sequence ($N = 30$), as predicted by Oligoanalyzer.[58]

The scaling of these rate constants is shown in Fig. 7(a). The predictions of Eq. (5) are included in the figure. The scaling law contains one adjustable prefactor, which was obtained by performing a least-squares fit to the data for the H sequence. After obtaining the prefactor, we find good agreement between the predicted and simulated rate constants for all three sequence types with $N = 30$. The predicted length scaling is also in good agreement with the simulated rate constants for H and poly(AC) sequences, serving to demonstrate the validity of the scaling relation. A slight discrepancy does appear between the poly(AC) simulations and the scaling law. This is most likely a consequence of the assumption that nuclei forming between strands out-of-register are equally likely

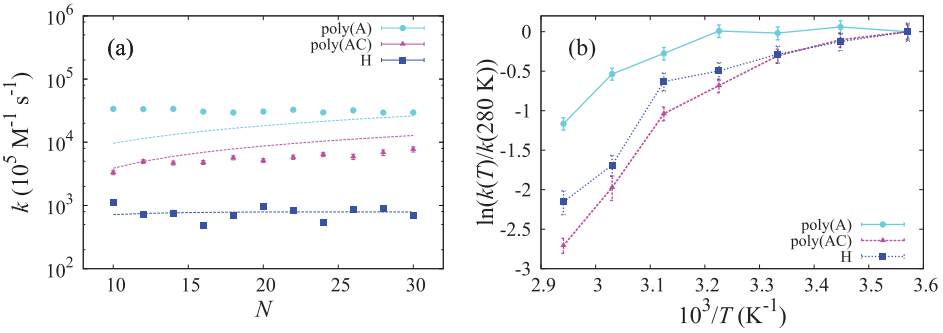

FIG. 7. (a) Hybridization rate constants and scaling theory as a function of sequence length. Points represent predictions from Forward Flux Sampling calculations; dashed lines show results from a scaling theory. The error bars represent the standard error of each data point. When not visible, the standard error can be approximated by the size of the symbol. poly(A): $T = 319.1$ K, poly(AC): $T = 331.0$ K, H: $T = 330.5$ K. $I = 1$ M. (b) Temperature-dependent scaling of hybridization rate constants for $N = 14$ sequences. $I = 1$ M.

to lead to duplex formation as nuclei formed between in-register strands.

The scaling of poly(A) is not predicted correctly. To investigate the source of this discrepancy, simulations were performed for each sequence type ($N = 14$) at various temperatures. The resulting temperature-dependence (Fig. 7(b)) shows that H and poly(AC) have similar dependence on temperature. As in Fig. 3(b), we observed non-Arrhenius behavior. We also find that hybridization rate constants of poly(A) are much less sensitive to temperature.

In Fig. 7(a), oligomers shorter than $N = 30$ are simulated with a lower degree of supercooling. However, poly(A) sequences appear to be much less susceptible to heightened temperatures. This leads to rate constants that are higher than expected, as observed in Fig. 7(a). To understand why this is, we look at the conformation of hybridizing homopolymers. These conformations (compared to other sequence types in Fig. S4 of the supplementary material[51]) show that associating poly(A)-poly(T) initially come together as a pseudo-globule. These globules exhibit increased stability when compared to initial hybridization configurations for H and poly(AC) sequences. This reveals a limitation of the 3SPN.2 model; the angle-dependent potentials modeling W–C base pairs often predict favorable interactions when adjacent bases of the same type are arranged in unphysical orientations.

Finally, we compare the scaling of simulated rate constants with length to experiment. We use the data from H sequences, as they are analogous to the random sequences discussed in Ref. 14. The results presented in Fig. 7(a) suggest that rate constants are roughly independent of N; however, experimental data for much longer sequences exhibit $N^{0.5}$ scaling. This difference between simulation and experiment can be reconciled by considering the relative influence of diffusion. The scaling theory presented earlier assumes that diffusive effects play a dominant role; however, calculated reaction probabilities from FFS are below 1%, indicating that the rate-limiting step is in fact the contact between the ssDNA, and that the contribution of diffusivity to length scaling can be ignored. If the effect of diffusion is removed from Eq. (5) ($\kappa = 0$), it leads to $N^{0.47}$ scaling, which is in good agreement with experiment. Fitting this relation to simulation data with diffusive effects removed also shows good agreement (see Fig. S6 of the supplementary material[51]), suggesting that the scaling theory presented here is plausible; note, however, that the small range of sequence lengths examined here and the uncertainty of the simulations precludes us from reaching a definite conclusion on this point.

D. Effect of ionic strength

Finally, we examine the effect of ionic strength on oligomer hybridization. This effect has been studied experimentally for long, polymeric DNA systems. At low ionic strengths ($I \leq 0.02$ M), the rate constant was found to be proportional to the cube of the ionic strength.[59] Between ionic strengths of 0.4 and 1.0 M, the hybridization rate constant was seen to increase by a factor of two.[9]

As depicted in Fig. 3(c), we can examine the evolution of the probability curves with increasing ionic strength (see Fig. 8). The probability curves reveal that the electrostatic repulsions between the complementary strands account for the majority of the dependence of rate constants on ionic strength (see $\xi = 0.2$ to $\xi = 0.4$). As the ionic strength decreases, the Debye length increases, leading to stronger electrostatic interactions. Thermal fluctuations must overcome this energetic barrier in order for base pairs to be formed.

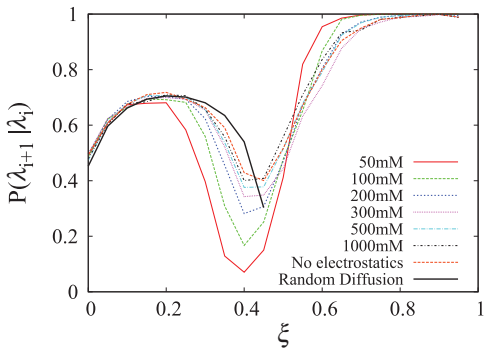

FIG. 8. Evolution of probability curves for H sequence as a function of ionic strength ($N = 20$, $T = 310.7$ K).

Interestingly, the probability curves climb more steeply in the vicinity of $\xi = 0.5$. As illustrated in Fig. 5, this region corresponds to the formation of the nucleus that precedes zippering. The trend can be understood in terms of the entropic penalty incurred upon hybridization. At low ionic strength, the hybridizing strands are more like rods than coils, due to the strong repulsion between phosphate sites along each strand. As a result, the loss of phase space incurred upon hybridization is smaller at low ionic strength than at higher ionic strength. The result is that the nucleus is effectively stabilized, and it is more likely for a nucleation event to lead to duplex formation.

The scaling of H sequence rate constants was also examined at low ionic strengths (see Fig. S7 of the supplementary material[51]). Our simulations predict rate constants a factor of 3–4 smaller than those at 1 M ionic strength. Due to the large uncertainty in the simulated rate constants, it is not possible to conclude that the scaling behavior differs from that at high ionic strength.

IV. DISCUSSION

First, we wish to compare the scaling of rate constants observed here with that observed in experiment. Craig *et al.*[10] studied self-complementary RNA block co-oligomers and did not observe monotonic scaling of rate constants for $N = 8-14$. Pörschke and Eigen[16] studied the hybridization of RNA homo-oligomers and observed that rate constants were "almost independent of chain length." This observation is consistent with the scaling behavior of poly(A) in Sec. III C and Fig. S6 of the supplementary material.[51] Further studies by Pörschke[15,17] examined RNA oligomers containing GC base pairs and homopolymers of RNA. In these studies, rates of hybridization were observed to decrease with increasing length[17] and behave non-monotonically.[15]

We do not see consistent trends in the scaling of H sequences over the range $N = 10-30$ bps. This is due, in part, to

the large variation in FFS probabilities from individual simulations (see Fig. S2 of the supplementary material[51]). As mentioned above, we are also constrained by the small range of sequence lengths examined here. FFS simulations of longer sequences are required to make definite conclusions but, unfortunately, they are beyond the reach of our current computational resources.

The scaling arguments developed in Sec. III C have limitations. The central assumption was that the formation of three contiguous W-C base pairs between hybridizing strands at any location would result in a fully formed duplex. This assumption implies that if two ssDNA come together misaligned, they can re-arrange via an "inchworm displacement" or "pseudoknot" mechanism without destabilizing the nascent duplex. This assumption is perhaps too liberal; simulations with another CG model[45] indicate that more than four contiguous base pairs are necessary to make it unlikely that misaligned strands detach. It was also assumed that the stability of the nucleus is independent of the length of the hybridizing strands. This assumption appears reasonable but merits further examination.

Finally, it was assumed that only one nucleation event occurs between the hybridizing strands. This assumption seems reasonable given the conclusion that DNA hybridization is reaction-controlled and the displacement mechanisms described in the literature.[45] However, the competition between displacement events has not been examined in great detail. Additional simulations examining the hybridization of long (>100 bps) ssDNA are necessary to explore these assumptions regarding the size, stability, and number of nucleation events.

It is also instructive to discuss these results in the context of previous studies examining DNA hybridization. Sambriski *et al.*[42,43] first considered DNA oligomer hybridization with the 3SPN.1 model.[22] It was observed that the predominant mechanism for hybridization consisted of states in which ssDNA strands were slightly offset or out-of-register. This mechanism was also observed for DNA hybridization on a surface.[40,41,44] This behavior was the result of an overly stiff representation of the ssDNA backbone and the isotropic nature of base pairing interactions in 3SPN.1. By incorporating a more permissive backbone dihedral potential and angle-dependent interactions,[32] 3SPN.2 can capture the correct stiffness of both ssDNA and dsDNA. With the correct flexibility of ssDNA and the more realistic representations of base pairing interactions, the predicted mechanisms changed from slithering in 3SPN.1 to zippering in 3SPN.2.

This qualitative change in the hybridization mechanism highlights a fundamental challenge of developing CG models, of any type, for study of dynamic processes. The question of how much detail to incorporate into a CG model for description of dynamic, non-equilibrium phenomena is still open. This concern could be partly addressed by repeating the analysis presented here with other models, including more detailed representations (for example, the models described in Refs. 24 and 27). Consistent results for more elaborate DNA models would demonstrate that 3SPN.2 captures all of the relevant physics. In the case of DNA, such models have often not been rigorously parameterized to capture the effect of sequence and length on persistence length and melting temperature; thus, additional validation would be necessary before applying them to the problem at hand.

It is also encouraging to note that the results presented in our paper are consistent with those of oxDNA,[25] which is another model that has recently been used to examine hybridization of short sequences.[45] More specifically, oxDNA consists of 3 collinear interaction sites. Simulations of 14 bp oligomers with that model suggest three distinct pathways of hybridization, namely, zippering in heterogeneous sequences, and pseudoknot and inchworm internal displacements in repetitive sequences. These mechanisms are consistent with those reported here. Indeed, reaction probabilities reported here for a 14 bp sequence (see Fig. 3(b)) are in good agreement with those of Ref. 45. That study did not examine the length-dependence of the rate constants or the effect of ionic strength, and it is therefore difficult to make additional comparisons to the analysis presented here.

A key difference between 3SPN.2 and oxDNA is the nature of the sugar-phosphate backbone. 3SPN.2 relies on a structured backbone constrained by angle and dihedral potentials in order to recover the correct persistence lengths of ssDNA and dsDNA, while oxDNA uses an unstructured backbone; this is possible due to the rigid nature and native anisotropy of oxDNA's coarse-grained nucleotides. While the lack of structure facilitates the use of oxDNA in DNA nanotechnology applications,[60] it does represent an approximation and a departure from "true," atomistic DNA. For biological applications, in particular, we anticipate that backbone structure is likely to play an important role. Indeed, our recent studies of DNA-histone interactions indicate that such structure is important for accurate description of the resulting complexes.[61] Another CG model actually uses five beads, instead of the two of 3SPN.2, to represent the backbone.[62] Such recent developments illustrate that additional work may be needed to identify the ideal coarse-graining of the phosphate-sugar backbone in order to preserve all of the essential physics.

An additional difference between 3SPN.2 and oxDNA is the presence of electrostatic interactions. oxDNA does not include explicit electrostatics and was parameterized for a single ionic strength (500 mM), where it was assumed that electrostatic repulsions could be modeled with a repulsive Lennard-Jones potential. In contrast, 3SPN.2 assumes a uniform electrostatic continuum by incorporating electrostatics via DH theory. That level of description predicts a short-ranged electrostatic repulsion at 1 M ionic strength. As shown in Fig. 8, such repulsion does not have a large effect on the calculated rate constants. At low ionic strength the electrostatic repulsion leads to a significant reduction in the rate constant, though the mechanisms for hybridization appear to be similar. Note, however, that the discrepancy in Fig. 3(c) between simulated and experimental scaling of hybridization rate constants with ionic strength suggests that important physics are being omitted by modeling the ions implicitly. Future work will study DNA hybridization in the presence of

explicit ions, in an attempt to identify the role of ions in DNA hybridization.

V. CONCLUSION

In this work, we have used the 3SPN.2 coarse-grained DNA model to examine the effect of sequence, length, and ionic strength on DNA oligomer hybridization rate constants. We observe that sequence has a significant effect on the observed mechanism of hybridization. The formation of non-native base pairs leads to cooperative effects that increase the magnitude of rate constants, relative to sequences where only native base pairs are permitted. Consistent with experiment, we find that ionic strength has a strong effect on the overall rate constant. We also find that the mechanism of hybridization does not change qualitatively with ionic strength.

We observe scaling behavior consistent with that observed for longer lengths of ssDNA when assuming that oligomer hybridization is a reaction-controlled process. Previous authors have formulated scaling theories as a function of molecular weight based on excluded volume arguments. We have extended their theories to account for the effect of sequence and nucleation events occurring between strands both in- and out-of-register. This new theory can predict the relative magnitude of rate constants from simulation and may find use in nanotechnology predicting the relative rates of hybridization of DNA oligomers with different sequence lengths and identities.

ACKNOWLEDGMENTS

D.M.H. thanks Jonathan K. Whitmer and Gordon S. Freeman for helpful discussions. The University of Wisconsin–Madison Center for High Throughput Computing is gratefully acknowledged for providing resources and computer expertise. D.M.H. was funded by a Graduate Research Fellowship from the National Science Foundation (NSF) (Grant No. DGE-1256259). The development of the DNA model employed here was supported by the National Science Foundation through the UW-NSEC (DMR0832760). The development of the biased forward flux technique and the corresponding rate constant calculations was supported by the U.S. Department of Energy, Office of Science, Materials Sciences and Engineering Division.

[1]N. C. Seeman, Mol. Biotechnol. **37**, 246 (2007).
[2]S. J. Tan, M. J. Campolongo, D. Luo, and W. Cheng, Nat. Nanotechnol. **6**, 268 (2011).
[3]R. J. Macfarlane, B. Lee, M. R. Jones, N. Harris, G. C. Schatz, and C. A. Mirkin, Science **334**, 204 (2011).
[4]R. J. Macfarlane, M. R. Jones, B. Lee, E. Auyeung, and C. A. Mirkin, Science **341**, 1222 (2013).
[5]S. Y. Park, A. K. R. Lytton-Jean, B. Lee, S. Weigand, G. C. Schatz, and C. A. Mirkin, Nature (London) **451**, 553 (2008).
[6]P. Rothemund, Nature (London) **440**, 297 (2006).
[7]H. Gu, J. Chao, S. Xiao, and N. Seeman, Nature (London) **465**, 202 (2010).
[8]J. Marmur and P. Doty, J. Mol. Biol. **3**, 585 (1961).
[9]J. G. Wetmur and N. Davidson, J. Mol. Biol. **31**, 349 (1968).
[10]M. E. Craig, D. M. Crothers, and P. Doty, J. Mol. Biol. **62**, 383 (1971).
[11]D. C. Rau and L. C. Klotz, J. Chem. Phys. **62**, 2354 (1975).
[12]L. E. Morrison and L. M. Stols, Biochemistry **32**, 3095 (1993).
[13]C. Chen, W. Wang, Z. Wang, F. Wei, and X. S. Zhao, Nucl. Acids Res. **35**, 2875 (2007).
[14]J.-L. Sikorav, H. Orland, and A. Braslau, J. Phys. Chem. B **113**, 3715 (2009).
[15]D. Pörschke, Biophys. Chem. **2**, 83 (1974).
[16]D. Pörschke and M. Eigen, J. Mol. Biol. **62**, 361 (1971).
[17]D. Pörschke, O. C. Uhlenbeck, and F. H. Martin, Biopolymers **12**, 1313 (1973).
[18]M. F. Hagan, A. R. Dinner, D. Chandler, and A. K. Chakraborty, Proc. Natl. Acad. Sci. U.S.A. **100**, 13922 (2003).
[19]S. Piana, J. Phys. Chem. A **111**, 12349 (2007).
[20]D. Bashford and D. A. Case, Annu. Rev. Phys. Chem. **51**, 129 (2000).
[21]T. A. Knotts IV, N. Rathore, D. C. Schwartz, and J. J. de Pablo, J. Chem. Phys. **126**, 084901 (2007).
[22]E. J. Sambriski, D. C. Schwartz, and J. J. de Pablo, Biophys. J. **96**, 1675 (2009).
[23]A. Morriss-Andrews, J. Rottler, and S. S. Plotkin, J. Chem. Phys. **132**, 035105 (2010).
[24]P. D. Dans, A. Zeida, M. R. Machado, and S. Pantano, J. Chem. Theory Comput. **6**, 1711 (2010).
[25]T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, J. Chem. Phys. **134**, 085101 (2011).
[26]M. C. Linak, R. Tourdot, and K. D. Dorfman, J. Phys. Chem. **135**, 205102 (2011).
[27]A. V. Savin, M. A. Mazo, I. P. Kikot, L. I. Manevitch, and A. V. Onufriev, Phys. Rev. B **83**, 245406 (2011).
[28]C. W. Hsu, M. Fyta, G. Lakatos, S. Melchionna, and E. Kaxiras, J. Chem. Phys. **137**, 105102 (2012).
[29]L. E. Edens, J. A. Brozik, and D. J. Keller, J. Phys. Chem. B **116**, 14735 (2012).
[30]O. Gonzalez, D. Petkevičiūtė, and J. H. Maddocks, J. Chem. Phys. **138**, 055102 (2013).
[31]Y. He, M. Maciejczyk, S. Ołdziej, H. A. Scheraga, and A. Liwo, Phys. Rev. Lett. **110**, 098101 (2013).
[32]D. M. Hinckley, G. S. Freeman, J. K. Whitmer, and J. J. de Pablo, J. Chem. Phys. **139**, 144903 (2013).
[33]S. Plimpton, J. Comp. Phys. **117**, 1–19 (1995).
[34]See https://uchic.ag/3spn2 for the USER-3SPN2 package needed to run 3SPN.2 simulations in LAMMPS.
[35]F. Oosawa and M. Kasai, J. Mol. Biol. **4**, 10 (1962).
[36]G. S. Manning, J. Chem. Phys. **51**, 924 (1969).
[37]G. Bussi and M. Parrinello, Phys. Rev. E **75**, 056707 (2007).
[38]A. E. Nkodo, J. M. Garnier, B. Tinland, H. Ren, C. Desruisseaux, L. C. McCormick, G. Drouin, and G. W. Slater, Electrophoresis **22**, 2424 (2001).
[39]We chose to normalize the rate constants near physiological conditions ($I = 200$ mM) because the model is more likely to be used under such conditions.
[40]T. J. Schmitt and T. A. Knotts IV, J. Chem. Phys. **134**, 205105 (2011).
[41]T. J. Schmitt, J. B. Rogers, and T. A. Knotts IV, J. Chem. Phys. **138**, 035102 (2013).
[42]E. J. Sambriski, D. C. Schwartz, and J. J. de Pablo, Proc. Natl. Acad. Sci. U.S.A. **106**, 18125 (2009).
[43]E. J. Sambriski, V. Ortiz, and J. J. de Pablo, J. Phys.: Condens. Matter **21**, 034105 (2009).
[44]M. J. Hoefert, E. J. Sambriski, and J. J. de Pablo, Soft Matter **7**, 560 (2011).
[45]T. E. Ouldridge, P. Šulc, F. Romano, J. P. K. Doye, and A. A. Louis, Nucl. Acids Res. **41**, 8886 (2013).
[46]R. J. Allen, C. Valeriani, and P. R. ten Wolde, J. Phys.: Condens. Matter **21**, 463102 (2009).
[47]F. A. Escobedo, E. E. Borrero, and J. C. Araque, J. Phys.: Condens. Matter **21**, 333101 (2009).
[48]S. H. Northrup, S. A. Allison, and J. A. McCammon, J. Chem. Phys. **80**, 1517 (1984).
[49]R. J. Britten, D. E. Graham, and B. R. Neufeld, Methods Enzymol. **29**, 363 (1974).
[50]S. Redner, *A Guide to First-passage Processes* (Cambridge University Press, 2001).
[51]See supplementary material at http://dx.doi.org/10.1063/1.4886336 for additional details regarding hybridization theory, predicted mechanisms, and scaling of rate constants.
[52]J. N. Onuchic and P. G. Wolynes, Curr. Opin. Struct. Biol. **14**, 70 (2004).
[53]J. H. van de Sande, N. B. Ramsing, M. W. Germann, W. Elhorst, B. W. Kalisch, E. von Kitzing, R. T. Pon, R. C. Clegg, and T. M. Jovin, Science **241**, 551 (1988).

[54]J. G. Wetmur, Annu. Rev. Biophys. Biol. **5**, 337 (1976).
[55]J. Des Cloizeaux, J. Phys. **41**, 223 (1980).
[56]A. R. Khokhlov, Macromol. Rapid. Commun. **2**, 633 (1981).
[57]A. G. Hinnebusch, V. E. Clark, and L. C. Klotz, Biochemistry **17**, 1521 (1978).
[58]R. Owczarzy, A. V. Tataurov, Y. Wu, J. A. Manthey, K. A. McQuisten, H. G. Almabrazi, K. F. Pedersen, Y. Lin, J. Garretson, N. O. McEntaggard, C. A. Sailor, R. B. Dawson, and A. S. Peek, Nucl. Acids Res. **36**, W163 (2008).
[59]F. W. Studier, J. Mol. Biol. **41**, 199 (1969).
[60]J. Doye, T. Ouldridge, A. A. Louis, F. Romano, P. Šulc, C. Matek, B. Snodin, L. Rovigatti, J. Schreck, R. Harrison, and W. P. J. Smith, Phys. Chem. Chem. Phys. **15**, 20395 (2013).
[61]G. S. Freeman, D. M. Hinckley, J. P. Lequieu, J. K. Whitmer, and J. J. de Pablo, "DNA shape dominates sequence affinity in nucleosome formation," Phys. Rev. Lett. (submitted).
[62]T. Cragnolini, P. Derreumaux, and S. Pasquali, J. Phys. Chem. B **117**, 8047 (2013).