


## Lattice model on the rate of DNA hybridization

R. Murugan \**Department of Biotechnology, Indian Institute of Technology Madras, Chennai 600036, Tamil Nadu, India*

(Received 31 January 2022; accepted 2 June 2022; published 21 June 2022)

We develop a lattice model on the rate of hybridization of the complementary single-stranded DNAs (c-ssDNAs). Upon translational diffusion mediated collisions, c-ssDNAs interpenetrate each other to form correct (cc), incorrect (icc), and trap correct contacts (tcc) inside the reaction volume. Correct contacts are those with exact registry matches, which leads to nucleation and zipping. Incorrect contacts are the mismatch contacts which are less stable compared to tcc, which can occur in the repetitive c-ssDNAs. Although tcc possess registry match within the repeating sequences, they are incorrect contacts in the view of the whole c-ssDNAs. The nucleation rate ( $k_N$ ) is directly proportional to the collision rate and the average number of correct contacts ( $\langle n_{cc} \rangle$ ) formed when both c-ssDNAs interpenetrate each other. Detailed lattice model simulations suggest that  $\langle n_{cc} \rangle \propto L/V$  where  $L$  is the length of c-ssDNAs and  $V$  is the reaction volume. Further numerical analysis revealed the scaling for the average radius of gyration of c-ssDNAs ( $R_g$ ) with their length as  $R_g \propto \sqrt{L}$ . Since the reaction space will be approximately a sphere with radius equals to  $2R_g$  and  $V \propto L^{3/2}$ , one obtains  $k_N \propto \frac{1}{\sqrt{L}}$ . When c-ssDNAs are nonrepetitive, the overall renaturation rate becomes as  $k_R \propto k_N L$ , and one finally obtains  $k_R \propto \sqrt{L}$  in line with the experimental observations. When c-ssDNAs are repetitive with a complexity of  $c$ , earlier models suggested the scaling  $k_R \propto \frac{\sqrt{L}}{c}$ , which breaks down at  $c = L$ . This clearly suggests the existence of at least two different pathways of renaturation in the case of repetitive c-ssDNAs, *viz.*, via incorrect contacts and trap correct contacts. The trap correct contacts can lead to the formation of partial duplexes which can keep the complementary strands in the close proximity for a prolonged timescale. This is essential for the extended 1D slithering, inchworm movements, and internal displacement mechanisms which can accelerate the searching for the correct contacts. Clearly, the extent of slithering dynamics will be inversely proportional to the complexity. When the complexity is close to the length of c-ssDNAs, the pathway via incorrect contacts will dominate. When the complexity is much less than the length of c-ssDNA, pathway via trap correct contacts would be the dominating one.

DOI: [10.1103/PhysRevE.105.064410](https://doi.org/10.1103/PhysRevE.105.064410)

### I. INTRODUCTION

The reversible unwind-rewind property of the double-stranded helical structure of DNA is critical for its various biological functions [1,2]. The double helical structure of DNA (dsDNA) is stabilized by the hydrogen bonding network and hydrophobic base stacking at the core [1]. Denaturation is the process of melting of dsDNA into the corresponding complementary single strands [2]. These single strands of DNA (ssDNA) zip back spontaneously to form the original dsDNA upon removal of the denaturant, which is known as renaturation or hybridization [2–4]. Transcription, translation, and replication of the genomic DNA and several *in vitro* laboratory techniques are based on the denaturation-renaturation property of dsDNA [5]. Clear understanding on the mechanism of DNA hybridization in solution is important to design efficient primers for polymerase chain reaction (PCR), design of oligonucleotide probes for microarray chips, design and construction of versatile nanostructures over single-stranded DNA scaffolds using the DNA origami method [6,7] and various DNA fingerprinting technologies [5]. Detailed understanding of the mechanism of hybridization of ssDNAs at the

microscopic level is a contemporary issue in the biological physics field.

Several models of DNA renaturation have been developed and experimentally tested [2,8–17]. There are two different views, *viz.*, one- and two-step models. Hybridization of the complementary ssDNAs (c-ssDNAs) was initially described as a one-step diffusion controlled bimolecular collision process as in Scheme I of Fig. 1(a) [4,18–20]. Although this scheme was simple enough to capture most of the underlying dynamics [8,14,17,19], it could not explain several other polymer physics-related scaling relationships. Particularly, Scheme I predicted a linear scaling of the bimolecular collision rate with respect to the length of c-ssDNAs, whereas the experimental data revealed approximately a square-root-type scaling [8,14]. Further experiments revealed an inverse scaling of the bimolecular collision rate with the sequence complexity of ssDNA.

Wetmur and Davidson [8] suggested a detailed two-step model with nucleation and zipping as in Scheme II of Fig. 1(a). In their model, the overall renaturation rate was directly proportional to the product of nucleation rate and length of c-ssDNAs and inversely proportional to the sequence complexity. The nucleation rate scales with the length of c-ssDNA in an inverse square root manner. As a result, the overall renaturation rate scales with the length of ssDNA in a square root manner. They further argued that the renatura-

\*rmurugan@gmail.com

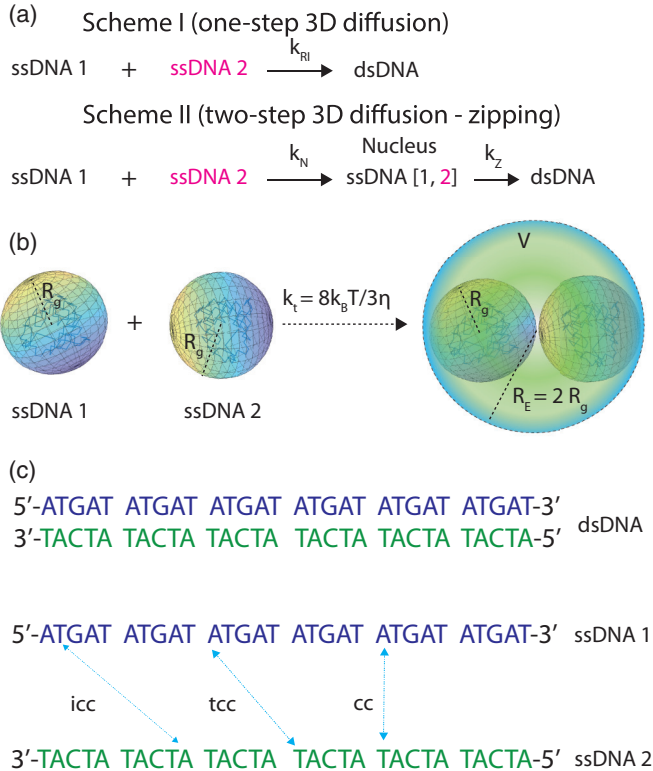


FIG. 1. Models of DNA renaturation. (a) One- and two-step models. In the one-step model described in Scheme I, the complementary strands form duplex with a second-order rate  $k_{RI}$  directly by diffusion-controlled 3D collisions. In the two-step model described in Scheme II, the renaturation progress occurs by the formation of stable nucleus with a rate  $k_N$  via 3D diffusion-mediated collisions and interpenetrations of c-ssDNAs and then subsequently by zipping with a rate  $k_Z$ . (b) The complementary ssDNA strands can be thought as spherical-shaped and loosely packed nucleotide clusters with average radius of gyration  $R_g$ . Upon translational diffusion, they arrive inside the reaction volume  $V$  with a rate  $k_t$  and interpenetrate each other. The reaction space is assumed to be a spherical one with radius  $2R_g$ . When the complementary strands are confined inside  $V$ , there may be contacts between them. (c) When the contact occurs with exact registry match, it is a correct contact that can lead to nucleation and zipping. When contacts occur between nonidentical registers, they are all incorrect contacts. When c-ssDNAs contain repeats, there may be contacts between identical registries of repeats placed at two nonidentical locations. These are trap correct contacts which can lead to the formation of partial duplexes with single-strand overhangs. In the given example, there are six repeats. When position 1 of the first repeat interacts with position 1 of the second repeat, it is a tcc. Sequence complexity is defined as the number of bases in a unique sequence. For example, the given sequence contains the repeats of five bases. Therefore, its complexity is  $c = 5$  where the total length is  $L = 30$ .

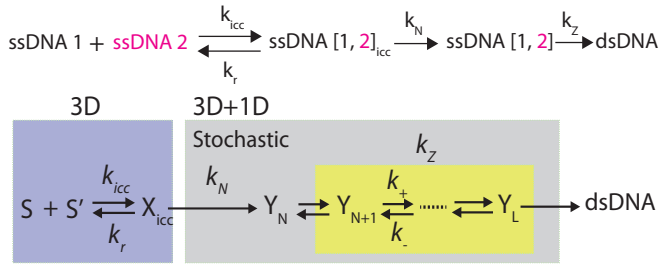
tion was not a diffusion-controlled process, and the inverse square root scaling of the nucleation rate with the length of c-ssDNA must be due to either the excluded volume effects associated with the intrastrand dynamics or steric hindrance associated with the interpenetration of c-ssDNAs that is essential for the nucleation [8]. However, this argument did not explain the inverse scaling of the renaturation rate with the

viscosity of the medium. To overcome this issue, they further assumed a diffusion-mediated growth of the nucleus with a rate that is inversely related with the viscosity coefficient of the medium [21]. This assumption ensured the inverse scaling of the overall renaturation rate with the viscosity of the medium.

The c-ssDNAs are random coils with average radius of gyration  $R_g$  [Fig. 1(b)]. At coarse-grained level, one can think both the complementary strands as loosely packed spherical clusters of nucleotide monomers with radius  $R_g$  performing translational diffusion. The base clusters of c-ssDNAs interpenetrate each other upon collision and subsequently nucleation occurs inside the reaction volume. Clearly, there are two distinct dynamical regimes in the process of renaturation, *viz.*, (1) translational diffusion mediated collision of base clusters of c-ssDNAs and (2) their interpenetration. Translational diffusion brings both strands within the reaction radius. Reaction radius will be the sum of the radius of gyration of c-ssDNAs. Subsequently, these c-ssDNA base clusters interpenetrate each other within the reaction volume to achieve the nucleation and zipping. Therefore, the translational diffusion component cannot be ignored. Further, the expression for the reaction rate corresponding to the two-step model [Eq. (3) in Ref. [8]] will be inconsistent whenever the sequence complexity has the same magnitude as that of the length of c-ssDNAs. Under such a scenario, the overall renaturation rate would inversely scale with the length of c-ssDNAs in a square root manner that is inconsistent with the experimental observations [8]. The *sequence complexity* is defined as the length of DNA with unique nucleotide sequence pattern [Fig. 1(c)]. Several theoretical and computational models [13,22–28] were developed recently to explain the observed scaling behaviors of the overall second-order renaturation rate on the size of c-ssDNAs, temperature, ionic strength, and viscosity of the reaction medium.

The nucleation step can be modeled as Kramer's escape problem over a free energy barrier [24]. However, the nature of the reaction coordinate and the entropic component of potential energy barrier associated with the renaturation process is unclear. According to the recently proposed three-step model [15] (Scheme III in Fig. 2), the renaturation process comprises (1) formation of nonspecific contact, (2) nucleation or correct contact formation, and (3) zipping. In the first step, c-ssDNAs perform three-dimensional collisions which result in the formation of Watson-Crick (WC) base pairs at random nonspecific contacts. Such nonspecific contacts randomly translocate along c-ssDNAs via either thermally driven one-dimensional (1D) slithering, inchworm movements, or internal displacement [25] mechanisms until finding the correct contact to initiate the nucleation process. Nucleation will be followed by spontaneous zippering of c-ssDNAs. In this model, the square root dependency of the renaturation rate on the length of c-ssDNAs mainly originates from the nonspecific contact formation step. However, it is still not clear whether the mode of nucleation is via 1D or 3D diffusion. It is also not clear how the inverse scaling on the sequence complexity arises in the case of repetitive c-ssDNAs. Further, the retardation effects of the repetitive DNA sequences were not considered in Scheme III of Ref. [15].

Scheme III (three-step 3D-1D diffusion, nonrepetitive DNA)



Scheme IV (three-step 3D-1D diffusion, repetitive DNA)

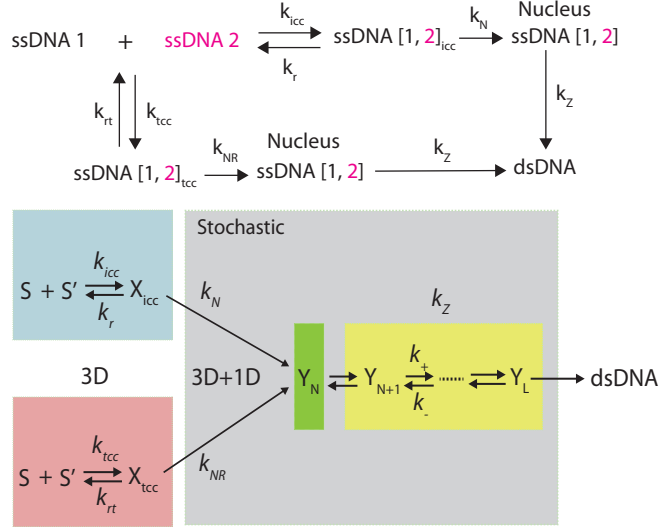


FIG. 2. Three-step models on DNA renaturation.  $S$  and  $S'$  represent ssDNA1 and ssDNA2, respectively,  $X_{icc} = \text{ssDNA [1, 2]}_{icc}$ ,  $X_{tcc} = \text{ssDNA [1, 2]}_{tcc}$ ,  $Y_N$  represents the nucleus with  $N$  nucleotides. In the three-step model given by Scheme III, c-ssDNAs renature via a combination of 3D and 1D diffusions. Here  $k_{icc}$  and  $k_{tcc}$  are the number of incorrect and trap correct contacts formed per second,  $k_r$  and  $k_{rt}$  are respective dissociation rate constants,  $k_N$  and  $k_{NR}$  are the respective nucleation rates, and  $k_+$  and  $k_-$  are the forward and reverse microscopic rate constants associated with the zipping steps. In Scheme III, 3D collisions result in the formation of incorrect contacts which bring the strands into the close proximity. Nucleation occurs via various 1D diffusion dynamics such as slithering, inchworm movements, and internal displacement mechanisms. When c-ssDNAs are repetitive, there are always chances for the formation of partial duplexes with single-strand overhangs. This in turn will retain the c-ssDNAs in the close proximity for prolonged amount of time that results in  $k_r > k_{rt}$  and long slithering lengths. This will be a parallel pathway along with Scheme III as demonstrated in Scheme IV in the case of repetitive c-ssDNAs.

Understanding the role of the conformational state of DNA on the rate of hybridization is critical to unravel the underlying mechanism. To understand the equilibrium thermodynamic properties of the DNA hybridization, several course-grained lattice models were developed and investigated using Monte Carlo simulation methods [28]. However, simulation of such systems with detailed base-pairing and base-stacking interactions were limited to very short DNA sequences [25,28]. Recently, Qu *et al.* [29] have simulated up to around 85 nucleotides using the oxDNA GPU-enabled standalone code.

Their simulated time-dependent system energy profile exhibited a four-state hybridization mechanism in line with earlier theoretical models [15]. Using similar codes, Snodin *et al.* [30] simulated the hybridization kinetics of DNA origami for a system of 384 nucleotides. Such studies are useful to obtain the equilibrium thermodynamic properties rather than the nonequilibrium kinetic scaling aspects. In this context, short strands have been studied using nonequilibrium methods for enhanced forward-flux sampling at much reduced computational requirements. Particularly, Schreck *et al.* [31] have studied 25-nucleotide oxDNA, Gravina *et al.* [32] have studied 42-nucleotide 3SPN.2 DNA, and Jones *et al.* [33] have studied a DNA with 10 nucleotides AT repeats. Apart from these, Sidky *et al.* [34] have created a Software Suite for Advanced General Ensemble Simulation (SSAGES) which can use the parallel computational workflows for forward-flux sampling. Clearly, all these computational methods assume that the interacting c-ssDNAs already arrived at the reaction volume through translational diffusion, whereas the kinetic scaling laws with respect to varying lengths of DNA, sequence complexity, and viscosity of the reaction medium associated with the hybridization are mainly dependent on the translational diffusion of c-ssDNAs along with their interpenetration dynamics rather than the fine details of base-pairing and base-stacking interactions, which are the main focus of all the coarse-grained Monte Carlo simulation methods. Further, though such computational studies revealed several interesting qualitative aspects of hybridization [29,33], derivation of scaling laws through a computational route requires the simulation of very large lengths of ssDNAs with different levels of sequence complexities that still limits the application of molecular dynamics methods. In this background, we model c-ssDNAs as self-avoiding random walks (SARWs) confined in a cubic lattice box which represents the reaction volume under *in vitro* conditions. Using this lattice model along with the observations from the computational studies [29,33], we will unravel the origin of various kinetic scaling relationships associated with the renaturation dynamics in detail.

## II. THEORETICAL METHODS

### A. Translational diffusion of c-ssDNAs

We model c-ssDNAs [we denote them as ssDNA 1 and 2 as in Fig. 1(a)] as loosely packed and approximately spherical-shaped nucleotide clusters with a radius of gyration  $R_g$ . The 3D collisions between these c-ssDNAs are mediated via their translational diffusion. When the length of both these strands are equal to  $L$ , one can approximately assume equal radius of gyration  $R_g$  for both strands. In this situation, the maximum possible steady-state Smolochowski rate of 3D diffusion-controlled collisions between these c-ssDNAs will be  $k_t = \frac{8k_B T}{3\eta} \phi$  [35] [Fig. 1(b)]. Here  $k_B$  is the Boltzmann constant,  $\eta$  is the viscosity coefficient of the medium,  $T$  is the absolute temperature in degrees  $K$ , and  $\phi \cong \left| \frac{(\kappa/2R_g)}{\exp(\kappa/2R_g) - 1} \right| \exp\left(\frac{\kappa}{\lambda}\right)$  [15,36] is the multiplication factor associated with the electrostatic repulsions between the negatively charged phosphate backbones of c-ssDNAs along with the shielding effects of solvent molecules and other ions present at the DNA-DNA

interface under dilute conditions. Here  $\kappa = \frac{\xi_1 \xi_2 e^2}{\Omega k_B T}$  is the Onsager radius, which is defined as the distance between the colliding c-ssDNA strands at which the overall electrostatic interaction energy will be equal to that of the background thermal energy ( $k_B T$ ),  $\xi_1, \xi_2$  are the overall charge numbers on the base clusters of strand 1 and 2, and  $\lambda$  is the thickness of the ionic shell present over the charged groups of c-ssDNAs. Here  $\kappa$  will be a positive quantity since there is a charge-charge repulsion between c-ssDNAs. Since the electrostatic interactions act only at short ranges, in general one finds that  $|\kappa| \ll 2R_g$  where the reaction radius is  $2R_g$ . As a result,  $|\frac{(\kappa/2R_g)}{\exp(\kappa/2R_g)-1}| \cong 1$  and therefore the multiplication factor  $\phi \cong \exp(\frac{\kappa}{\lambda})$  will be independent of the radius of gyration of c-ssDNAs. One also should note that  $\kappa < 0$  in the case of site-specific DNA-protein interactions since the DNA binding domains of the DNA binding proteins are usually rich in positively charged amino acids. Under such conditions,  $\phi$  will be dependent on  $R_g$  [37].

In the calculation of  $k_t$ , we have assumed an absorbing boundary condition for the associated diffusion equation at the reaction radius that is approximately equals  $2R_g$ . By definition, this is the rate of arrival of the base clusters of c-ssDNAs within the reaction radius. Here  $k_t$  does not enumerate the number of contacts formed upon such collisions and subsequent interpenetrations. We will show in the later sections that there is always a nonzero probability for the occurrence of zero contacts upon collision and interpenetration of c-ssDNAs. Clearly, the translational diffusion brings c-ssDNAs inside the reaction volume, and there is no physical confinement of the polymer within a closed space here. The reacting c-ssDNAs enter and exit the reaction volume freely by diffusion. We will show in the later sections that this assumption is essential since the radius of gyration of c-ssDNAs will be strongly dependent on the confinement volume.

## B. Interpenetration dynamics of c-ssDNAs

With this background, the nucleation rate ( $k_N$ ) will be directly proportional to the number of correct contacts formed between the c-ssDNAs per second. On the other hand, the average number of correct contacts formed per second ( $k_{cc}$ ) will be equal to the number of collisions between the base clusters of c-ssDNAs per second ( $k_t$ , measured in  $M^{-1}s^{-1}$ ) multiplied by the average number of correct contacts ( $\langle n_{cc} \rangle$ ) formed after each collision and interpenetration inside the reaction volume ( $V$ ), i.e.,  $k_{cc} = k_t \langle n_{cc} \rangle$ . As a result, one obtains  $k_N \propto k_t \langle n_{cc} \rangle$ . Only the correct contact is the perfect registry match between the c-ssDNAs that can lead to nucleation and zipping. Apart from the correct contacts, several incorrect contacts (icc) can also be formed between the c-ssDNAs upon collision. When the c-ssDNAs are repetitive, there are possibilities for the trap correct contacts (tcc) as described in Fig. 1(c). Trap correct contacts possess registry matches within the repeats, but they are all incorrect contacts in the view of the entire c-ssDNAs. Here tcc can lead to partial duplexes with single-strand overhangs which are actually kinetic traps in the pathway of DNA renaturation. We denote the number of cc, icc, and tcc as  $n_{cc}$ ,  $n_{icc}$ , and  $n_{tcc}$ , respectively. Clearly,  $n_{cc}$ ,

$n_{icc}$ , and  $n_{tcc}$  are all random variables since they vary from collision to collision and interpenetrations. We denote the corresponding ensemble averages as  $\langle n_{cc} \rangle$ ,  $\langle n_{icc} \rangle$ , and  $\langle n_{tcc} \rangle$ , respectively. With these definitions, one can derive the following expressions:

$$k_{cc} \cong k_t \langle n_{cc} \rangle, \quad k_{icc} \cong k_t \langle n_{icc} \rangle, \quad k_{tcc} \cong k_t \langle n_{tcc} \rangle. \quad (1)$$

In these equations,  $k_{cc}$ ,  $k_{icc}$ , and  $k_{tcc}$  are respectively the number of correct, incorrect, and trap correct contacts formed between c-ssDNAs strands per second. Clearly, the nucleation rate  $k_N \propto k_{cc}$ . The correct contacts can form anywhere on the entire stretch of c-ssDNAs with length equal to  $L$  number of nucleotides (nt). Therefore, the probability of getting a correct contact upon each collision will be  $p_{cc} = 1/L$ . This also means that the average number of correct and incorrect contacts is approximately connected via  $\langle n_{cc} \rangle \cong p_{cc} \langle n_{icc} \rangle$ . To understand various scaling behaviors of  $\langle n_{cc} \rangle$ ,  $\langle n_{icc} \rangle$ , and  $\langle n_{tcc} \rangle$ , we model c-ssDNAs as self-avoiding random walks (SARWs) confined in a lattice cube box that represents the reaction volume  $V$  of the standard bimolecular collision model. In this setting, the volume of a monomeric unit is  $v = 1$ , and the volume occupied by the c-ssDNA strand of length  $L$  will be  $vL$ . When the complexity of ssDNA is  $c$  nt, there will be at least  $L/c$  number of repeating sequences in each c-ssDNA strand. We assume that the respective c-ssDNA base clusters have already arrived at the reaction volume through the 3D translational diffusion with rate  $k_t$ . The average number of cc, icc, and tcc can be obtained by repeated generation of two independent SARWs of length  $L$  inside a fixed cubic lattice box with volume  $V$ . In each iteration, the number of cc, icc, and tcc will be counted, and these counts are averaged over  $10^5$  SARW trajectories. Detailed stochastic simulations with fixed  $L$  for both the c-ssDNA strands and reaction volume  $V$  as given in the Simulation Methods section revealed the following scaling relationships:

$$\langle n_{cc} \rangle \cong \frac{vL}{V}, \quad \langle n_{icc} \rangle \cong \frac{vL^2}{V}, \quad \langle n_{tcc} \rangle \cong \frac{vL^2}{cV}. \quad (2)$$

Here  $v/V$  is the probability of finding any nucleotide of c-ssDNAs inside  $V$  upon each movement and  $vL$  is the total chain volume. Upon assuming c-ssDNAs as a spherical-shaped nucleotide clusters, the reaction space can be thought approximately as a sphere with radius of  $2R_g$  where  $R_g$  is the average radius of gyration of the individual c-ssDNA. One can straightforwardly interpret these results as follows. Since  $L$  number of monomers of the c-ssDNA of interest search for the complementary nucleotides on the other strand at the same time, one obtains  $\langle n_{icc} \rangle \cong \frac{vL^2}{V}$ . Out of these  $n_{icc}$  numbers of incorrect contacts, the probability of finding the trap correct contacts among the repetitive c-ssDNAs will be  $1/c$ . Therefore, one obtains  $\langle n_{tcc} \rangle = \langle n_{icc} \rangle / c$ . In the same way, the probability of finding the correct contact will be  $1/L$  that is independent of the number of repeats. From this one can deduce that  $\langle n_{cc} \rangle = \langle n_{icc} \rangle / L$ . This expression for  $\langle n_{tcc} \rangle$  will work only when  $L$  is much higher than the sequence complexity  $c$  so that  $(\frac{L}{c}) \gg 1$ . When  $c \rightarrow L$ , there are always chances for the occurrence of partial repeats, and one can obtain the approximation  $\lim_{c \rightarrow L} \langle n_{tcc} \rangle \cong \frac{v(\bar{L}^2 + l^2)}{cV}$  where  $\bar{L} = c \lfloor \frac{L}{c} \rfloor$  is the length of c-ssDNAs containing full repeats and  $l = L - c \lfloor \frac{L}{c} \rfloor$

is the length of c-ssDNAs with partial repeats. For example, when  $L = 20$  and  $c = 6$ , one obtains  $\bar{L} = 18$  and  $l = 2$ . Here  $\lfloor \frac{L}{c} \rfloor$  is the floor function operator. For example,  $\lfloor \frac{20}{6} \rfloor = 3$ . When the main template strand length is fixed at  $L$  and only the length of the probe  $u$  is varied as in the case of PCR reactions, we find the following scaling relationships:

$$\langle n_{cc} \rangle \cong \frac{vu}{V}, \quad \langle n_{icc} \rangle \cong \frac{Lvu}{V}, \quad \langle n_{tcc} \rangle \cong \frac{Lvu}{cV}. \quad (3)$$

When the length of both c-ssDNAs is randomized within  $(0, L)$  with equal probabilities ( $= 1/L$ ) emulating the uniformly sheared DNA [8] so that the average length will be  $\langle L \rangle = L/2$ , one finds the following scaling relationships:

$$\begin{aligned} \langle n_{cc} \rangle &\cong \frac{v\langle L \rangle}{2V} = \frac{vL}{4V}, & \langle n_{icc} \rangle &\cong \frac{v\langle L \rangle^2}{V} = \frac{vL^2}{4V}, \\ \langle n_{tcc} \rangle &\cong \frac{v\langle L \rangle^2}{cV} = \frac{vL^2}{4cV}. \end{aligned} \quad (4)$$

When the complexity  $c = L$ , then irrespective of the sheared nature of c-ssDNAs, the probability of finding the correct contact will be  $p_{cc} = 1/L$ . As a result, one obtains  $\langle n_{cc} \rangle \cong p_{cc} \langle n_{icc} \rangle = Lv/4V$  for the case of sheared DNA. When  $c \rightarrow \langle L \rangle$  one can obtain the approximation  $\lim_{c \rightarrow \langle L \rangle} \langle n_{tcc} \rangle \cong \frac{v(\langle L \rangle^2 + l^2)}{cV}$  where  $\langle \bar{L} \rangle = c \lfloor \frac{\langle L \rangle}{c} \rfloor$  and  $\langle l \rangle = \langle L \rangle - c \lfloor \frac{\langle L \rangle}{c} \rfloor$  by definition as in the case of nonsheared and repetitive c-ssDNAs.

### C. Radius of gyration of c-ssDNAs

The ensemble average of the radius of gyration of a chain molecule can be defined as  $R_g = \langle |\sqrt{\frac{\sum_{i,j=1}^L (\bar{w}_i - \bar{w}_j)^2}{2L^2}}| \rangle$  where  $\bar{w}_i$  is the vector coordinates ( $X, Y, Z$ ) of the  $i$ th monomer and  $L$  is the length [38,39]. The averaging is done over several possible spatial configurations of the polymer. Detailed numerical simulations suggested that the scaling of  $R_g$  with  $L$  deviates significantly from the standard scaling  $R_g \propto \sqrt{L}$  when the ratio  $(\frac{vL}{V_S}) \rightarrow 1$  where  $V_S$  is the volume in which the SARW is confined. In other words, the scaling  $R_g \propto \sqrt{L}$  will be valid only in the limit  $(\frac{vL}{V_S}) \rightarrow 0$ . In this context, nonlinear least-square fittings of the values of  $R_g$  obtained over various lengths of SARWs confined in different volumes  $V_S$  using the Marquart-Levenberg algorithm [40] suggested the following functional form:

$$R_g \cong \alpha \sqrt{L}(1 + \beta L)^{-1}, \quad \beta = \left( \varepsilon \left( \frac{v}{V_S} \right)^\delta + \sigma \right),$$

$$\lim_{V_S \rightarrow \infty} R_g \cong \alpha \sqrt{L}, \quad \alpha \cong \sqrt{l_p^2/6}. \quad (5)$$

Here  $\alpha$  is the preexponent and  $l_p$  is the physical distance between the monomers. Clearly, in the limit as  $V_S \rightarrow \infty$  that represents the dilute *in vitro* conditions, we recover the well-known scaling relationship for linear chain polymers as  $R_g \cong l_p \sqrt{L/6}$ . The excluded volume effects will be prominent when the intrinsic volume of the polymer  $vL$  is close to the confinement volume  $V_S$  [8]. Under such conditions, the interpenetration of c-ssDNAs among each other will be very much limited. In our model, c-ssDNAs are not physically confined in space. Rather, they are allowed to enter or exit the reaction volume freely. Hence, we assume  $R_g \cong l_p \sqrt{L/6}$  and

subsequently derive the following expressions for the reaction and monomer volumes:

$$V \cong \frac{4}{3} \pi (2R_g)^3 = \frac{32}{3} \pi l_p^3 \left( \frac{L}{6} \right)^{3/2}, \quad v \cong \frac{4}{3} \pi l_p^3 n_D^3. \quad (6)$$

In these equations,  $l_p n_D$  is the radius of gyration of the nucleotide monomer in terms of number ( $n_D$ ) of distances between adjacent nucleotides [ $l_p \cong 3.4 \times 10^{-10}$  m for dsDNA and  $l_p \cong (5.9 \text{ to } 7) \times 10^{-10}$  m for c-ssDNAs; since  $l_p$  corresponding to c-ssDNAs is a fluctuating quantity, we denote  $l_p$  corresponding to dsDNA as the standard base-pair unit, bp], and the volumes are measured in  $\text{bp}^3$ . When we model DNA as chain of beads where each bead represents a monomer, the radius of the monomer bead will be approximately equal to the radius of the DNA cylinder, which is approximately equals  $n_D l_p \approx 3$  bp. Using Eqs. (6), when  $L \gg c$ , one can derive the following expressions for the number of cc, icc, and tcc formed upon interpenetration of c-ssDNAs inside the reaction volume:

$$\begin{aligned} \langle n_{cc} \rangle &\cong \frac{3\sqrt{6}n_D^3}{4\sqrt{L}}, & \langle n_{icc} \rangle &\cong \frac{3\sqrt{6}n_D^3\sqrt{L}}{4}, \\ \langle n_{tcc} \rangle &\cong \frac{3\sqrt{6}n_D^3\sqrt{L}}{4c}. \end{aligned} \quad (7)$$

When c-ssDNAs are equal in length, using Eqs. (6) and (7) one can derive expressions for the overall rate associated with the formation of cc, icc, and tcc as follows:

$$\begin{aligned} k_{cc} &\cong k_t \frac{3\sqrt{6}n_D^3}{4\sqrt{L}} \propto \frac{1}{\eta\sqrt{L}}, \\ k_{icc} &\cong k_t \frac{3\sqrt{6}n_D^3\sqrt{L}}{4} \propto \frac{\sqrt{L}}{\eta}, \\ k_{tcc} &\cong k_t \frac{3\sqrt{6}n_D^3\sqrt{L}}{4c} \propto \frac{\sqrt{L}}{\eta c}. \end{aligned} \quad (8)$$

### D. One-step DNA hybridization model

Using the scaling results presented in Eqs. (7) and (8), we now revisit the Wetmur-Davidson model [8]. When the nucleation occurs via pure 3D diffusion controlled collision route as described in Scheme I of Fig. 1(a), the nucleation occurs with a rate  $k_N \propto k_{cc}$  and one can straightforwardly derive the scaling as  $k_N \propto \frac{1}{\eta\sqrt{L}}$ . This means that  $k_N = \zeta k_{cc}$  where  $\zeta$  is the dimensionless proportionality constant. In this background, the differential rate equations for the renaturation of a nonrepetitive c-ssDNAs can be written as follows:

$$\begin{aligned} \frac{d[2LM_D]}{dt} &= k_N [LM_1][LM_2], & \frac{d[M_D]}{dt} &= \left( \frac{k_N L}{2} \right) [M_1][M_2], \\ k_{RI} &\cong \left( k_{cc} \zeta \frac{L}{2} \right) \cong k_t \zeta \frac{3\sqrt{6}n_D^3}{8} \sqrt{L}. \end{aligned} \quad (9)$$

In this equation, we have substituted  $k_N = \zeta k_{cc}$ . Here  $k_{RI}$  is the overall hybridization rate corresponding to Scheme I,  $M_D$  (mol/lit) is the concentration of dsDNA molecules and  $M_1$  and  $M_2$  (mol/lit) are the concentrations of the c-ssDNA molecules. Upon multiplying by  $L$  (or  $2L$  in the case of dsDNA) one can convert these concentrations of the polymer molecules into the

concentrations of nucleotides in each category which are actually the experimentally observed variables. In the derivation of Eqs. (9), one assumes that the zipping is spontaneous and the nucleation is the rate-limiting step. Remarkably, Eqs. (9) associated with the Wetmur-Davidson model correctly predicts the length-dependent scaling as  $k_{RI} \propto \frac{\sqrt{L}}{\eta}$  in line with the experimental observations [8,14].

### E. Two-step mechanism of DNA hybridization

When the concentration of the nucleus is  $W$  and the average size of a nucleus is  $N$  nt (so that the concentration of nt in the nucleus form dsDNA is  $[2NW]$ ), from the two-step mechanism given in Scheme II of Fig. 1(a), one can derive the following rate equations:

$$\begin{aligned} \frac{d[2NW]}{dt} &= k_N[LM_1][LM_2] - k_Z[2NW], \\ \frac{d[2LM_D]}{dt} &= k_Z[2NW], \quad \frac{dW}{dt} \rightarrow 0, \\ \frac{d[M_D]}{dt} &\cong \left(\frac{k_N L}{2}\right)[M_1][M_2], \quad k_{RI} \cong \left(\zeta k_{cc} \frac{L}{2}\right). \end{aligned} \quad (10)$$

In this equation,  $k_{RI}$  is the overall hybridization rate corresponding to Scheme II,  $k_Z$ (nt/s) is the zipping rate constant. When zipping is much faster than the nucleation step, the rate equations given in Eqs. (10) attain steady state so that one can set  $\frac{dW}{dt} \rightarrow 0$ . Nucleation will be the rate-limiting step. In such scenario, Eqs. (10) reduce to Eqs. (9) where  $k_{cc} = k_{icc}/L$  for nonrepetitive c-ssDNA. Therefore one obtains the expression  $k_{RI} \cong (k_{cc} \zeta \frac{L}{2}) \propto \sqrt{L}$ .

### F. Limitations of one-step and two-step hybridization models

The main flaws of Eqs. (9) and (10) are as follows:

(1) Clearly, Eqs. (9) and (10) will work only for short DNA segments for which one can ignore the zipping times. However, for long c-ssDNAs, the zipping time  $\tau_Z$  scales with the length of DNA in a linear or square [15] or power-law manner [27]. In general, one observes the scaling for the zipping rate (which is the inverse of the zipping time) as  $k_Z \propto L^{-\rho}$ . There are two extreme possibilities: (1) When the zipping is diffusion like and not energetically driven or hampered by the chain entropy of the single-stranded overhangs, the zipping will be a diffusion-like process and  $k_Z \propto L^{-2}$ . (2) When the zipping is irreversible and an energetically driven process over the entropic barriers, one finds that  $k_Z \propto L^{-1}$ . In reality, the chain entropy barrier decreases in the process of zipping and the stability of duplex increases along the transition from ssDNA to dsDNA form. In general, when the entropic barrier associated with the single-strand overhangs is significant, one observes the exponent as  $1 \leq \rho \leq 2$  depending on the type of prevailing conditions. For example, when the zipping is similar to that of the forced translocation of a polymer through a nanopore [27], one finds that  $\rho \cong 1.37$ .

(2) When  $L = 1$ , one finds from the extrapolation of experimental data [21] that  $\lim_{L \rightarrow 1} k_{RI} \cong 5 \times 10^{-3} k_{sm}$  where  $k_{sm} = k_t/\phi$  is the maximum possible Smoluchowski diffusion-controlled bimolecular collision rate limit across neutral molecules. However, Eqs. (9) and (10) associated with the 3D

diffusion model predict only that  $\lim_{L \rightarrow 1} k_{RI} \cong k_{sm} \phi \frac{3\sqrt{6}n_D^3}{8}$ . This means that the contributions from the electrostatic repulsions across the negatively charged phosphate backbones over the rate enhancement at the DNA-DNA interface of c-ssDNAs should be  $\phi \cong \frac{4 \times 10^{-2}}{3\sqrt{6}n_D^3}$  where  $n_D \cong 3$  nt.

(3) When the mode of nucleation is via pure 3D diffusion, the number of correct contacts or the nucleation rate should be independent of the number of repeats in the c-ssDNA strands. Since the zipping step directly follows from the 3D diffusion-mediated nucleation, the overall renaturation rate should be independent of the complexity of c-ssDNAs. Further, the presence of trap correct contacts would slow down the nucleation process since tcc across the c-ssDNAs need to be broken before exploring other locations for the correct contacts [13,25]. This incurs a significant amount of time lapse, which in turn delays the renaturation process. On the other hand, such tcc keeps the complementary strands in the close proximity for an extended amount of time compared to icc. This is essential for the efficient operation of various one-dimensional facilitating processes such as slithering, inchworm movements, and internal displacement mechanisms [13,25], which can speed up the rate of nucleation. Detailed molecular dynamics simulations also reveal that the metastable states arising out of trap correct contacts among the short repetitive c-ssDNAs can significantly speed up the hybridization process [33].

(4) For repetitive ssDNA with complexity  $c$  and  $L/c$  number of repeats, the Wetmur-Davidson model directly assumed the scaling  $k_{RI} \propto \frac{\sqrt{L}}{c\eta}$ , which predicted that  $k_{RI} \propto \frac{1}{\sqrt{L}}$  at  $c = L$  is not consistent with their experimental observations [8,14]. This critical flaw and other arguments arising out of the simulation results [25] clearly suggest that the nucleation process must involve at least two different pathways, viz., 3D diffusion-only mode and a mode with the combination of 1D and 3D diffusions [15] as described in Schemes III and IV of Fig. 2. The 3D diffusion-only pathway progresses directly via correct contacts, and it works only for short c-ssDNAs where the zipping times can be ignored. The 3D-1D diffusion pathway can progress via either incorrect contacts or trap correct contacts depending on the repetitive nature of c-ssDNAs.

### G. Three-step models of DNA hybridization

According to Scheme III, the base clusters of c-ssDNAs interpenetrate each other upon collision to the form incorrect contacts in the first step with a rate  $k_{icc}$ . Subsequently, the correct contact will be formed via various 1D facilitating processes such as slithering, inchworm movements, and internal displacement mechanisms. These are all analogous to the facilitating processes such as sliding, hopping, and intersegmental transfers exhibited by the DNA binding proteins in the process of searching for their cognate sites on DNA [41]. Nucleation occurs upon finding the correct contacts over several rounds of incorrect contact formation, 1D slithering movements, and dissociations. Remarkably, detailed molecular dynamics simulation on the hybridization of short ssDNA segments also revealed the presence of a four-state (three-step) mechanism [33]. With this background, Eqs. (10) can be rewritten for the case of nonrepetitive c-ssDNAs as follows

(Scheme III in Fig. 2):

$$\begin{aligned} \frac{d[2UY]}{dt} &= k_{icc}[LM_1][LM_2] - (k_r + k_{NZ})[2UY], \\ \frac{d[2LM_D]}{dt} &= k_{NZ}[2UY], \quad \frac{d[Y]}{dt} \rightarrow 0, \\ \frac{d[M_D]}{dt} &\cong \frac{k_{NZ}k_{icc}L}{2(k_r + k_{NZ})}[M_1][M_2], \\ k_{RIII} &\cong \frac{k_{NZ}k_{icc}L}{2(k_r + k_{NZ})} = \frac{k_{icc}L}{2(1 + k_r\tau_{NZ})}. \end{aligned} \quad (11)$$

In these equations,  $k_{RIII}$  is the overall hybridization rate corresponding to Scheme III,  $[2UY]$  is concentration of the nucleotides involved in the icc,  $k_r$  (1/s) is the dissociation rate constant connected with the icc,  $\tau_{NZ} = \frac{1}{k_{NZ}}$  is the total time required for the nucleation and zipping through incorrect contact formation route, and its inverse will be the overall nucleation-zipping rate. After each incorrect contact, c-ssDNAs perform 1D slithering on each other over  $q$  nucleotides on average and then dissociate. The time required for such 1D diffusion will be  $\tau_q \cong \frac{l_p^2 q^2}{6D_o}$  [15]. In this time, c-ssDNAs scan  $q$  nt on each other for the presence of correct contacts. To completely scan the entire c-ssDNAs of length  $L$ , at least  $L/q$  numbers of such cycles of incorrect contact formation, 1D slithering, and dissociations are required. Therefore, the nucleation time can be expressed as  $\tau_N \cong \frac{L\tau_q}{q}$ , which is the minimum amount of time that is required by the c-ssDNAs to find a correct contact. When c-ssDNAs are nonrepetitive, one finds the total time required for the nucleation and zipping steps as  $\tau_{NZ} \cong \tau_Z + \tau_N$  where  $\tau_Z \cong \frac{L}{k_+}$  and  $\tau_N \cong \frac{l_p^2 q L}{6D_o}$  are the zipping and nucleation times through the incorrect contact route [15]. The nucleation rate will be the inverse  $k_N = \frac{1}{\tau_N}$  and the zipping rate will be  $k_Z = \frac{1}{\tau_Z}$ . Noting that fact that nucleation will be immediately followed by zipping, we combine the nucleation and zipping times.

Since the stabilizing effects of the already formed Watson-Crick base pairs are much stronger than the entropic barriers arising out of the freely moving single-stranded overhangs, one can assume the linear scaling for the zipping time with the length of c-ssDNAs as  $\tau_Z \propto L$ . Here  $k_+$  (nt/s) is the microscopic zipping rate (Fig. 2),  $l_p = 1$  bp,  $q$  (nt) is the average slithering length, and  $D_o$  (bp<sup>2</sup>/s) is the phenomenological 1D diffusion coefficient associated with the slithering dynamics of c-ssDNA segments. Upon inserting the expressions for the nucleation and zipping times into the expressions for  $k_{RIII}$  as given in Eq. (11) one obtains the following result:

$$k_{RIII} \cong \frac{k_{icc}L}{2(1 + k_r\tau_{NZ})} = \frac{k_r \frac{3\sqrt{6}m_p^3\sqrt{L}}{4} L}{2\left[1 + k_r\left(\frac{L}{k_+} + \frac{l_p^2 q L}{6D_o}\right)\right]} \propto \frac{\sqrt{L}}{\eta}. \quad (12)$$

One should note that  $D_o$  will be influenced only by the local viscosity at the DNA-DNA interface of the c-ssDNAs with icc or tcc, and it will not be much influenced by the global viscosity that is connected with the translational diffusion of the entire c-ssDNA nucleotide cluster. In the presence of repetitive sequences, apart from icc and cc, trap correct contacts are also formed as shown in Scheme IV of Fig. 2. Clearly, the nucleation can occur via both icc and tcc in

the case of repetitive c-ssDNAs, which is also substantiated by the coarse-grained molecular dynamics simulations [29]. Particularly, these tcc can influence the renaturation in two possible ways: (1) they keep the complementary strands in the close proximity for prolonged amount of time that enhances the slithering times, which in turn increases the efficiency of searching for the correct contacts [29,33] or (2) these kinetic traps prolong the overall renaturation timescales, which will be more prominent especially when the length of ssDNA is very long. As a result of these, the dissociation rate connected with the partial duplexes decreases ( $k_r \rightarrow k_{rt}$ ) and the slithering length increases ( $q \rightarrow q_t$ ). The overall renaturation rate corresponding to the tcc route can be obtained by replacing ( $k_r \rightarrow k_{rt}$ ,  $q \rightarrow q_t$ ,  $k_{icc} \rightarrow k_{tcc}$ ) in Eq. (12). Upon combining the contributions of these parallel pathways through icc and tcc routes, one finally obtains the following expression for the overall renaturation rate for the repetitive c-ssDNAs:

$$\begin{aligned} k_{RIV} &\cong \frac{k_{icc}L}{2(1 + k_r\tau_{NZ})} + \frac{k_{tcc}L}{2(1 + k_{rt}\tau_{NZR})} \\ &\cong \frac{3k_t\sqrt{6}n_D^3\sqrt{L}}{4} \left\{ \frac{L}{2\left[1 + k_r\left(\frac{L}{k_+} + \frac{l_p^2 q L}{6D_o}\right)\right]} \right. \\ &\quad \left. + \frac{L}{2c\left[1 + k_{rt}\left(\frac{L}{k_+} + \frac{l_p^2 q_t L}{6D_o}\right)\right]} \right\}. \end{aligned} \quad (13)$$

Equation (13) is the central result of this paper. In this equation,  $k_{RIV}$  is the overall hybridization rate corresponding to the repetitive c-ssDNAs as described by Scheme IV,  $\tau_{NZR} \cong \tau_Z + \tau_{NR}$  is the nucleation-zipping time via trap correct contact route where  $\tau_{NR} \cong \frac{l_p^2 q_t L}{6D_o}$  is the average nucleation time,  $k_{rt}$  is the dissociation rate connected with tcc, and  $q_t$  is the 1D slithering length associated with the repetitive c-ssDNAs with partial duplexes and single-strand overhangs. The nucleation rate will be the inverse  $k_{NR} = \frac{1}{\tau_{NR}}$ . When  $L > c$ ,  $k_r > k_{rt}$  since the partial duplexes are more stable than the incorrect contacts and  $q < q_t$  since the c-ssDNAs stay close to each other for a prolonged amount of time, which allows extended amount of slithering. As a result of these, we find that  $\left[k_r\left(\frac{1}{k_+} + \frac{l_p^2 q}{6D_o}\right)\right] \ll \left[k_{rt}\left(\frac{1}{k_+} + \frac{l_p^2 q_t}{6D_o}\right)\right]$  from which one obtains the generally observed scaling as  $k_{RIV} \propto \frac{\sqrt{L}}{c\eta}$ . Remarkably, when  $c = L$ , Eq. (13) predicts the correct scaling  $k_{RIV} \propto \frac{\sqrt{L}}{\eta}$  in line with the experimental observations [8,14].

Equation (13) will be still valid when c-ssDNAs are unequal in length similar to that of the template and probe DNAs used in the PCR reactions. However, in this case the length of duplex formed upon hybridization will be equal to the length of the probe. Let us denote the lengths of the template and probe as  $L$  and  $u$  respectively. When  $L \gg u$ , the template strand diffuses slower than the probe strand, and one can show that the bimolecular collision rate associated with the template-probe system scales as  $k_t \propto \sqrt{\frac{L}{u}}$ . This follows from the fact that  $k_t \cong \frac{2k_B T}{3\eta} \frac{(R_u + R_L)^2}{R_u R_L}$  where  $R_u \propto \sqrt{u}$  and  $R_L \propto \sqrt{L}$  are the radius of gyration of the probe and template strands, respectively [35]. When  $u \ll L$ , the reaction volume will be almost independent of the volume of the probe. This results in

the scaling as  $k_{cc} \propto \frac{\sqrt{u}}{L}$  for the rate of formation of the correct contacts. Since the nucleation rate is directly proportional to the rate of formation of the correct contacts across c-ssDNAs, this means that the nucleation rate scales with the lengths of the probe and template c-ssDNAs as  $k_N \propto \frac{\sqrt{u}}{L}$  since  $k_N \propto k_{cc}$ , which reduces to the observed scaling corresponding to the nucleation rate  $k_N \propto \frac{1}{\sqrt{L}}$  as  $c \rightarrow L$ .

### III. SIMULATION METHODS

To understand various scaling relationships associated with the correct, incorrect, and trap correct contacts in the process of renaturation, we modeled the c-ssDNAs as self-avoiding random walks confined inside a 3D lattice box that mimics the reaction volume.

(1) When the box is a cube with side  $b$ , the box volume will be  $V = b^3$  cubic units. We denote the position on this lattice box with coordinated  $(X, Y, Z)$ . Here  $(X, Y, Z) = (0, 0, 0)$  and  $(b, b, b)$  are the reflecting boundaries for the generation of SARWs 1 and 2, which mimic c-ssDNA strands 1 and 2 confined inside the reaction volume.

(2) Consider a cubic lattice with boundaries  $(0, b)$  in all  $(X, Y, Z)$  dimensions. The first positions of SARWs 1 and 2 will be randomly chosen by calling three uniform distributed random integers inside  $(0, b)$ , which are the initial positions denoted as  $(X_0, Y_0, Z_0)$ . The total number of points in a given SARW represents the number of monomers in c-ssDNAs. We denote these points by the index starting from 1 to  $L$  where  $L$  is the total length.

(3) Let us assume that the current position of the growing SARW of interest at the end of  $k$ th step is  $(X, Y, Z)$ . The subsequent point of this SARW will be generated by sequentially calling three real random numbers from the uniform distribution inside  $r_{1,2,3} \in (0, 1)$ . If  $r_1 < 0.5$ , then  $X \rightarrow (X + 1)$  else  $X \rightarrow (X - 1)$  and similar rules were set for  $Y$  and  $Z$  variables. The next  $(k + 1)$ th position of a SARW could be any one from the eight possibilities  $(X \pm 1, Y \pm 1, Z \pm 1)$ . Hence the obtained new position will be checked against the earlier  $k$  positions of the growing SARW for self-intersections. Only those moves without self-intersections will be allowed.

(4) The boundaries at  $(0$  and  $b)$  are reflecting ones so that those moves result in  $(X, Y, Z) = (-1, -1, -1)$  or  $(b + 1, b + 1, b + 1)$  will not be allowed. For example, those moves which result in the new position  $(X, Y, Z) = (-1, 1, 2)$  and  $(1, 2, b + 1)$  will not be allowed.

(5) When all these eight possible moves are self-intersecting, such dead-end walks will be dropped and a new SARW will be started from step 2 by choosing random initial position similar to the Rosenbluth algorithm [42].

(6) SARW 1 and SARW 2 cannot self-intersect, but they can cross-intersect each other that mimics the interpenetration and contact formation among the pair of c-ssDNAs. These cross-intersection points can be classified into correct, incorrect, or trap correct contacts.

(7) When the sequence is repetitive with complexity  $c$ , in such a polymer of length  $L > c$ , the positions from 1 to  $c$ ,  $c + 1$  to  $2c$ ,  $2c + 1$  to  $3c$ , and so on are identical in sequence. To model such repetitive c-ssDNAs, SARW of size  $L$  will be fragmented into  $L/c$  number of blocks, and the sequence

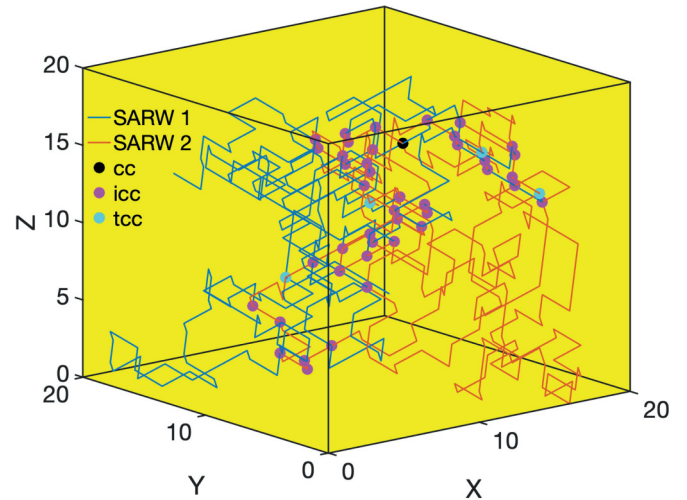


FIG. 3. Lattice model of DNA renaturation. The complementary single strands are modeled as self-avoiding random walks [SARWs 1 (blue) and 2 (red)] confined inside the reaction volume defined by the lattice box. The positions of the nucleotides are marked from 1 to  $L$ . Both strands arrive at the reaction volume via 3D diffusion. Correct contact between these SARWs occurs when there is an exact registry match that can further lead to nucleation and zipping. For example, when there is a contact between position 7 of SARW 1 with position 7 of SARW 2, it is defined as a correct contact (cc). When there is a contact between position 7 of SARW 1 with position 5 of SARW 2, it is an incorrect contact (icc). When the complexity is  $c < L$ , the SARW has  $L/c$  number of repeats, i.e., the sequence spanned across position 1 to  $c$ ,  $c + 1$  to  $2c$ ,  $2c + 1$  to  $3c$ , and so on will be the same. Therefore, when there is a contact between position 1 of SARW 1 with position  $c + 1$  (or  $2c + 1$  and so on) of SARW 2, it is defined as a trap correct contact (tcc). Here tcc can lead to the formation of partial duplexes with single-stranded overhangs which are actually kinetic traps in the renaturation pathway. The settings are  $L = 250$ ,  $c = 5$ , and reaction volume  $V = 20^3$ , the number of incorrect contacts ( $n_{icc}$ ) = 25, trap correct contacts  $n_{tcc} = 4$ , and the number of correct contacts  $n_{cc} = 1$  that occur at  $(X, Y, Z) = (9, 8, 17)$ .

indices of all the blocks will be from 1 to  $c$  apart from the original overall sequence index, which runs from 1 to  $L$ .

(8) Those contacts between c-ssDNAs without exact registry matches are the incorrect contacts. The positions from 1 to  $c$  of SARW 1 can form duplex with position stretch  $c + 1$  to  $2c$  of SARW 2 and so on. These cross-intersections can be classified as correct contact (cc), incorrect contact (icc), or trap correct contact (tcc) depending the position of the registry match. For example, when there is a cross-intersection between the overall sequence index 4 of SARW 1 and 7 of SARW 2, it is an incorrect contact in the case of nonrepetitive c-ssDNAs. When there is a cross-intersection between the original sequence index 7 of SARW 1 and 7 of SARW 2, it is a correct contact. When there is a cross-intersection between the block 1 sequence index 3 of SARW 1 and block 5 sequence index 3 of SARW 2, it is defined as a trap correct contact. When there is a cross-intersection between the block 1 sequence index 3 of SARW 1 and block 5 sequence index 5 of SARW 2, it is an incorrect contact. In this setting, the volume of a SARW with length  $L$  will be  $vL$  where  $v = 1$  is the volume of the monomer unit.



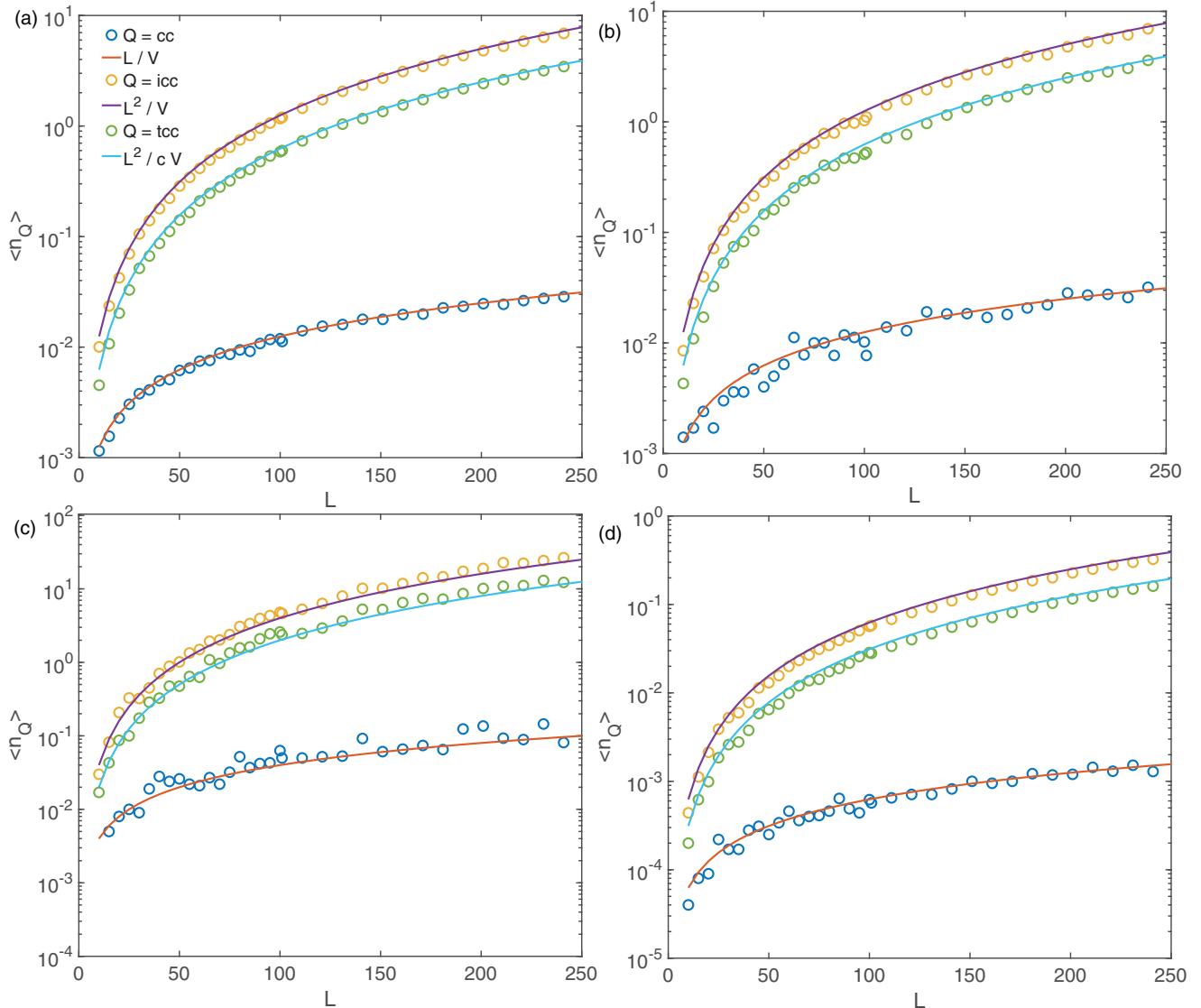


FIG. 4. Variation of the average number of correct ( $n_{cc}$ ), incorrect ( $n_{icc}$ ), and trap correct contacts ( $n_{tcc}$ ) with respect to the length  $L$  of SARWs and dimensionality. Average number of various contacts were obtained over  $10^5$  SARW trajectories. Hollow circles are the stochastic simulations results, and solid lines are the predictions by Eqs. (2). Irrespective of the dimensionality and shape of the lattice box where the volume of the monomer unit is  $v = 1$ , the expressions given in Eqs. (2) are valid. (a) 3D SARW confined in volume  $V = 20^3$ , length iterated in  $L = (10, 250)$ , and  $c = 2$ . (b) 3D SARW confined inside a box with base area =  $10^2$  (10 by 10) and length = 80 so that  $V = 8000$ . The length of the SARW is iterated inside  $L = (10, 250)$  with complexity  $c = 2$ . (c) 2D SARW confined in  $V = 50^2$ , and the length of the SARW iterated inside  $L = (10, 250)$  and  $c = 2$ . (d) 4D SARW confined in  $V = 20^4$ ,  $L = (10, 250)$  and  $c = 2$ .

(9) To understand the effects of sheared c-ssDNAs on the overall rate of renaturation, we randomized the lengths of SARWs. The length of the original c-ssDNA is  $L$ , and we denote the index of the monomers from 1 to  $L$ . Shearing of c-ssDNAs will generate fragments of this template strand with random lengths. To mimic this, two uniform distributed random integers  $r_1, r_2$ , were generated inside  $(1, L)$ , which are the random lengths of SARW 1 and SARW 2. The starting sequence index of SARW 1 was defined by calling a uniform random integer  $s_1$  inside  $(1, L - r_1)$  and the end point will be  $s_1 + r_1$ . Similarly the starting sequence index of SARW 2 was defined by calling a random integer  $s_2$  inside  $(1, L - r_2)$ , and the end point will be  $s_2 + r_2$ . These start and end locations of SARWs 1 and 2 were used to compute the number of

correct, incorrect and trap correct contacts as described in step 8.

We assume that the pair of c-ssDNAs have already reached the reaction volume  $V$  via 3D translational diffusion with a bimolecular collision rate  $k_t$ . The quantities that we are interested to compute here are the average number of correct, incorrect, and trap correct contacts formed between a pair of c-ssDNAs of length  $L$  and complexity  $c$  inside the reaction volume  $V$ . To achieve this, several pairs of SARWs were generated inside a closed cubic lattice box as in Fig. 3, and the number of cc, icc, and tcc were enumerated and averaged over  $10^5$  trajectories. The average number of cc, icc, and tcc seems to be influenced by the length of c-ssDNA ( $L$ ), its sequence complexity ( $c$ ), and the reaction volume ( $V$ ). To investigate

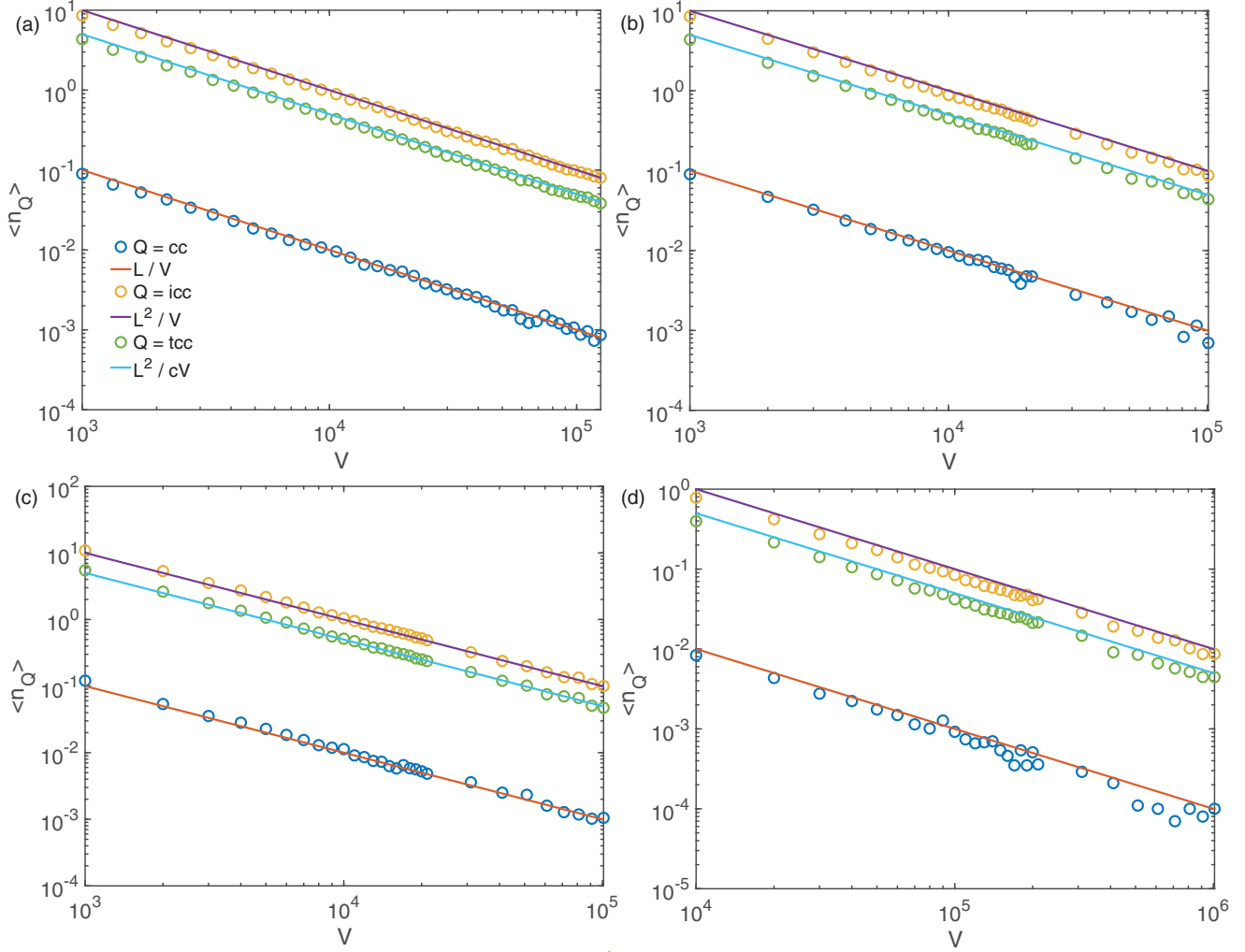


FIG. 5. Variation of the average number of correct ( $n_{cc}$ ), incorrect ( $n_{icc}$ ), and trap correct contacts ( $n_{tcc}$ ) with respect to change in confinement volume of SARW and dimensionality. Average number of various types of contacts was obtained over  $10^5$  trajectories. Hollow circles are the stochastic simulations results, and solid lines are the predictions by Eqs. (2). Irrespective of the dimensionality and shape of the lattice box where the volume of the monomer unit is  $v = 1$ , the expressions given in Eqs. (2) are valid. (a) 3D SARW confined in cubic box with volumes iterated inside  $V = (10^3, 50^3)$ , length of SARW set as  $L = 100$ , and  $c = 2$ . (b) 3D SARW confined inside a box with base area  $= 10^2$ , and length iterated from 10 to 1000 so that  $V$  varies from  $10^3$  to  $10^5$ . The length of SARW is  $L = 100$  with complexity  $c = 2$ . (c) 2D SARW confined in box with base side  $= 100$  and length iterated from 10 to 1000 so that  $V$  varies from  $10^3$  to  $10^5$ ,  $L = 100$ , and  $c = 2$ . (d) 4D SARW confined in a box with base  $10^3$  ( $10 \times 10 \times 10$ ) and the other side iterated from 10 to 1000 so that the volume  $V$  varies from  $10^4$  to  $10^6$ , length of SARW set as  $L = 100$ , and the complexity set as  $c = 2$ .

the effects of these variables, we iterated one factor over a range of values at a time by fixing other two factors at constant values. In the real situations, one can translate the results of the lattice model by setting the reaction volume as the volume of the sphere with radius equals to  $2R_g$  where  $R_g$  is the average radius of gyration of c-ssDNAs. To understand the distribution of icc, cc, and tcc, we constructed the histogram of samples drawn from large population of counts over  $10^4$  numbers of SARW trajectories.

#### IV. RESULTS AND DISCUSSION

The c-ssDNA strands can be modeled as self-avoiding random walks with the average radii of gyration  $R_g$ . Renaturation of c-ssDNAs requires the formation of correct

contacts between them that leads to nucleation and zipping. Here c-ssDNAs reach the reaction volume via 3D translational diffusion. Upon entering the reaction volume, they can interpenetrate each other to form various types of contacts such as cc, icc, tcc, or no contact at all. Clearly, when the size of c-ssDNAs is approximately equal, then the rate at which they enter the reaction volume via translational diffusion ( $k_t = \frac{8k_B T}{3\eta} \phi$ ) will be independent of the radius of gyration. However, the extent of interpenetration of these strands will be strongly dependent on the length of c-ssDNAs  $L$  and the confinement volume  $V$ .

Detailed lattice model simulations revealed the general scaling of various types of contacts between c-ssDNA strands as  $n_Q \propto \frac{vL^v}{V}$  for nonrepetitive c-ssDNAs where  $Q =$  (icc, cc) and  $v$  is the monomer volume. For cc, we find the

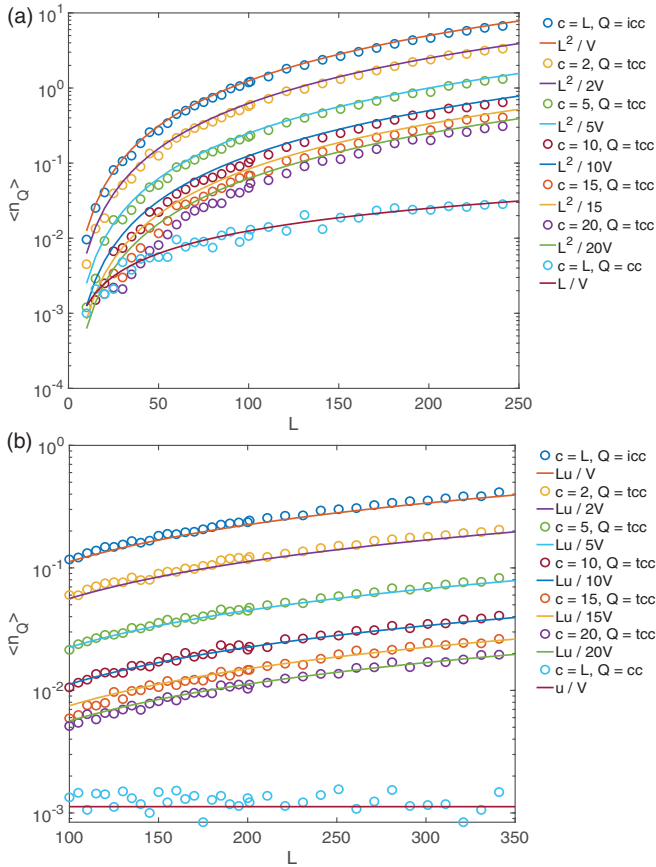


FIG. 6. Variation of the average number of correct ( $n_{cc}$ ), incorrect ( $n_{ice}$ ), and trap correct contacts ( $n_{tcc}$ ) with respect to changes in length  $L$  and complexity  $c$  of SARWs. Average number of various types of contacts was obtained over  $10^5$  trajectories. Hollow circles are the stochastic simulation results, and solid lines are the predictions by Eqs. (3). The volume of the monomer unit is set as  $v = 1$ . (a) The volume is fixed at  $V = 20^3$ . (b) The template strand length is fixed at  $L = 100$ , and the length of the probe is fixed at  $u = 10$ . The probe is set to form duplex with the template at position stretch 40 to 50. This is the recognition stretch of the probe ssDNA. Both these strands are embedded inside the volume  $V = 20^3$ .

exponent  $\gamma = 1$ , and for icc we find that  $\gamma = 2$ . For repetitive c-ssDNAs one finds that  $n_{tcc} \propto \frac{vL^\gamma}{cV}$  with  $\gamma = 2$ . When  $c = L$ , this reduces to  $n_{tcc} \propto \frac{vL^{\gamma-1}}{V}$  resembling the nonrepetitive case. Variations of the number of tcc, cc, and icc with respect to changes in  $L$  with fixed  $V$  and  $c$  are demonstrated in Fig. 4. Variations of the number of tcc, cc, and icc with respect to changes in  $V$  with fixed  $L$  and  $c$  are demonstrated in Fig. 5. Results clearly suggest that the scaling  $n_Q \propto \frac{vL^\gamma}{V}$  where  $Q = (\text{icc}, \text{cc})$  is independent of the shape and dimension of the confining lattice box as shown in Figs. 4(b)–4(d) and 5(b)–5(d). Similar results for the repetitive c-ssDNAs are demonstrated in Fig. 6(a). We can conclude from these results that the number of correct contacts scales with  $L$  and  $V$  as  $n_{cc} \propto \frac{vL}{V}$  irrespective of the complexity of c-ssDNA strands and shape and dimension of the confining lattice box.

These scaling relationships are not affected when the lengths of c-ssDNAs are unequal as demonstrated in Fig. 6(b) or random as demonstrated in Fig. 7. Particularly, when the

length of the template ssDNA is  $L$  and probe ssDNA is  $u$  as in the case of annealing phase of polymerase chain reactions, then one obtains the scaling  $n_{cc} \propto \frac{vu}{V}$ . When the c-ssDNA lengths are random with equal distribution inside  $(0, L)$ , one observes the scaling  $n_{cc} \propto \frac{v\langle L \rangle}{2V}$  where  $\langle L \rangle = \frac{L}{2}$  here. These results are demonstrated in Fig. 7. Under *in vitro* conditions,  $V$  will be the reaction volume. Since the reaction radius here is approximately equal to  $2R_g$ , one can consider the reaction volume as  $V \cong \frac{4}{3}\pi(2R_g)^3$ .

To understand the effect of varying the confinement volume on the radius of gyration, we iterated the length  $L$  of SARWs at different confinement volumes. From Flory's theory, one can conclude that the radius of gyration of the spatially unconfined 3D SARW approximately scales with the length as  $R_g \propto L^{3/5}$  [43]. One naturally expects the scaling exponent  $\theta$  in  $R_g \propto L^\theta$  as  $\lim_{V_S \rightarrow \infty} \theta \rightarrow \frac{3}{5}$ . However, when the SARW trajectory is confined inside a cubic box, then the magnitude of the scaling exponent decreases in a complicated manner since the SARW trajectory gets reflected at boundaries of the box, which in turn results in tight packaging of SARWs. In this line, several asymptotic functions for the exponent were tried to fit the data on the computed  $R_g$  versus the confinement volume of SARWs. However, the volume and length dependency of  $R_g$  seems to best fit the functional form  $R_g \cong \alpha\sqrt{L}(1 + \beta L)^{-1}$  where  $\beta$  is a volume-dependent parameter. Nonlinear least-square fitting using the Marquardt-Levenberg algorithm [40] to this function revealed an approximate functional form for the parameter  $\beta \cong [\varepsilon(\frac{v}{V_S})^\delta + \sigma]$  where  $v$  is the monomer volume,  $V_S$  is the confinement volume with the fit parameters  $\varepsilon = 0.5 \pm 0.04$ ,  $\delta = 0.61 \pm 0.1$ , and  $\sigma = -10^{-4} \pm 10^{-5}$  at 95% confidence level. These results are summarized in Fig. 8. The parameter  $\alpha$  seems to be independent of  $V_S$  as shown in Fig. 8(b). Clearly, one obtains the limiting condition  $\lim_{V_S \rightarrow \infty} R_g \cong \alpha\sqrt{L}$  noting the fact that  $\beta \rightarrow 0$  under such conditions as shown in Fig. 8(c). Since this limiting condition is valid under *in vitro* conditions, one can use this limiting expression to calculate the molecular and reaction volumes. Remarkably, the numbers of cc, icc, and tcc show a bimodal-type distribution with zero spike as demonstrated in Fig. 9. The reason for the zero spike could be that the confinement volume here is much larger than the intrinsic volume of the c-ssDNA polymer. This means that the magnitude of the zero spike is inversely proportional to the volume ratio  $vL/V$ . The number of tcc decreases as the sequence complexity increases, which is evident from Figs. 9(c)–9(f).

When the average radius of gyration of c-ssDNAs scales as  $R_g \propto \sqrt{L}$ , the reaction volume scales with  $L$  as  $V \propto L^{3/2}$ . As a result, one observes the scaling associated with the average number of correct contacts formed between the c-ssDNAs upon interpenetration as  $\langle n_{cc} \rangle \propto \frac{1}{\sqrt{L}}$ . Since the rate of nucleation is directly proportional to the number of correct contacts, one observes  $k_N \propto \frac{1}{\sqrt{L}}$  in line with the experimental observations. When c-ssDNAs reach the reaction volume element via 3D translational diffusion, one finally obtains the viscosity dependence as  $k_N \propto \frac{1}{\eta\sqrt{L}}$ . The overall renaturation rate in the case of pure 3D diffusion model [Scheme I of Fig. 1(a)] is directly proportional to the nucleation rate as well as  $L$ . As a result, we finally arrive at the scaling for overall renaturation

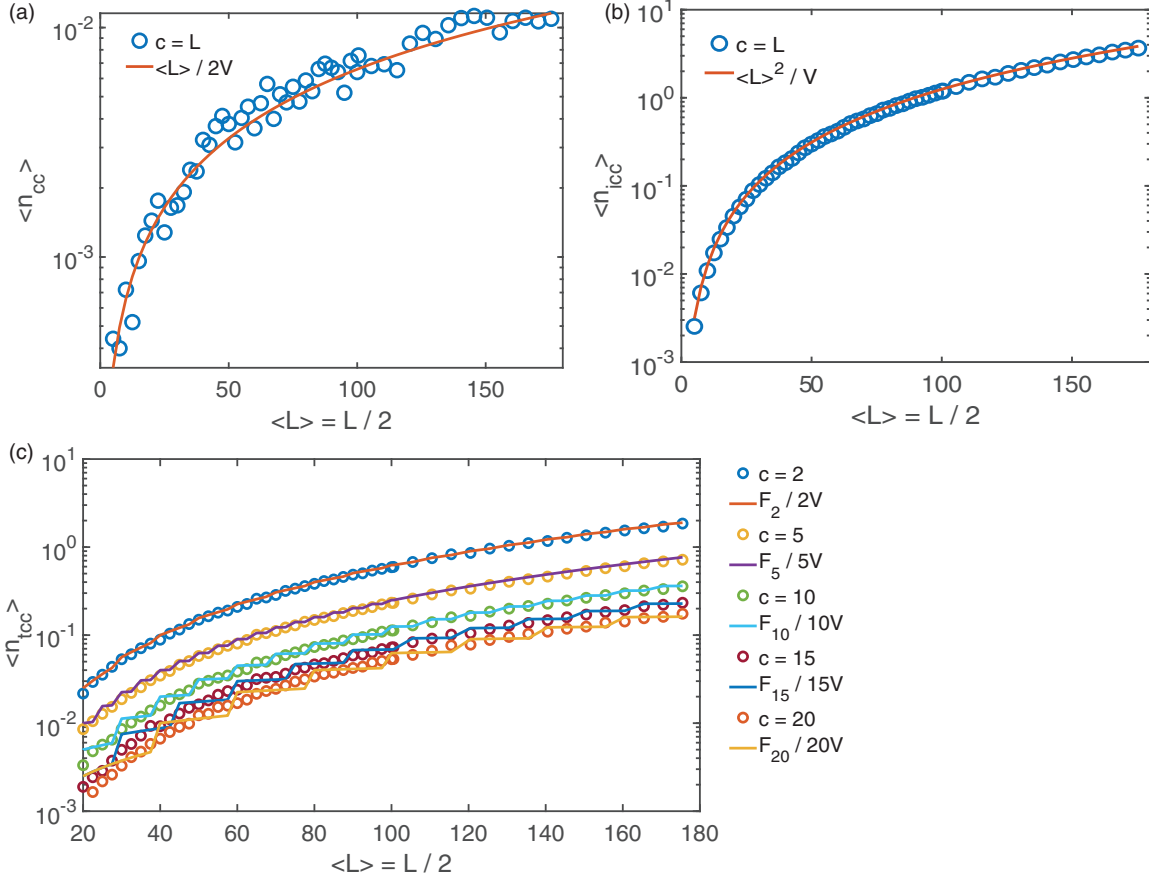


FIG. 7. Variation of the average number of correct ( $n_{cc}$ ), incorrect ( $n_{icc}$ ), and trap correct contacts ( $n_{tcc}$ ) with respect to changes in the randomized lengths and complexity  $c$  of ( $c$ -ssDNAs) SARWs. Average number of various types of contacts obtained over  $10^5$  SARW trajectories. Hollow circles are the stochastic simulation results, and solid lines are the predictions by Eqs. (4). The volume is set at  $V = 20^3$ , and the length of the SARW was chosen randomly within  $(0, L)$  with equal probabilities so that the average length is  $\langle L \rangle = L/2$ .  $L$  is iterated inside  $(10, 360)$  so that the average length  $\langle L \rangle$  will vary from 5 to 180. Results suggest that the probability of finding the correct contact upon collision is  $1/L$  rather than  $1/\langle L \rangle$ , which leads to  $\langle n_{cc} \rangle = L/4V$ . (a)  $\langle n_{cc} \rangle$ ; (b)  $\langle n_{icc} \rangle$ ; (c)  $\langle n_{tcc} \rangle$ . When the sequence complexity is close to the sequence length, one finds the approximation  $\lim_{c \rightarrow \langle L \rangle} \langle n_{tcc} \rangle \cong \frac{v(\langle \bar{L} \rangle^2 + \langle l \rangle^2)}{cV}$  where  $\langle \bar{L} \rangle = c \lfloor \frac{\langle L \rangle}{c} \rfloor$  is the length of DNA containing full repeats and  $\langle l \rangle = \langle L \rangle - c \lfloor \frac{\langle L \rangle}{c} \rfloor$  is the length of DNA left over with partial repeat. Here  $v = 1$  and  $F_c = \langle \bar{L} \rangle^2 + \langle l \rangle^2$ . One also should note that Eqs. (4) will be valid for the repetitive DNA sequences only when  $(L/c) \gg 1$ .

rate as  $k_{RI} \propto \frac{\sqrt{L}}{\eta}$  in line with the experimental observations [8] on the hybridization of nonrepetitive  $c$ -ssDNAs.

In the case of repetitive  $c$ -ssDNAs, the average number of correct contacts is independent of the sequence complexity or the number of repeating elements since it requires only the exact registry match. Clearly, models based on only the 3D diffusion cannot explain the complexity dependence of the renaturation rate. For example, when we assume the scaling  $k_{RI} \propto \frac{\sqrt{L}}{c\eta}$  as in the case of the Wetmur-Davidson model, then we will end up with the inconsistency when  $c \rightarrow L$ , which predicts the scaling  $k_{RI} \propto \frac{1}{\sqrt{L}}$ , which is not in line with the experimental observations [8]. These arguments clearly suggest that the renaturation follows multiple and parallel pathways in the presence of repetitive sequences. Particularly, the 1D slithering dynamics plays critical roles in the renaturation of the repetitive sequences. The trap correct contacts between the repetitive  $c$ -ssDNAs can lead to the formation of partial duplexes with single-strand overhangs. Although the entropy component of the single-strand overhangs will destabilize

these kinetic traps in the pathway of renaturation, tcc keep the  $c$ -ssDNAs in the close proximity for a prolonged timescale, which is essential for efficient slithering dynamics and internal displacement mechanisms, which in turn accelerates the search for the correct contacts. Clearly, the extent of possible slithering dynamics is directly proportional to the number of trap correct contacts, which is inversely proportional to the sequence complexity.

The pathway of renaturation via an incorrect contact route always operates irrespective of the presence of repeating elements. Therefore, the overall renaturation rate associated with the repetitive  $c$ -ssDNAs will be the sum of rates corresponding to the incorrect contact and trap correct contact routes as described in Eq. (13). When the sequence complexity is close to the length of the  $c$ -ssDNAs, the number of tcc and the associated slithering lengths will be much limited. This follows from the fact that the stability of incorrect contacts will be much lower than the trap correct contacts. As a result, the pathway via incorrect contacts will be the dominating one when  $c$  is close to  $L$ , which is characterized by the scaling  $k_{RI} \propto \frac{\sqrt{L}}{\eta}$ .

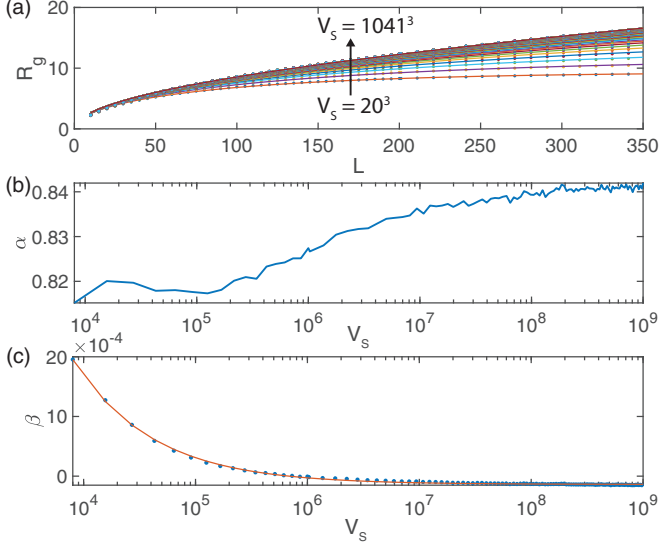


FIG. 8. Variation of the average radius of gyration  $R_g$  of SARWs with respect to changes in the length  $L$  and the confinement volume  $V_s$  of the SARW. The radius of gyration of a chain molecule was calculated as  $R_g = \langle |\sqrt{\frac{\sum_{i,j=1}^L (\bar{w}_i - \bar{w}_j)^2}{2L^2}}| \rangle$  where  $\bar{w}_i$  is the vector coordinates  $(X, Y, Z)$  of the  $i$ th monomer and  $L$  is the total length. Average of  $R_g$  is obtained over  $10^5$  numbers of SARW trajectories. Filled circles in (a) and (c) are the numerical results, and solid lines are the nonlinear least-square (NLS) fittings. (a) NLS fits to function  $R_g \cong \alpha \sqrt{L(1 + \beta L)^{-1}}$  with respect to  $L$  at various  $V_s$  values reveal the parameter  $\alpha$  does not change much with  $V_s$  as given in (b) and  $\beta = [\varepsilon(\frac{v}{V_s})^\delta + \sigma]$  as given in (c) where  $\varepsilon = 0.5 \pm 0.04$ ,  $\delta = 0.61 \pm 0.1$ , and  $\sigma = -10^{-4} \pm 10^{-5}$ . These values were obtained with 95% confidence level. The monomer volume is set to  $v = 1$  cubic unit.

The number of tcc and the extent of slithering lengths will be much higher when the sequence complexity is much lesser than the length of c-ssDNAs. Under such conditions, the pathway via trap correct contacts will be the dominating one, which is characterized by the scaling  $k_{RIV} \propto \frac{\sqrt{L}}{c\eta}$  in line with the experimental observations [8]. Clear differences between one-step, two-step, and three-step hybridization models are summarized in Table I.

### A. Computational evidence for the three-step model

The single-molecule coarse-grained molecular dynamic trajectory of the hybridization of oxDNA shows a four-state mechanism which comprises [29] random collisions between the complementary ssDNAs with oscillations in the systematic energy, which is followed by nucleation, zipping, and formation of duplex DNA. The oscillations in the systematic energy upon random collisions between c-ssDNAs represents the existence of several rounds of incorrect contact formations and dissociations of our model. The step corresponding to the initialization of hybridization represents the nucleation. Remarkably, the average zipping time seems to be almost independent of the electrostatic interactions arising at the DNA-DNA interface of c-ssDNAs, which implies that the zipping is not a rate-determining step, especially for short c-ssDNA segments [29]. However, our theory predicts a linear

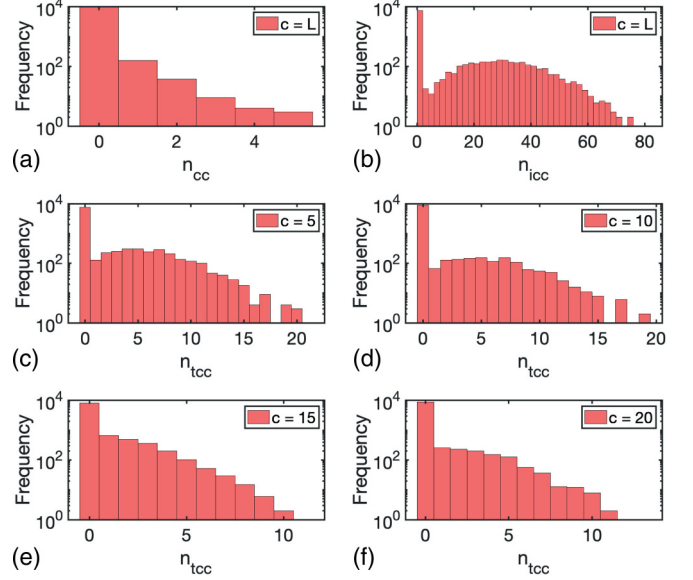


FIG. 9. Distributions of the number of correct ( $n_{cc}$ ), incorrect ( $n_{icc}$ ), and trap correct contacts ( $n_{tcc}$ ) at various sequence complexity  $c$  values. The volume of the lattice cube is set as  $V = 20^3$ , the length of the SARW is fixed at  $L = 250$ , and the sequence complexity  $c$  is varied as  $c = (5, 10, 15, 20, 250)$ . Histograms of various types of contacts are constructed over  $10^4$  numbers of SARW trajectories. Clearly, all these distributions show a bimodal type with zero spike that represents no contact cases. (a), (b) Complexity is set to  $c = 250$ . (c)  $c = 5$ ; (d)  $c = 10$ ; (e)  $c = 15$ ; (f)  $c = 20$ .

scaling of the zipping time with the length of c-ssDNAs, which means that the zipping time cannot be ignored when the length of c-ssDNAs is very large.

Coarse-grained molecular dynamics simulation [33] of the hybridization of repetitive ATATATATAT oligomer revealed the existence of out-of-register shifted metastable states (partially hybridized DNA molecules at tcc with single-strand overhangs of our model). The rate of formation of dsDNA from these metastable states seems to be twice faster than the formation rate from the corresponding fully dissociated c-ssDNAs [33]. This observation is in line with our theoretical prediction that the partially hybridized metastable states via tcc retain the c-ssDNAs in close proximity for prolonged timescales, which enhances the 1D slithering-type search for the correct contacts to initiate the nucleation. This enhancement effect can be represented by the inequality condition  $q < q_t$  in Eq. (13). However, our theory also predicted the retardation effects of such metastable trap configurations on the overall hybridization rate, which can be represented by the inequality  $k_r > k_{rt}$  in Eq. (13). Insertion of a GC pair at the terminal or middle position of this poly-AT oligomer drastically reduced the number of out-of-register shifted metastable states leading to a two-state hybridization scheme [33]. This observation suggested that the ruggedness of the interaction energy landscape of the single-stranded poly AT oligomers gets smoothed and funneled upon insertion of a GC pair. That is to say, those out-of-state shifted metastable states are funneled towards the stable GC contact-nucleus, which indirectly reveals the presence of 1D slithering dynamics.

TABLE I. Comparison of various hybridization models.

Model	Summary	Predictions and limitations
One-step model (Scheme I)	Hybridization occurs via one-step 3D diffusion-controlled collision	(1) This model predicts an inverse scaling of the hybridization rate with the length of DNA, which contradicts the experimental findings that the hybridization rate scales with the length of DNA in a square root manner. (2) Zipping time is ignored. Therefore, this model will work only for short DNAs. (3) Correctly predicts the viscosity dependence of the hybridization rate.
Two-step model (Scheme II)	Hybridization occurs via nucleation in the first step, which is followed by zipping in the second step	(1) This model predicts correct scaling of the hybridization rate with the length and complexity of DNA as long as the complexity is much lower than the length. However, when the complexity approaches the length, this model predicts inverse square root scaling of the hybridization rate with the length of DNA, which is against the experimental observations. (2) Translational diffusion of DNA is ignored, and the microscopic diffusion dynamics at the time of nucleation is used to explain the viscosity dependence of the hybridization rate.
Three-step model presented (Schemes III and IV)	Hybridization occurs via 3D diffusion-mediated incorrect as well as trap correct contacts in the first step, followed by nucleation via 1D slithering in the second step, followed by zipping in the third step. In Scheme IV, slithering pathways are different for incorrect and trap correct contact-mediated nucleation.	(1) Detailed contributions of trap correct contacts arising from the repetitive DNA over the hybridization rate were not considered in Scheme III.  (2) (a) Scheme IV correctly predicts the length and complexity dependence of the hybridization rate irrespective of the relative levels of complexity and length of the interacting DNA sequences. (b) It correctly predicts the viscosity dependence of the hybridization rate.

The possibility of bipedal-directed walks of short fragments of DNA over a template DNA track have been demonstrated by several groups [44–48]. The directional-dependent movement of these DNA walkers requires the input of an external energy. In this context, according to our three-step model, the correct contact formation across the c-ssDNAs is via a combination of 1D and 3D diffusion routes, which is an unbiased random walk that is mainly driven by the background thermal energy.

Although the computational studies could reveal the finer details about the mechanism of DNA hybridization at the molecular level, our simplified lattice model is able to capture the overall dynamical aspects of hybridization phenomenon. Particularly, the three-step Scheme IV corresponding to the repetitive c-ssDNA along with Eq. (13) can successfully explain how the overall hybridization rate scales with the viscosity of the reaction medium and relative levels of lengths and sequence complexities of c-ssDNAs (particularly in the limit as the level of sequence complexity approaches the length) where most of the earlier theoretical models and computational studies failed.

## B. Assumptions and limitations

Although the presented lattice model could recover several kinetic scaling laws associated with the DNA renaturation dynamics, there are several assumptions and limitations, *viz.*, (1) the rigidities of both c-ssDNAs and dsDNA are ignored; (2) the same intermonomer distance for both c-ssDNA and dsDNA strands was assumed; (3) detailed base-pairing and base stacking interactions were ignored; (4) formation of intrastrand loops was ignored; (5) GC base compositions of the c-ssDNA were not considered; and (6) there is no direct computational or experimental evidence for the entire connected Scheme IV.

Translation diffusion of c-ssDNAs and their interpenetration upon collision mainly decide the kinetic scaling laws rather than the fine details of base-pairing and base-stacking interactions. Therefore, these factors will not affect the main scaling results much. Variation in the GC composition of c-ssDNAs will influence the microscopic zipping rates  $k_+$  and  $k_-$  and subsequently the overall zipping rate  $k_z$ . However, this will not affect the scaling of the zipping rate

with the length of c-ssDNAs. Variation of the intermonomer distances of c-ssDNAs will not affect the obtained scaling laws since we measure the lengths in terms of number of monomer units (nt). For example, we consider only the number of nucleotides scanned by the c-ssDNAs over slithering events rather than the number of intermolecular distances. The rigidity of both c-ssDNA and dsDNA will be mainly compensated by the conformational fluctuations driven by the chain entropy [49]. Formation of intrastrand loop structures will influence our results in two ways: (1) significant fraction of c-ssDNA segments will not be exposed for the exploration of correct contacts, which in turn increases the time required for nucleation and (2) upon forming correct contacts at the nonloop regions, zipping will be hindered by the presence of intrastrand loops, which in turn increases the overall zipping time. This means that additional time components will be added up to the overall nucleation and zipping times, which will not change the overall kinetic scaling laws much. Although there is no direct computational or experimental evidence for the entire connected Scheme IV, recent simulation studies have revealed the presence of its various components, *viz.*, incorrect and trap correct contact formation, slithering, nucleation, and zipping steps.

## V. CONCLUSION

We have developed a lattice model on the rate of hybridization of the complementary ssDNA (c-ssDNA) strands. These c-ssDNAs can be thought as loosely packed and spherical-shaped nucleotide clusters, and the collisions between these clusters are mediated via the 3D translational diffusion. Upon each collision, the base clusters of c-ssDNAs interpenetrate each other to form three different types of contacts among them: correct, incorrect, and trap correct contacts. Correct contacts are those with exact registry matches which can lead to nucleation and zipping of c-ssDNAs. Incorrect contacts are the mismatch contacts which are less stable compared to the trap correct contacts, which can occur in the repetitive c-ssDNAs. Trap correct contacts possess exact registry match within the repeats. However, they are incorrect contacts in the view of the whole c-ssDNAs.

Trap correct contacts can form partial duplexes with single-stranded overhangs. The nucleation rate ( $k_N$ ) is directly proportional to the average number of correct contacts ( $\langle n_{cc} \rangle$ ) formed when c-ssDNAs interpenetrate each other inside the reaction volume. The c-ssDNAs reach the reaction volume via translational diffusion with rate  $k_t$ . This rate will be independent of the length of c-ssDNAs when both c-ssDNAs are equal in size. As a result, the nucleation rate will be directly

proportional to  $k_t$  times the number of correct contacts. For short c-ssDNAs, the zipping times will much lower than the nucleation times. In such conditions, the overall renaturation rate ( $k_R$ ) will be directly proportional to  $k_N$  and  $L$  as  $k_R = k_N L \propto k_t \langle n_{cc} \rangle L$ , which is the 3D diffusion model.

To understand the scaling properties of the average number of various types of contacts with respect to the length ( $L$ ) of c-ssDNAs, sequence complexity ( $c$ ), and reaction volume ( $V$ ), we modeled the c-ssDNAs as a pair of self-avoiding random walks confined in a cubic lattice box which resembles the reaction volume ( $V$ ). Our lattice model simulations suggested the scaling  $\langle n_{cc} \rangle \cong \frac{vL}{V}$ ,  $\langle n_{icc} \rangle \cong \frac{vL^2}{cV}$ , and  $\langle n_{icc} \rangle \cong \frac{vL^2}{V}$  where  $v = 1$  is the monomer volume,  $\langle n_Q \rangle$  are the average number of correct ( $Q = cc$ ), incorrect ( $Q = icc$ ), and trap correct contacts ( $Q = tcc$ ). Further numerical analysis and nonlinear least-square fitting results suggested the scaling for the average radius of gyration of c-ssDNAs with their length as  $R_g \propto \sqrt{L}$ . Since the reaction space will be approximately a sphere with radius equals to  $2R_g$ , one obtains the scaling for the nucleation rate with length of c-ssDNA as  $k_N \propto \frac{1}{\sqrt{L}}$ , and one finally obtains  $k_R \propto \sqrt{L}$  in line with the experimental observations. However, this expression works only for a non-repetitive and short c-ssDNAs. When c-ssDNAs are repetitive with a complexity of  $c < L$ , earlier models suggested the scaling  $k_R \propto \frac{\sqrt{L}}{c}$ . This scaling will break down when  $c = L$ . These observations clearly suggested the existence of at least two different pathways of renaturation, *viz.*, through incorrect contact and trap correct contact.

The trap correct contacts occurring between repetitive c-ssDNAs can lead to the formation of partial duplexes with single-strand overhangs. These partial duplexes keep the c-ssDNAs in close proximity for prolonged timescales, which is essential for the extended 1D slithering that can speed up the searching for the correct contacts. Clearly, the extent of slithering dynamics will be inversely proportional to the sequence complexity  $c$ . When the complexity is close to the length of c-ssDNAs, the pathway through incorrect contact will be the dominating one with minimal level of slithering. When the complexity is much less than the length of c-ssDNAs, the pathway via the trap correct contact will be the dominating one.

## ACKNOWLEDGMENT

The funding was provided by Science and Engineering Research Board (Grants No. CRG/2019/001208 and No. MTR/2019/00002).

The author declares no conflict of interest.

- 
- [1] B. Alberts, *Molecular Biology of the Cell* (Garland Science, New York, 2002).
  - [2] B. Lewin, J. E. Krebs, S. T. Kilpatrick, E. S. Goldstein, and B. Lewin, *Lewin's Genes X* (Jones and Bartlett, Sudbury, MA, 2011).
  - [3] E. L. Duggan, *Biochem. Biophys. Res. Commun.* **6**, 93 (1961).
  - [4] L. F. Cavalieri, T. Small, and N. Sarkar, *Biophys. J.* **2**, 339 (1962).
  - [5] T. Maniatis, E. F. Fritsch, and J. Sambrook, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1982).
  - [6] S. Woo and P. W. Rothemund, *Nat. Commun.* **5**, 4889 (2014).
  - [7] H. T. Maune, S. P. Han, R. D. Barish, M. Bockrath, W. A. Goddard III, P. W. Rothemund, and E. Winfree, *Nat. Nanotechnol.* **5**, 61 (2010).
  - [8] J. G. Wetmur and N. Davidson, *J. Mol. Biol.* **31**, 349 (1968).

- [9] D. C. Rau and L. C. Klotz, *Biophys. Chem.* **8**, 41 (1978).
- [10] M. D. Chilton, *Nat. New Biol.* **246**, 16 (1973).
- [11] R. Murugan, *Biophys. Chem.* **104**, 535 (2003).
- [12] E. J. Sambriski, D. C. Schwartz, and J. J. de Pablo, *Biophys. J.* **96**, 1675 (2009).
- [13] E. J. Sambriski, V. Ortiz, and J. J. de Pablo, *J. Phys. Condens. Matter* **21**, 034105 (2009).
- [14] F. W. Studier, *J. Mol. Biol.* **41**, 199 (1969).
- [15] G. Niranjani and R. Murugan, *PLoS ONE* **11**, e0153172 (2016).
- [16] G. A. Galau, R. J. Britten, and E. H. Davidson, *Proc. Natl. Acad. Sci. USA* **74**, 1020 (1977).
- [17] M. J. Smith, R. J. Britten, and E. H. Davidson, *Proc. Natl. Acad. Sci. USA* **72**, 4805 (1975).
- [18] J. A. Subirana, *Biopolymers* **4**, 189 (1966).
- [19] J. A. Subirana and P. Doty, *Biopolymers* **4**, 171 (1966).
- [20] K. J. Thrower and A. R. Peacocke, *Biochim. Biophys. Acta* **119**, 652 (1966).
- [21] J. G. Wetmur, *Biopolymers* **10**, 601 (1971).
- [22] J. C. Araque and M. A. Robert, *J. Chem. Phys.* **144**, 125101 (2016).
- [23] P. J. Sanstead and A. Tokmakoff, *J. Chem. Phys.* **150**, 185104 (2019).
- [24] J. L. Sikorav, H. Orland, and A. Braslau, *J. Phys. Chem. B* **113**, 3715 (2009).
- [25] T. E. Ouldridge, P. Sulc, F. Romano, J. P. Doye, and A. A. Louis, *Nucleic Acids Res.* **41**, 8886 (2013).
- [26] A. Ferrantini, M. Baiesi, and E. Carlon, *J. Stat. Mech. Theory Exp.* (2010) P03017.
- [27] A. Ferrantini and E. Carlon, *J. Stat. Mech. Theory Exp.* (2011) P02020.
- [28] A. Cumberworth, A. Reinhardt, and D. Frenkel, *J. Chem. Phys.* **149**, 234905 (2018).
- [29] Z. Qu, Y. Zhang, Z. Dai, Y. Hao, Y. Zhang, J. Shen, F. Wang, Q. Li, C. Fan, and X. Liu, *Angew. Chem. Int. Ed.* **60**, 16693 (2021).
- [30] B. E. K. Snodin, F. Romano, L. Rovigatti, T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, *ACS Nano* **10**, 1724 (2016).
- [31] J. S. Schreck, T. E. Ouldridge, F. Romano, P. Šulc, L. P. Shaw, A. A. Louis, and J. P. K. Doye, *Nucleic Acids Res.* **43**, 6181 (2015).
- [32] N. M. Gravina, J. C. Gumbart, and H. D. Kim, *J. Phys. Chem. B* **125**, 4016 (2021).
- [33] M. S. Jones, B. Ashwood, A. Tokmakoff, and A. L. Ferguson, *J. Am. Chem. Soc.* **143**, 17395 (2021).
- [34] H. Sidky *et al.*, *J. Chem. Phys.* **148**, 044104 (2018).
- [35] P. W. Atkins, *Physical Chemistry* (W. H. Freeman, San Francisco, 1978).
- [36] P. Debye, *Trans. Electrochem. Soc.* **82**, 265 (1942).
- [37] G. Niranjani and R. Murugan, *Phys. Biol.* **13**, 046003 (2016).
- [38] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics* (Clarendon Press, Oxford, 1988).
- [39] P.-G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, NY, 1979).
- [40] D. W. Marquardt, *J. Soc. Ind. Appl. Math.* **11**, 431 (1963).
- [41] G. Niranjani and R. Murugan, *J. Stat. Mech. Theory Exp.* (2016) 053501.
- [42] M. N. Rosenbluth and A. W. Rosenbluth, *J. Chem. Phys.* **23**, 356 (1955).
- [43] P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, NY, 1953).
- [44] R. Masoud, R. Tsukanov, T. E. Tomov, N. Plavner, M. Liber, and E. Nir, *ACS Nano* **6**, 6272 (2012).
- [45] J.-S. Shin and N. A. Pierce, *J. Am. Chem. Soc.* **126**, 10834 (2004).
- [46] F. C. Simmel, *ChemPhysChem* **10**, 2593 (2009).
- [47] T. E. Ouldridge, R. L. Hoare, A. A. Louis, J. P. K. Doye, J. Bath, and A. J. Turberfield, *ACS Nano* **7**, 2479 (2013).
- [48] P. Yin, H. Yan, X. G. Daniell, A. J. Turberfield, and J. H. Reif, *Angew. Chem. Int. Ed.* **43**, 4906 (2004).
- [49] R. Murugan, *J. Stat. Mech. Theory Exp.* (2020) 013501.