DNA unzipping phase diagram calculated via replica theory

C. Brian Roland
Chemical Physics Program, Harvard University, Cambridge, Massachusetts 02138, USA

Kristi Adamson Hatch
Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA

Mara Prentiss
Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA

Eugene I. Shakhnovich
Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA

(Received 1 April 2007; revised manuscript received 7 December 2008; published 28 May 2009)

We show how single-molecule unzipping experiments can provide strong evidence that the zero-force melting transition of long molecules of natural dsDNA should be classified as a phase transition of the higher-order type (continuous). Toward this end, we study a statistical-mechanics model for the fluctuating structure of a long molecule of dsDNA, and compute the equilibrium phase diagram for the experiment in which the molecule is unzipped under applied force. We consider a perfect-matching dsDNA model, in which the loops are volume-excluding chains with arbitrary loop exponent c. We include stacking interactions, hydrogen bonds, and main-chain entropy. We include sequence heterogeneity at the level of random sequences; in particular, there is no correlation in the base-pairing (bp) energy from one sequence position to the next. We present heuristic arguments to demonstrate that the low-temperature macrostate does not exhibit degenerate ergodicity breaking. We use this claim to understand the results of our replica-theoretic calculation of the equilibrium properties of the system. As a function of temperature, we obtain the minimal force at which the molecule separates completely. This critical-force curve is a line in the temperature-force phase diagram that marks the regions where the molecule exists primarily as a double helix versus the region where the molecule exists as two separate strands. We compare our random-sequence model to magnetic tweezer experiments performed on the 48 502 bp genome of bacteriophage λ. We find good agreement with the experimental data, which is restricted to temperatures between 24 and 50 °C. At higher temperatures, the critical-force curve of our random-sequence model is very different for that of the homogeneous-sequence version of our model. For both sequence models, the critical force falls to zero at the melting temperature T_c like $|T-T_c|^{\alpha}$. For the homogeneous-sequence model, $\alpha=1/2$ almost exactly, while for the random-sequence model, $\alpha\approx 0.9$. Importantly, the shape of the critical-force curve is connected, via our theory, to the manner in which the helix fraction falls to zero at T_c. The helix fraction is the property that is used to classify the melting transition as a type of phase transition. In our calculation, the shape of the critical-force curve holds strong evidence that the zero-force melting transition of long natural dsDNA should be classified as a higher-order (continuous) phase transition. Specifically, the order is 3rd or greater.

DOI: 10.1103/PhysRevE.79.051923 PACS number(s): 87.14.G−, 82.35.Gh, 64.60.F−, 64.60.Cn

I. INTRODUCTION

Despite the simplicity of Watson and Crick's double-helical structure of crystallized dsDNA [1], when long dsDNA is in contact with a constant-temperature fluid, as is genomic DNA in a cell, spontaneous thermal fluctuations can cause a small region of the double helix to unwind and separate (denature). Clearly, the prevalence of locally denatured regions bears on questions regarding the mechanisms by which a cell's proteins achieve the maintenance of, and data acquisition from, the genome. Thus, since the 1950s, it has been the objective of both experimental and theoretical endeavors to understand how the physical structure of DNA constrains the space of possible biological activities. One step toward this objective is to understand the manner in which dsDNA's physical structure changes as temperature is varied.

A standard practice for determining the prevalence of locally denatured regions (loops) is by spectroscopically measuring the helical content of a solution of DNA molecules, and comparing the temperature dependence of the reading with the helix fraction [fraction of base pairs (bps) involved in complementary base pairing] predicted by models. There is evidence from bulk spectroscopy measurements that as the temperature is increased over an interval of 10–15 °C, long dsDNA melts in a sequence of local unwinding events [2,3]. Recently, new experimental methods have been developed that complement this spectroscopic information.

Single-molecule force experiments have been performed in which a polynucleotide double helix is "unzipped." In this experimental configuration, a dsDNA molecule or an RNA hairpin is pulled apart by an apparatus that exerts a constant force that tends to separate the base pair at one end of the double helix [4,5]. Danilowicz *et al.* experimentally deter-

mined the minimal force, termed the critical force, to completely separate the strands of a 1500 bp subsequence of the 48 502 bp genome of the bacteriophage λ. The critical force is a measure of the stability of the duplex state as compared to the strand-separated state. This critical force was determined as a function of temperature, and the resulting critical-force curve is a line in a temperature-force phase diagram that marks the regions where the molecule exists primarily as a double helix versus the region where the molecule exists as two separate strands. We study a model for unzipping the entire genome of phage λ (not just the 1500 bp subsequence), and show how the measurement of the critical-force curve in the neighborhood of the melting temperature will allow us to classify the genome's melting transition as a type of phase transition (see Appendix, Sec. 1 for the classification of phase transitions).

In a fundamental work toward the classification of the melting transition of long dsDNA, Poland and Scheraga considered the perfect-matching (PM) model—loops are allowed, but base pairing can only occur between bps that are paired in the native state—with a homogeneous sequence [6]. A loop of length ℓ has a free energy that increases (destabilizes) like $c \ln \ell$, resulting in a "long-range interaction" between the helix-state bps flanking the loop-state bps; c is a parameter termed the loop exponent that depends upon the model for the structure of a loop. Despite the approximate one dimensionality of dsDNA, they showed that the long-range interaction causes the melting transition to be a true phase transition, and the order depends upon the parameter c. Because this work used analytical methods that employed the infinite-chain limit, it produced precise and succinct statements about the order of the phase transition (a phase transition is defined only for infinite systems [7]), in the homogeneous-sequence approximation. But, the homogeneous PM model fails to describe long molecules of natural dsDNA, e.g., the genome of phage λ, because the width of the model's transition is much smaller than observed in spectroscopic studies [7]. Thus, we turn to models in which sequence heterogeneity is incorporated with random sequences.

But is the long-range interaction present in the loop entropy considered by Poland and Scheraga a necessary component of a successful random-sequence model? There have been many investigations of random-sequence one-dimensional nearest-neighbor (NN) Ising models for dsDNA [8–11], which are loopless approximations to the PM model. But every heterogeneous one-dimensional (1D) NN Ising model is less cooperative than the homogeneous version, which does not have a true phase transition [12]. Thus every heterogeneous 1D NN model differs qualitatively from the homogeneous PM model. So, we cannot say, *a priori*, if a random-sequence NN Ising model is a safe approximation to the random PM model, which includes loops. Therefore, as a model for the 48 502 bp genome of phage λ, we study the random-sequence PM model in the long-chain limit, and attempt to determine if this model melts with a true phase transition, and thus resembles more closely the homogeneous PM model or the random NN Ising model. If the random PM model does have a true phase transition, we attempt to classify it by the order.

Our model includes (a) perfect-matching, (b) an uncorrelated random sequence of base-pairing free energies, (c) loops with arbitrary loop exponent c, (d) end sequence that couples to external unzipping force, and (e) long-chain limit. The self-consistency of the perfect-matching and long-chain approximations is discussed in Appendix, Sec. 7. We give a heuristic argument that the low-temperature macrostate does not exhibit degenerate ergodicity breaking. We use the no ergodicity-breaking statement to interpret the results of our replica calculation, which give the model's thermodynamics for all experimentally relevant temperatures. We find that the presence of sequence heterogeneity changes the order of the melting transition from first order to higher order (continuous).

The organization of the paper is as follows. In Sec. II, we describe our microscopic mechanical model and how we model the sequence of the λ genome as random. In Sec. III, we present a version of the replica method that is amenable to our model. In Sec. IV, we present the numerical solutions to the equations obtained from the replica method; these solutions give the temperature-force phase diagram. We also present the methods and results of zero-force melting experiments and unzipping experiments that provide a consistency check for the results of our theory. In Sec. V, we show that sequence heterogeneity changes the order of the melting transition from first order to higher order, when the loop exponent $c=2.115$. Also, we show how the shape of the critical-force curve near the melting temperature is determined by an effective value of the loop exponent c^{ef}, which evidences the order of the melting transition as 3rd or greater.

II. MODEL

A. Mechanical model

1. Sequence partition functions

In our minimal model of dsDNA unzipping, we retain only those features of molecular structure necessary to analyze the interplay between base-pairing heterogeneity and thermally induced local melting (loops), and whether this interplay has consequences for the temperature-force phase diagram. First, we choose a partitioning of the strands into "nucleotides" that is convenient to describe the physics of unzipping, as shown in Fig. 1(a). We label the "nucleotides" with sequence position t, running from 1 to N. Next we assume that each complementary bp may exist in one of two microstates: the helix state or the coil state. The helix state of bp t is defined by the activation of the complementary hydrogen bonds (h-bonds), and to achieve this, the bases are close and properly oriented [see Fig. 1(b)]. We assume that if bps t and $t-1$ are both in the helix state, then for each strand, the main-chain bonds linking the deoxyribose of bp t to the deoxyribose of bp $t-1$ are frozen into a special rotation state. We ignore the rotation states provided by the bond connecting deoxyribose to the nitrogenous base, and lump this entropy into that of the main-chain bonds. The coil state of bp t is defined by the deactivation of the h-bond, and thus the bases are far apart or misoriented. Thus, if bp t or $t-1$ are in

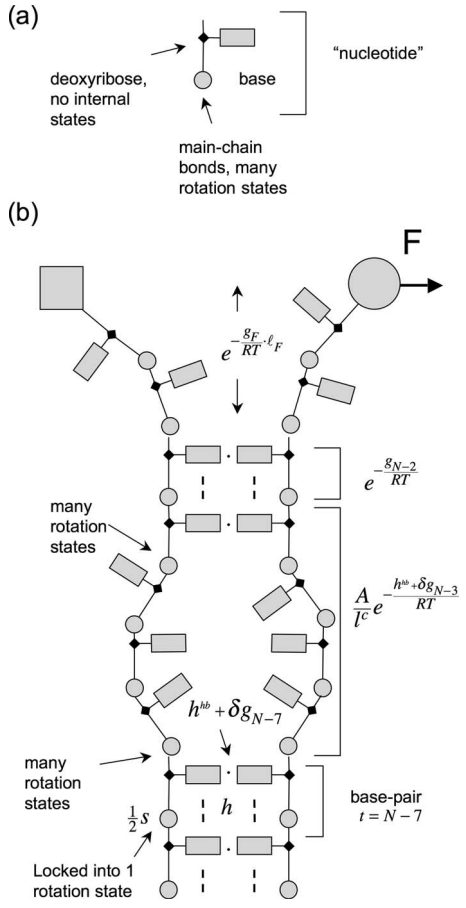

FIG. 1. Our model for dsDNA under unzipping force. (a) We choose a partitioning of ssDNA into "nucleotides" that is convenient to describe unzipping. (b) Base-pair $t=N-7$ is in the helix state, as is the preceding bp, so bp $N-7$ contributes entropy s due to the loss of rotation states of the main-chain bonds, enthalpy h due to stacking with the previous bp, and free energy $h^{\mathrm{hb}}+\delta g_t$ due to hydrogen bonding. Bp $t=N-7$ is followed by a bound sequence of length $\ell=4$, which is followed by a bound sequence of length one, which is followed by the unbound sequence, having length $\ell_F=2$. The big square represents the glass surface (a capillary tube [4]) to which the DNA is attached; the square is fixed in space. The large circle represents the bead, which is pulled by the magnetic field with constant force in a direction perpendicular to the glass surface.

the coil state, then the main-chain bonds of bp t are free to sample many rotation states. In the experiment [4], a hairpin is synthetically added to the not-unzipped end of the double helix; this is modeled by imagining a fictitious bp at sequence position $t=0$ that remains forever in the helix state.

From the definitions of single-bp microstates, we may state our algorithm for assembling the statistical weight to observe a particular microstate of the entire molecule, when at equilibrium. To do this, we apply a physical approximation that is common in the statistical mechanics of polynucleic acids [7]: in any given microstate of the molecule, the only bps that interact with a given bp are those in its mechanical sequence.

We define two types of mechanical sequences. A *bound sequence* is a contiguous subchain of bps such that all but the last (according to sequence position t) are in the coil state; the last bp is in the helix state [see Fig. 1(b)]. The bp preceding (in sequence position) the bound sequence is in the helix state, and a bound sequence of length one is a single helix-state bp. For any given microstate of the model, all bps up to and including the last helix-state bp may be assigned to a bound sequence. All bps following the last helix-state bp belong to the second type of mechanical sequence, which is called the *unbound sequence*, in which all bps are in the coil state, and experience the stability offered by the external force F. In the statistical mechanics of polynucleic acids, there are two distinct schemes for partitioning the bound section of the model into mechanical sequences. We follow the scheme of Gibbs and DiMarzio [13], where the bound section of the model is made up of a series of bound sequences. Alternatively, the bound section can be described as an alternating series of *helix* and *loop* sequences, as in the manner of Hill [14]. For mathematical convenience, we use the scheme of Gibbs and DiMarzio.

According to our physical approximation, the statistical weight to observe any given polymer microstate is a product of the statistical weight factors (partition functions) for each mechanical sequence present in that microstate. The statistical weight factor (partition function) for a bound sequence of length one (a single helix-state bp preceded by a helix-state bp) is

$$e^{-g_t/RT}, \tag{1}$$

where

$$g_t = h - Ts + h^{\mathrm{hb}} + \delta g_t \tag{2}$$

is called the *helix-propagation free energy* and is the energy to transfer a bp at sequence position t from the coil to the helix state, given that bp $t-1$ is in the helix state [see Fig. 1(b)]. The average of g_t, over all bps in the entire molecule is

$$h - Ts + h^{\mathrm{hb}} = \frac{1}{N}\sum_{t=1}^{N} g_t, \tag{3}$$

where $h+h^{\mathrm{hb}}$ and s are the enthalpic and entropic components of this average. The parameter h^{hb} gives the enthalpy of hydrogen-bond formation; the entropy change due to hydrogen-bond formation is neglected because it is small [15]. For real dsDNA, the parameters h and s may each have contributions from stacking interactions and main-chain entropy, but in our model, we associate h only with the stacking interaction between bps t and $t-1$, and associate s only with the main-chain bonds of bp t [see Fig. 1(b)]. For real DNA, the sequence-dependent term δg_t may have contributions from hydrogen bonding, stacking interactions, and main-chain entropy, but in our model we associate δg_t only with hydrogen bonding.

The statistical weight (partition function) for a bound sequence with $\ell \geq 2$ bps, terminated at sequence position t, is

$$\frac{A}{\ell^c} e^{-(h^{\mathrm{hb}}+\delta g_t)/RT}, \tag{4}$$

where the parameters have the following thermodynamic meaning. Suppose we have some sequence of ℓ coil-state bps, preceded by a bp in the helix state, and followed by no

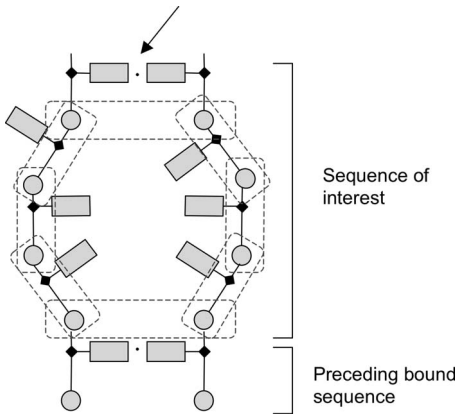

FIG. 2. Thermodynamics of a loop, i.e., a bound sequence with $\ell \geq 2$ bps. For the sequence of interest, $\ell=4$. Here, we imagine that no bps follow the sequence of interest. If the terminal bp in the sequence is removed from the helix state (breaking the indicated h-bond), the nucleotides in both strands have the maximal number of rotation states; we call this the free coil macrostate. This macrostate has a free energy, purely entropic, that we set to be the zero of free energy (defined for $\ell \to \infty$). If the terminal bp is constrained to the helix state, then the sequence of bps forms a loop in which the individual monomers are indicated with boxes. The number of boxes is twice the number of bps (2ℓ) in the sequence of interest. As ℓ becomes large, the ring of boxes sample a number of configurations (relative to the number of configurations in the free coil state) that is approximated by the scaling law A/ℓ^c. The T-independent prefactor A is associated with the configuration-space volume of the helix state (arrow) of the terminal bp.

other base pairs (see Fig. 2). The macrostate of the sequence—the free-coil macrostate—has a T-independent free energy (per bp) that we set to zero, i.e., we measure all free energies with respect to the free coil macrostate (here, we imagine $\ell \to \infty$). If we constrain the terminal bp in the sequence to be in the helix state, we have formed a loop and this action will raise the free energy of the sequence above the zero of free energy. The factor A/ℓ^c gives the entropic cost to limit the rotation states of the "nucleotides" of the strands so that they may form a loop. The factor $\exp[-(h^{\mathrm{hb}}+\delta g_t)/RT]$ represents the free energetic gain/loss upon transferring the terminal bp from the coil to the helix state (the hydrogen bond is activated), given that bp $t-1$ is in the coil state.

In the solution of our model, we found that the manner in which we distributed the average hydrogen-bonding enthalpy between the terms h and h^{hb} had little impact on the numeric results. For convenience, we set $h^{\mathrm{hb}}=0$, absorbing all of the hydrogen-bonding interaction into h. Nonetheless, we retain h^{hb} in our notation because its variation mathematically identifies the fraction of base pairs in the helix state.

The T-independent prefactor A is associated with the configuration-space volume of the helix state of the terminal bp. This configuration-space volume is achieved by restricting the main-chain bonds of all "nucleotides" in the bound sequence, but the associated entropy is numerically different from the entropy s in Eq. (3), as there is no helix-state bp preceding the terminal bp. Our notation A emphasizes a purely entropic origin, and is consistent with [6,16,17]. The parameter A is given the name (constant) in Section IXD of [7]. Because A plays a role similar to the cooperativity parameter σ in models of polyalanine [18], it is often called σ_c, as in [19], or σ as in [20]; note that in polyalanine models, σ has an enthalpic origin.

The parameter c governs the manner in which the number of configurations—relative to the coil state—of a loop decreases as the length of the loop increases. Fisher showed that if we model the loop as a volume-excluding polymer, but one that is isolated from other portions of the chain, then the appropriate value is $c=1.75$ [21]. Kafri *et al.* [16] showed that if we include interactions with the other portions of the chain outside the loop, the appropriate value is at least $c=2.115$. Recently, Blossey *et al.* remarked that the Kafri model may be inappropriate for real dsDNA because the difference in the persistence lengths of double-stranded (in helix regions) and single-stranded DNA (in loops) was ignored [20]. For this reason, we consider both the Fisher and Kafri values, $c=1.75$ and $c=2.115$, respectively. Employing the numerical algorithm MELTSIM of Blake *et al.* [22], Blossey *et al.* fit the parameter A (in those works it is called σ) to each value of c: for $c=2.115$, $A=1.75\times10^{-4}$, and for $c=1.75$, $A=1.26\times10^{-5}$ (see Table I).

The statistical factor (partition function) for an unbound sequence of ℓ_F bps is

TABLE I. Microscopic parameters for our model. Notice that all parameter values are obtained from studies external to ours. Thus, our model contains zero adjustable parameters. The helix energy parameters h and s are obtained by averaging the Blake-Delcourt parameters over the sequence of the λ-phage genome, see Sec. II B. The standard deviation of the helix-propagation free energy, $g_{\lambda}^{(2)}$, is calculated as a function of temperature [see Fig. 5(b)], and so is not given here.

Parameter	Value		Phenomena		Literature source
			Interactions	No interactions	
c, loop exponent	2.115	1.75	between loops	between loops	Kafri *et al.* [16] Fisher [21]
A, loop termination factor	1.75×10^{-4}	1.26×10^{-5}	Limited movement of helix-state bp at end of loop		Blossey and Carlon [20]
h, helix enthalpy	−9.404 kcal/mol		Favorable stacking of nearest-neighbor bases		Blake and Delcourt [19]
s, helix entropy	−25.87 cal/mol·K		Unfavorable constraints on the rotation states of main-chain bonds		Blake and Delcourt [19]
a, Kuhn length of ssDNA	1.9 nm		Characterizes the flexibility of the ssDNA		Danilowicz *et al.* [4]
b, length of a base in ssDNA	0.7 nm		Estimate for contour length of a single monomer base in ssDNA		Smith *et al.* [4,49]

$$e^{-g_F \ell_F / RT} \tag{5}$$

and g_F is the free energy per bp in this sequence. The statistical factor in Eq. (5) factorizes (exactly) by bp because we model the unbound sequence as a freely jointed-chain under stretching force [23], obtaining

$$g_F(T,F) = -\frac{a}{b} RT \log\left[\frac{k_B T}{Fb} \sinh\left(\frac{Fb}{k_B T}\right)\right], \tag{6}$$

where F is the applied force in pN, T is the temperature in Kelvin, and $b=1.9$ nm is the Kuhn length of ssDNA in the solvent conditions of the experiment [4]. Because g_F is the free energy per bp, we include the prefactor a/b, where $a=0.7$ nm is the length of a "nucleotide" in ssDNA (see Table I). The parameter k_B is Boltzmann's constant and the appearance of the gas constant R means that g_F is in units of per mol of bps.

2. Polymer in one dimension

With the sequence partition functions described above, we can give the statistical factor for any given microstate. But it proves convenient to describe the equilibrium statistics of our mechanical model in terms of the following Hamiltonian:

$$H_g[\rho] = \sum_{i=1}^{N} u_A(\rho_i, \rho_{i-1}) + \sum_{i=1}^{N} u_g(\rho_i, \rho_{i-1}) + \sum_{i=1}^{N} [h^{\mathrm{hb}}_{\rho_i} + \delta g_{\rho_i}] + g_F \cdot (N - \rho_N), \tag{7}$$

which describes the fictitious one-dimensional polymer shown in Fig. 3; Equation (7) can be derived from the sequence partition function description. Hamiltonian $H_g[\rho]$ represents the following approach: instead of describing a microstate of our DNA model by stating that each bp is in the helix or coil state, we instead give the DNA sequence locations where the bp is in the helix state. This collection of sequence locations has the statistics of the monomers of a polymer with Hamiltonian $H_g[\rho]$. The monomers of the 1D polymer are labeled $i \in \{0,\ldots,N\}$, each having location variable $\rho_i \in \{0,\ldots,N\}$. Degree of freedom ρ_0 is fixed at 0, while for $i \in \{1,\ldots,N\}$, ρ_i is a fluctuating variable (of the annealing type). For each monomer i, its location ρ_i is confined to the set of locations $\{0,\ldots,N\}$ by walls at locations $t=0$ and $t=N$. The initial configuration of the polymer is any arrangement satisfying $\rho_{i-1} \leq \rho_i$ for each $i \in \{1,\ldots,N\}$.

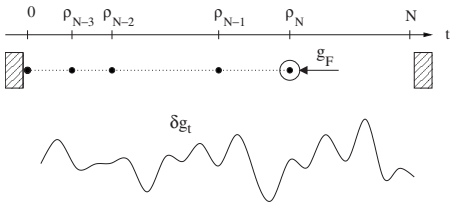

FIG. 3. 1-dimensional polymer representation of our model. There is a 1-to-1 correspondence between each microstate of this fictitious one-dimensional polymer and a microstate of the DNA model described in terms of the helix/coil state of each bp (see Sec. II A 1). In the microstate shown, monomers $\{N-3,\ldots,N\}$ represent the DNA sequence locations at which the bps are in the helix state, and the dot at $t=0$ indicates that monomers $\{0,\ldots,N-4\}$ all reside at location 0. Each monomer i, that is at a sequence location $t>0$, interacts with monomer $i-1$ with pair potentials u_A and u_g. Each monomer interacts with the external potential field δg_t; we show a cartoon of a random realization of this potential field.

The polymer connectivity is described with a pair interaction between neighboring monomers along the polymer chain,

$$u_A(\rho_i, \rho_{i-1}) = \begin{cases} 0 & \rho_i = 0 \\ \Lambda & \rho_i > 0 \text{ and } |\rho_i - \rho_{i-1}| = 0 \\ 0 & \rho_i > 0 \text{ and } |\rho_i - \rho_{i-1}| = 1 \\ RT \log A^{-1} + cRT \log|\rho_i - \rho_{i-1}| & \rho_i > 0 \text{ and } |\rho_i - \rho_{i-1}| \geq 2, \end{cases} \tag{8}$$

where the first line of Eq. (8) indicates that when a monomer is at location $t=0$, the interaction with the previous monomer is turned off. When a monomer is at any $t>0$, it experiences a strong repulsion with the previous monomer at short separations and a weak attraction at large separations (see Fig. 4). The parameter Λ is a large number that makes it very unfavorable to find two monomers at the same location (if that location is not $t=0$). As related to the dsDNA model, the potential u_A represents the effective interaction between the helix-state bps terminating and preceding a bound sequence and is due to the entropy of the intervening coil-state bps.

Two monomers that have adjacent locations share an interaction

$$u_g(\rho, \rho') = \begin{cases} \text{h-Ts} & \rho,\ \rho' > 0 \text{ and } |\rho - \rho'| = 1 \\ 0 & \text{otherwise}. \end{cases} \tag{9}$$

This short-range attractive potential has the same form as that in the one-dimensional lattice gas [24]. Because monomers cannot move past one another, only monomers that are chain neighbors may share this interaction. As related to the dsDNA model, the interaction u_g corresponds to the stacking and main-chain entropy contributions to the helix-propagation free energy. Note that we ignore the helix-

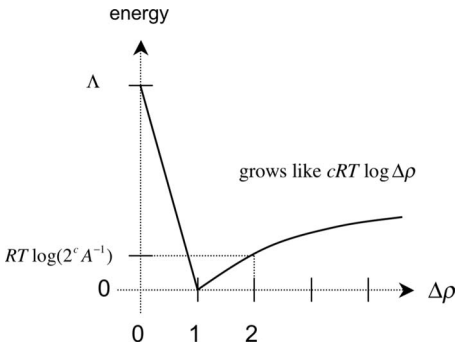

FIG. 4. In the one-dimensional polymer, sequence connectivity can be represented as a pair-potential u_A between neighbors along the chain. Chain-neighbor monomers cannot occupy the same DNA sequence location due to the large repulsive energy Λ. At large separations, a logarithmic attractive potential exists. These terms combine to give a pair potential with short-ranged repulsion and long-ranged attraction, and a minimum at $\Delta\rho=1$. In this figure, the base of the log is e.

propagation energy between monomer i at $t=1$ and monomer $i-1$ at $t=0$, but this interaction is negligible because it makes a nonthermodynamic contribution to the Hamiltonian [see Eq. (7)].

The monomers interact with an external potential field $h_t^{\mathrm{hb}}+\delta g_t$ (sequence-dependent hydrogen bonding in dsDNA model) that depends on position t. At $t=0$, both h_t^{hb} and δg_t vanish identically, whereas for $t>0$, each $h_t^{\mathrm{hb}}=h^{\mathrm{hb}}$ and $\delta g_t \neq 0$ [see Eq. (A4)]. Additionally, monomer N is driven to the left by compressing force g_F, which represents the interaction between the unbound sequence and the unzipping force in the dsDNA model [see Fig. 3].

For this model, the partition function is

$$Z_g(T,F) = Tr_\rho e^{-H_g[\rho]/RT}, \tag{10}$$

where the notation Tr_ρ represents the sum over all arrangements of the monomers that preserves $\rho_{i-1} \leq \rho_i$ for each pair of nearest neighbors.

B. Genetic model

1. Helix-propagation free energies

In addition to the above mechanical model, we must explain our model for the genetic sequence of λ-phage DNA, i.e., we must specify the helix-propagation free energies g_t. We assume that thermodynamic observables of interest can be well approximated by modeling the genetic sequence of λ-phage dsDNA as random. Specifically, we consider a large ensemble of sequences, in which each sequence is generated by choosing each g_t from a Gaussian distribution with mean $g^{(1)}$ and standard deviation $g^{(2)}$. As a result of training these parameters on experiment, described next, $g^{(1)}$ and $g^{(2)}$ depend on temperature T and sodium concentration $[Na^+]$, but we do not explicitly indicate these dependencies in our notation.

For convenience, our mechanical model is written in terms of *helix-propagation* parameters g_t, representing the free energy to constrain bp t to the helix state, given that bp $t-1$ is in the helix state. For each sequence in the sequence ensemble, if each bp is constrained to the helix state, the molecule will have a free energy (per bp)

$$g^{(1)} = \frac{1}{N}\sum_{t=1}^{N} g_t, \tag{11}$$

and the variance (from one sequence position to the next) in free energy (per bp) is

$$[g^{(2)}]^2 = \frac{1}{N}\sum_{t=1}^{N} g_t^2 - [g^{(1)}]^2 = \frac{1}{N}\sum_{t=1}^{N} (\delta g_t)^2. \tag{12}$$

Fluctuations in Σg_t decay like $1/N^{1/2}$; the second equality is obtain from Eqs. (3) and (11). The random variables g_t can be connected to the DNA melting literature by fixing the sequence-ensemble parameters to the statistical parameters of the genome, i.e., $g^{(1)}=g_\lambda^{(1)}(T)$, $g^{(2)}=g_\lambda^{(2)}(T)$, where

$$g_\lambda^{(1)}(T) \equiv \frac{1}{N}\sum_{t=2}^{48\,502} g^{\mathrm{Blake}}(b_{t-1},b_t;T),$$

$$[g_\lambda^{(2)}(T)]^2 \equiv \frac{1}{N}\sum_{t=2}^{48\,502} g^{\mathrm{Blake}}(b_{t-1},b_t;T)^2 - [g_\lambda^{(1)}(T)]^2, \tag{13}$$

where in Eqs. (12) and (13) the index t runs over the genome of λ phage, beginning at the "left" end of the "ℓ strand;" the ℓ strand has a $5'g$, and its left end is the $5'$ end, which turns out to have a higher GC content than the right end (notation from [25]). The parameter $g^{\mathrm{Blake}}(b_{t-1},b_t;T)$ is the contribution due to the dimer $5'-b_{t-1}b_t-3'$, where b_t is the letter of the base at sequence position t, to the total free energy to transfer the genome from the free coil (strands separated) state to the helix macrostate (each bp in the helix state). The parameters $g^{\mathrm{Blake}}(b,b';T)$, one for each type of NN dimer, were determined in the experimental study of Blake and Delcourt [19], by fitting the NN model of Tinoco and Uhlenbeck [26] to spectroscopic melting measurements. We choose the parameters obtained in the Blake-Delcourt study, as opposed to other studies, because the size of the molecules used in this experiment most closely resembles that of the 48 502 bp genome of λ phage. To the entropic component of the Blake-Delcourt parameters, we add a salt dependence using column 2 in Table IIa of [19]. In this paper, we fix $[Na^+]=0.14$ M, as this approximates the ionic condition of phosphate-buffered saline (PBS), the solvent used in the unzipping experiments [4].

Later, we will compute the thermodynamics of each sequence in the ensemble, and average the result over all members in the ensemble; these procedures are integrated into a single analytical process. We assume that the thermodynamics of each sequence self-averages (see [23,27] for discussions of self-averageness) so that the sequence-ensemble average is a good proxy for the thermodynamics of any given random sequence.

Note that in our statistical ensemble of sequences, each sequence exhibits zero correlation between g_t and g_{t-1}. This is an approximation because such correlation is clearly present in λ-phage dsDNA. Suppose that the base at position

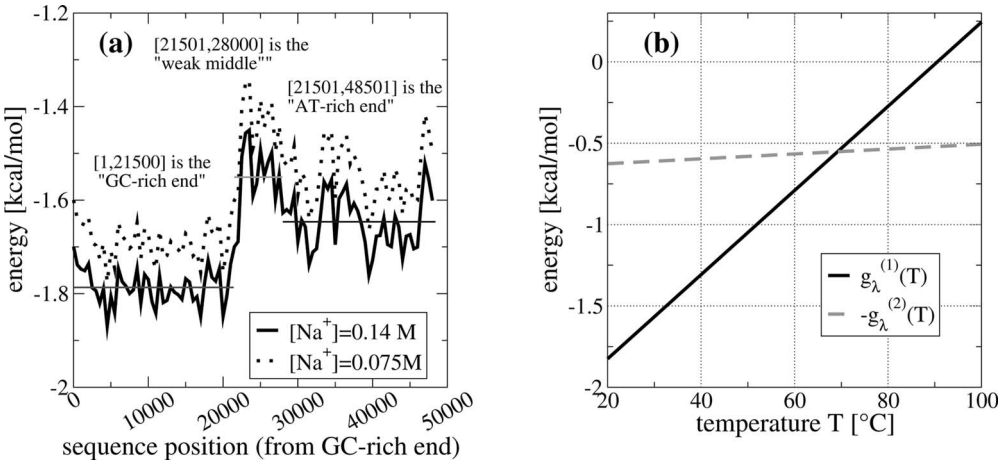

FIG. 5. Statistics of the NN free energies of the λ-phage genome. Unless indicated otherwise, all calculations employ $[Na^+]=0.14M$. (a) The NN energies are averaged over blocks of size 500 bps (black curve) at 25 °C. The average over the subsequence [1,21 500]] is −1.75 kcal/mol, the average over [21 501,28 000] is −1.55 kcal/mol, and the average over [28 001,48 501] is −1.65 kcal/mol (horizontal lines). (b) The mean $g_\lambda^{(1)}$ and standard deviation $g_\lambda^{(2)}$, averaged over the whole genome, of the NN energies are shown at each temperature T. Note that we have plotted the negative of $g_\lambda^{(2)}$. For T above ≈70 °C the magnitude of $g_\lambda^{(2)}$ is comparable to or larger than the magnitude of $g_\lambda^{(1)}$. Thus, at high T, we expect heteropolymeric behavior.

t of the ℓ strand is A. The NN interaction parameter between bp t and $t-1$ will likely be weak, and the interaction parameter for bp $t+1$ and t will also likely be weak. Such correlations are ignored.

2. Analysis of the λ-phage genome

We obtained the sequence of the enterobacteria phage λ genome (NC 001416) from http://www.ncbi.nlm.nih.gov, and computed $g_\lambda^{(1)}$ and $g_\lambda^{(2)}$ as functions of T (Fig. 5). The components of $g_\lambda^{(1)}(T)$ are $h=-9.404$ kcal/mol and $s=-25.87$ cal/mol K ($h^{\mathrm{hb}}=0$) (see Table I).

At this point, we have discussed all microscopic parameters of our model, as summarized in Table I. We emphasize that all microscopic parameters have been trained on experiment, in efforts external to this work. Thus, we do not adjust the values of any microscopic parameters.

C. Course-graining procedure that incorporates loops

To study the long-length-scale properties of our model near the forced unzipping transition, one may imagine course-graining the description of the DNA molecule in Sec. II A 1, or equivalently, the 1D polymer [see Eq. (7)]. We first reduce the description. At a temperature where there is a thermodynamic number of helix-state bps, the system free energy $E(k)$ is computed under the constraint that the unzipping fork, ρ_N, is restricted to interval $k+1$ in a lattice of intervals of size ℓ_{course}, labeled $1,\ldots,M$, where interval 1 is at the end under force. At $F=0$, we imagine turning off the base-pairing potential for every sequence position in intervals $k' \le k-1$, i.e., $g^{(1)}=0$, $\delta g_t=0$ for $t>N-(k-1)\ell_{\mathrm{course}}$, so that the fork naturally resides in interval k. We call $w(k)$ the change in the system free energy, at $F=0$, that results when we turn off the base-pairing potential in interval k given that the potential is off in all intervals k' with $k' \le k-1$ and that the potential is on in all intervals k' with $k' \ge k+1$. Thus, $w(k)$ is the work to transfer interval k from the bound to the free coil state at $F=0$, given that interval $k+1$ is in the bound state while interval $k-1$ is in the free coil state. Then we can write

$$E(k) = \sum_{k'=1}^{k} w(k') + k\ell_{\mathrm{course}} g_F + O(\ell_{\mathrm{course}}), \tag{14}$$

where the term $k\ell_{\mathrm{course}}g_F$ suggests that we do not let the fork position advance beyond interval $k+1$ in response to the unzipping force. The term $O(\ell_{\mathrm{course}})$, of order ℓ_{course}, accounts for the uncertainty in the fork position within interval $k+1$, and thus for the uncertainty in the term $k\ell_{\mathrm{course}}g_F$.

If we call ℓ_{seq} the length scale giving the decay of correlations in the sequence δg_t (in our case, $\ell_{\mathrm{seq}}=0$), and ℓ_{loop} is the length scale giving the decay of loop sizes, then the course-grained description is justified when we can choose some finite ℓ_{course} such that

$$\ell_{\mathrm{seq}}, \ell_{\mathrm{loop}} \ll \ell_{\mathrm{course}} \ll N. \tag{15}$$

We require $\ell_{\mathrm{seq}} \ll \ell_{\mathrm{course}}$ so that the $w(k)$ are not correlated via correlations in the $\delta g(t)$. We require $\ell_{\mathrm{loop}} \ll \ell_{\mathrm{course}}$, so that the $w(k)$ of adjacent intervals have correlations (due to the δg_t) that are weak in the sense that the interaction between neighboring intervals, mediated by the term u_A in Eq. (8), can be made arbitrarily small relative to the magnitude of the $w(k)$. Thus, if Eq. (15) is satisfied, then the free energies $w(k)$ are uncorrelated or weakly correlated. Since we are interested in a long-length-scale description, i.e., large k, we can utilize the central limit theorem to model the $w(k)$ as Gaussian distributed and uncorrelated from each other. This model for the $w(k)$ constitutes a course-graining step in our reduction procedure.

Also, ℓ_{loop} gives the typical length of the free coil portion of the chain in interval $k+1$, so $\ell_{\mathrm{loop}} \ll \ell_{\mathrm{course}}$ guarantee that the fork is indeed in interval $k+1$ at $F=0$.

The term $O(\ell_{\text{course}})$ becomes negligible when the length of the unzipped section is much larger than ℓ_{course}. So, in the system's thermodynamic limit, as the unzipping transition is approached and the length of the unzipped section diverges, we can ignore the $O(\ell_{\text{course}})$ term in our course-grained description, i.e., Eq. (14) with Gaussian and uncorrelated $w(k)$. It is because the length of the unzipped section becomes infinite that we call the location of the forced unzipping transition F_c a critical force. Note that the divergent length scale makes possible the course-graining step.

This course-graining procedure is one way to obtain the (uncorrelated) random zipper model: a zipper model [13,28,29], i.e., one sequence approximation [7], augmented with a random sequence with zero correlations among sequence positions. The (uncorrelated) random zipper model was employed by Lubensky and Nelson in their study of the DNA unzipping transition [30], and also by Kafri and Polkovnikov in connection to directed-polymers-in-random-media models [31]. Note that we claim that the (uncorrelated) random zipper model is justified as a course-grained version of our mechanical and genetic model (see Sec. II B 1) only when (a) the conditions in Eq. (15) are met and (b) the unzipped section becomes very large. We assume that Eq. (15) is satisfied at temperatures exhibiting a thermodynamic number of helix-state bps.

Our analysis is simplified by a consequence of the long-chain approximation, as applied to the unzipping behavior of the random zipper model [30]. For $F<F_c$, the unzipped section is finite, and thus negligible compared to the infinite bound (section) section of the molecule. Also, for $F>F_c$, the bound section is finite. Thus, the random zipper model predicts that the unzipping transition is first order. Because the random zipper model is an exact reduced description of our model in the neighborhood of F_c, the unzipping transition of our model must be first order. In the present work, we seek only to compute the location of F_c; thus our computation will depend only upon statistics of the entire sequence, as opposed to local sequence fluctuations that may occur near the end of the chain that is being unzipped. Consistent with this statement, the previous work finds that the location of F_c depends on sequence only through the value of the bound-state free energy (per base pair), averaged over the entire sequence. Our work goes a step further, in that we compute corrections to the bound-state free energy, as caused by loops. The self-consistency of the long-chain approximation is discussed in Appendix, Sec. 7.

III. METHODS

A. Sequence-averaged free energy of the bound phase

With the mechanical and genetic models in hand, we compute the thermodynamics of a typical random sequence. The replica method [27] offers an analytical route to perform the sequence-ensemble averaging operation.

In the replica method, we compute the sequence-ensemble average of the free energy (per bp) as

$$-RT\frac{1}{N}\overline{\log Z_g(T)} = -RT\lim_{n\to 0}\frac{1}{nN}\log\overline{Z_g^n}, \tag{16}$$

where the overbar denotes averaging over sequences in the ensemble. We first compute, for large integer values of n, the sequence-ensemble average of the partition function of n uncoupled copies (replicas) of the original system to obtain

$$\overline{Z_g^n} = Tr_{\rho^1}\cdots Tr_{\rho^n} e^{-\beta H_n[\rho^1,\ldots,\rho^n]}, \tag{17}$$

where

$$\beta H_n[\rho^1,\ldots,\rho^n] = \sum_{\alpha=1}^{n}\beta H_A[\rho^\alpha] + \beta E[\langle\rho|\rho\rangle] \tag{18}$$

and H_A is a primarily entropic Hamiltonian derived in Appendix, Sec. 2; note that $H_A[\rho^1]=H_A[\rho^1;A,c,h,s]$. The object

$$\beta E[\langle\rho|\rho\rangle] = N\beta(g^{(1)}+h^{\text{hb}})\sum_{\alpha=1}^{n}\langle\rho^\alpha|\rho^\alpha\rangle - N\frac{[\beta g^{(2)}]^2}{2}\sum_{\alpha=1}^{n}\sum_{\gamma=1}^{n}\langle\rho^\alpha|\rho^\gamma\rangle \tag{19}$$

shows that, as a result of the averaging procedure, every pair of replicas (α,γ) now interacts through their overlap, defined as

$$\langle\rho^\alpha|\rho^\gamma\rangle = \begin{array}{c}\text{fraction of sequence-positions}\\ \text{exhibiting a monomer}\\ \text{in both replicas } \alpha \text{ and } \gamma\end{array} \tag{20}$$

and illustrated in Fig. 6. See Appendix, Sec. 2 for derivation of Eqs. (17)–(20).

The partition function $\overline{Z_g^n}$ may be written

$$\overline{Z_g^n} = Tr_Q\, e^{-\beta E_n[Q]+S_n[Q]}, \tag{21}$$

where

$$e^{S_n[Q]} = Tr_{\rho^1}\ldots Tr_{\rho^n} e^{-\beta H_A[\rho^1]}\ldots e^{-\beta H_A[\rho^n]}\delta[Q-\langle\rho|\rho\rangle] \tag{22}$$

and $\delta[Q-\langle\rho|\rho\rangle]$ constrains the n-replica system to a particular realization of the matrix of overlaps $\langle\rho^\alpha|\rho^\gamma\rangle$. In the factor $e^{S_n[Q]}$, we allow the n-replica system to sample all microstates consistent with Q.

At this point, we assume that $\overline{Z_g^n}$ is dominated by a single value of Q (maximum term, i.e., saddle-point approximation) that is obtained by making an ansatz for the form of this matrix, and varying the corresponding parameters. The resulting Q constitutes our approximate solution to Eq. (16). We employ an ansatz for Q that is a restricted version of Parisi's one-step replica-symmetry-breaking (1RSB) scheme

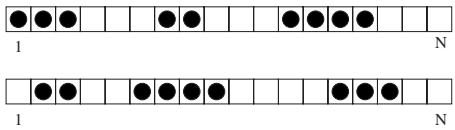

FIG. 6. Overlap between two replicas. Each dot represents the presence of a monomer in the one-dimensional polymer picture. The overlap is the fraction of positions at which both replicas exhibit a monomer; for the configuration shown, the overlap has value 1/3.

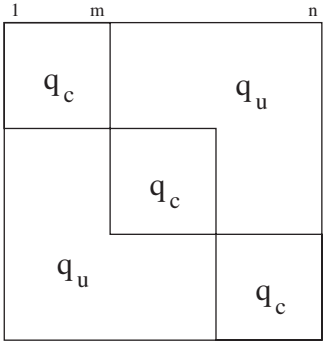

FIG. 7. The cuRSB overlap matrix. In our implementation of the replica method, the replicas are divided into groups of size m. Within a group, the overlap between any two replicas has the maximally correlated value q_c, which is equal to the helix fraction, $q_c=\theta$. The groups fluctuate independently of each other, and the intergroup overlap has the value associated with the maximally uncorrelated value $q_u=\theta^2$.

[27]. Our RSB scheme, which we call cuRSB, for "correlated-uncorrelated" RSB, was successfully applied in a previous theoretical study of a related random heteropolymer (RHP) model [32]. A RHP is a polymer in which the monomers are randomly selected from a variety of types. We restrict the summation in Eq. (21) to a subset of the space of Q matrices, and hope that the restricted sum well approximates the unrestricted sum. Elements, Q^c, of the restricted space have the form shown in Fig. 7 and are achieved by enforcing the cuRSB system of constraints on the n-replica system represented by H_n [see Eq. (18)]: (1) the self-overlap, i.e., helix fraction, of every replica is θ, (2) replicas are divided into groups of size m and the overlap between any two replicas in the same group is the perfectly correlated value $q_c=\theta$, and (3) the overlap between any two replicas that are in different groups is the perfectly uncorrelated value that corresponds to the situation in which the groups fluctuate independently of each other (see Fig. 8). Simply put, monomer i of replica α and monomer i of replica γ move (a) in unison, if α and γ are in the same group (b) without correlation if α and γ are in different groups.

The off-block matrix elements correspond to the overlap between replicas in distinct groups. The value of the intergroup overlap q_u is determined by considering the n-replica system H_n with constraints (1) and (2) imposed and with $g^{(1)}(T)=0$. The intergroup overlap is then the thermal-average overlap between two replicas in distinct groups. It is computed as a function of θ and m at $F=0$, resulting in $q_u=\theta^2$ as shown in the Appendix, Sec. 4.

At large integer n, we constrain the n-replica system to a point in the Q^c space of matrices, to obtain the free energy (per bp)

$$\beta f_n(\theta,m;T,F) = \beta g^{(1)}\theta - \frac{[\beta g^{(2)}]^2}{2}[mq_c + (n-m)q_u] - \frac{1}{mN}\log Z_{\text{group}}, \tag{23}$$

where Z_{group} is the partition sum of a single group in the cuRSB constrained n-replica system, with the coupling between replicas turned off ($g^{(2)}=0$. for all T). In this sense, Z_{group} gives the entropic contribution to the cuRSB free energy; its explicit analytical form is given in Appendix, Sec. 4 for $F=0$.

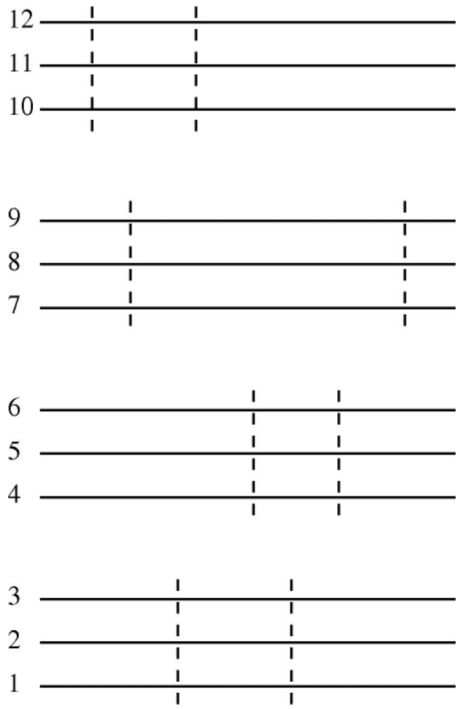

FIG. 8. A system of replicas subjected to cuRSB constraints. Each horizontal (solid) line represents the set of sequence positions $\{1,\ldots,N\}$ of a single replica, in a 12-replica system. The numerical labels on the left index the replicas, which are divided into groups of size $m=3$. In the replicas they intersect, the vertical (dashed) lines indicate the only sequence positions having a helix-state base pair (i.e., a monomer in the 1D polymer picture). Thus each vertical line indicates a sequence position exhibiting an overlap between each pair of replicas in the group that the vertical line touches. For each group of replicas, there are $p=2$ overlaps (vertical lines), where $\theta=p/N$ is the helix fraction (see Sec. III A). Shown is a particular microstate of the $n=12$, $p=2$, $m=3$ system that results from imposing cuRSB constraints upon the system represented in Eqs. (17) and (21). The reader should imagine all vertical lines fluctuating independently of each other, while maintaining $p=2$ distinct vertical lines in each group. The fluctuating system is represented by the free energy (per bp, per replica) in Eq. (23). The average overlap between each pair of replicas in distinct groups is the uncorrelated value $q_u=\theta^2$.

In the final step of this procedure, we obtain the equations that make βf_n stationary with respect to θ and minimized with respect to m, solving these equations with $n=0$. The complexities of this "minimization" are discussed in Appendix, Sec. 5.

B. Equilibria between the bound and the unbound phase

First we imagine the system at $F=0$ and $T<T_c$, where T_c is the largest temperature below which the helix fraction is nonzero (at $F=0$). In other words, T_c is the smallest tempera-

ture above which the strands are separated, except perhaps for a thermodynamically unimportant number of helix-state bps. We can call T_c a critical temperature, in the sense of a divergent correlation length (typical loop size), if the melting transition is continuous (higher-order). For $T<T_c$, we would like to compute the value of the unzipping force F_c above which the strands are separated. We call F_c the critical force because as $F \to F_c$ (from below), the system develops a divergent length scale. Specifically, the length of the unbound section of the molecule diverges (approaches the size of the long molecule).

In order to compute the critical force F_c, we apply the usual maximum-term method to determine the dominant partitioning of the molecule between the bound and unbound phase. At low forces, the molecule will be dominated by the bound phase. The critical force F_c is the point of phase-coexistence between the bound and unbound phases; here the free-energy density of the two phases must be equal. So, at each T, F_c is the solution of

$$f_0(\theta,m;T,F)|_{F=0} = g_F(T,F_c), \tag{24}$$

where g_F is the free energy (per bp) of the unbound phase, given in Eq. (6).

C. Properties of the bound macrostate

At each T, for $F<F_c(T)$, the majority of the molecule belongs to the bound phase, and the unbound section has an extent that is not thermodynamic (not proportional to N). For these values of T and F, we say that the molecule occupies the bound macrostate. As described in Sec. III A, the helix fraction θ is obtained as an optimized parameter from the replica method variational process. In terms of a general random-sequence theory, the helix fraction θ is the sequence-ensemble average of the thermal average of the fluctuating helix fraction,

$$\theta = \overline{\langle \theta[\rho] \rangle_g} = \frac{\partial}{\partial h^{\mathrm{hb}}} \overline{\left(-\frac{RT}{N} \log Z_g \right)} = \frac{\partial}{\partial h^{\mathrm{hb}}} \overline{f_g} = \frac{\partial \beta f_0}{\partial \beta h^{\mathrm{hb}}}, \tag{25}$$

as can be derived from the partition function of the one-dimensional polymer [see Eq. (10)]. In Eq. (25), $\theta[\rho]$ is the helix fraction for microstate ρ, and $\langle \ldots \rangle_g$ is the equilibrium thermal average for a fixed sequence $\{g_t\}$, and h_{hb} is the hydrogen-bonding component of $g^{(1)}(T)$. The derivative is taken with h, s, and $g^{(2)}$ constant when performed after sequence averaging, or with h, s, and every δg_t constant when performed before sequence averaging; we set $h^{\mathrm{hb}}=0$ after differentiation. In physical terms, deforming the value of h_{hb} away from zero (differentiation) changes the value of the hydrogen-bonding strength, which may be realized in experiment by changing the sequence of the DNA molecule. The last equality in Eq. (25) connects the general random-sequence definitions to our specific method using replicas.

To describe other properties of the bound macrostate, it is helpful to imagine the bulk of the molecule as a collection of helix or internal coil (loop) sequences, in the manner of Hill [14], rather than as a collection of bound sequences. A helix sequence is a set of contiguous sequence positions at which every bp is in the helix state. A loop sequence is a string of coil-state bps that is followed by a helix-state bp. We can compute the fraction of sequence positions where a helix sequence is followed by a loop, i.e., by a coil-state bp. We call this fraction the density of helix-coil junctions θ_{hc}, which is computed as

$$\theta_{\mathrm{hc}} = -\frac{\partial \beta f_0}{\partial \ln A}, \tag{26}$$

where θ and m are evaluated at the values obtained in the variational process. From θ_{hc}, we can compute the equilibrium average length of a helix sequence as $L_{\mathrm{helix}}=\theta/\theta_{\mathrm{hc}}$. Similarly, we can compute the average length of a loop sequence $L_{\mathrm{loop}}=1/\theta_{\mathrm{hc}}-L_{\mathrm{helix}}$ from the normalization condition $\theta_{\mathrm{hc}} \cdot (L_{\mathrm{helix}}+L_{\mathrm{loop}})=1$.

IV. RESULTS

A. Interpretation of the overlap matrix

Consider two uncoupled replicas of the one-dimensional polymer system H_g, see Figs. 3 and 6, at fixed finite N and $F=0$. Each replica has the same sequence $\{g_t\}$, which is picked at random from the sequence ensemble discussed in Sec. II B. The distribution of overlaps, for the sequence $\{g_t\}$, is defined as [27]

$$P_g(q) = Tr_{\rho^1} Tr_{\rho^2} \frac{e^{-\beta H_g[\rho^1]-\beta H_g[\rho^2]}}{Z_g^2} \delta(q - \langle \rho^1 | \rho^2 \rangle), \tag{27}$$

where $\delta(q-q')$ is a smoothed version of Dirac's delta and the overlap $\langle \rho^1 | \rho^2 \rangle$ is the fraction of sequence positions at which both replicas 1 and 2 have monomers.

In Appendix, Sec. 3, we give a detailed argument that for every $T \neq T_c$, the Gibbs state of the system does not exhibit ergodicity breaking, and thus at these temperatures, the overlap distribution $P_g(q)$ should exhibit a single peak. The location of the single peak should be a self-averaging quantity because (1) the overlap is an intensive property of the 2-replica system and (2) we argued that the loop and tail distributions have a finite cutoff. Thus, the sequence-ensemble-averaged distribution of overlaps should exhibit a single peak. We call the distribution obtained by averaging Eq. (27) over members of the sequence ensemble the true distribution

$$P(q) \equiv \overline{P_g(q)} = \delta(q - q_0) \tag{28}$$

in order to distinguish it from distributions computed in an approximate manner; q_0 is the sequence-ensemble average of the thermal average of $\langle \rho^1 | \rho^2 \rangle$. When the true distribution $P(q)$ has a single peak, we say that the system is replica symmetric [27] because $P(q)$ corresponds to an overlap matrix of the form shown in Fig. 9, in which each replica is related symmetrically to every other replica.

We can identify the location q_0 of this single peak, for the homopolymer (homogeneous-sequence) limit $g^{(1)}=g_\lambda^{(1)}(T), g^{(2)}=0$, by computing the thermal-average overlap of the 2-replica system with $g_t=g^{(1)}$ for each t. The distribu-

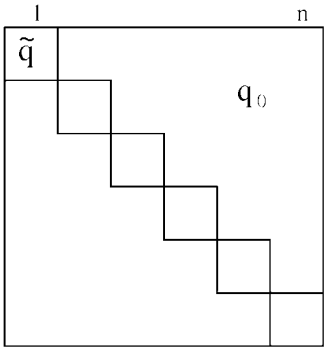

FIG. 9. The replica-symmetric matrix of overlaps. The diagonal elements give the self-overlap, i.e., the helix fraction $\tilde{q}=\theta$. Each off-diagonal element gives the overlap between a pair of replicas.

tion will be well localized around the thermal average due to the clustering property [27] of the relevant single-replica "component" (the term component is discussed in Appendix, Sec. 2). This value of the overlap, q_u, is the totally uncorrelated value in the sense that the two replicas are not "correlated" by the presence of sequence fluctuations. Thus, for the homopolymer $g^{(1)}=g_\lambda^{(1)}(T), g^{(2)}=0$, we have $q_0=q_u$. In Appendix, we compute $q_u=\theta^2$.

We naively expect that for the random heteropolymer (random-sequence model) case, $g^{(1)}=g_\lambda^{(1)}(T), g^{(2)}=g_\lambda^{(2)}(T)$, there will be some temperatures at which the peak in the true overlap distribution will shift toward values that are higher than q_u, i.e., $q_0>q_u$. That is, we expect that increasing the scale of the fluctuation in the potential g_t will cause the 1D monomers to be found more often in positions where the potential is low than positions where the potential is high. This tends to increase the thermal-average overlap, which gives the location q_0 of the peak of the true distribution $P(q)$.

In this work, we employ the replica method to test the stability of the realization of the overlap distribution that is localized at q_u. In Sec. III A, we constructed the cuRSB overlap matrix (Fig. 7), which predicts that the sequence-ensemble averaged overlap distribution should have the form

$$P_{\mathrm{cu}}(q) = m\delta(q-q_u) + (1-m)\delta(q-q_c), \tag{29}$$

where q_c is the value of the thermal-average overlap between two replicas that are constrained to have perfectly correlated motions of the 1D monomers, i.e., $q_c=\theta$. The distribution $P_{\mathrm{cu}}(q)$ is an approximation for the true distribution $P(q)$, which is useful because it allows for weight in a peak at $q>q_u$. We assume that if the variational procedure results in a weight at q_u that is less than unity, i.e., $m<1$, this suggests that the true distribution will exhibit a peak at some $q_0>q_u$, indicating that the system is taking advantage of the sequence heterogeneity (quenched disorder). This approximate overlap distribution allows us to study melting and unzipping behavior in a way that takes some account of the enhanced overlap due to sequence heterogeneity.

B. Homogeneous-sequence perfect-matching model treated with uRSB

We apply cuRSB to the homopolymer (homogeneous-sequence) case by setting $g^{(1)}=g_\lambda^{(1)}, g^{(2)}=0$ [see Eq. (A16)]. The variational process selects the replica group-size $m=1$ (see Figs. 10 or 11 for loop exponents $c=2.115$ or 1.75, respectively). Thus, two replicas of the homopolymer will show no tendency to move in unison and the overlap will have the uncorrelated value q_u [see Eq. (29)]. Fixing $m=1$ results in a special case of cuRSB that we call uRSB; we still require that βf_0 is stationary with respect to θ and maximized (see Appendix, Sec. 6) with respect to m.

The uRSB "minimization" conditions, for $g^{(2)}=0$, reduce to the single stationarity equation in Eq. (A18), which is identical to the equation that would be obtained if we solved the homopolymer using a maximum-term method in which we select the value of the helix fraction that minimizes the system's free energy (see Appendix, Sec. 5). Thus, our results are consistent with methods not based on replicas.

In Figs. 10 and 11, we also plot the values of the critical force F_c and the helix fraction θ obtained from the uRSB solution. Also, we plot F_c and the order parameter at the critical temperature

$$T_c^{\mathrm{homo}} = \frac{h}{s + R\ln\left\{\dfrac{1}{1-A[\zeta(c)-1]}\right\}} \cong T_{\mathrm{all}}\left\{1+\frac{R}{|s|}A[\zeta(c)-1]+O(A^2)\right\}, \tag{30}$$

which can be derived from the stationarity equation, see Eq. (A18), by setting $\hat{x}=1$. The notation $\zeta(s)$ is the Riemann zeta function. The logarithmic term, which is positive, shows that loops provide a small entropic stabilization of the bound phase. The temperature $T_{\mathrm{all}}=h/s$ is the location of the melting transition of the all-or-none version of our model. In the all-or-none model [7,33], which is obtained by taking the $A\to 0$ limit of our model, the molecule may exist in only one of two microstates: (1) all bps are in the helix state (2) all bps are in the coil state. The location T_{all} of the first-order melting transition of the all-or-none model is the temperature at which the helix-propagation free energy equals the free energy of the free coil macrostate, i.e., $g^{(1)}(T)=0$. We write T_c^{homo} in terms of T_{all} to emphasize that they are very close in numerical value; for the genome of λ-DNA $T_{\mathrm{all}}=90.5478$, $T_c^{\mathrm{homo}}=90.5505$ for $c=2.115$, and $T_c^{\mathrm{homo}}=90.5481$ for $c=1.75$ (significant figures chosen to discriminate among the temperatures).

To characterize the melting transition of the homopolymer, we say that the molecule is essentially a complete helix, $\theta=1$, up to T_{all}. Here, loops emerge in the structure of the molecule, and the helix fraction falls from unity to

$$\theta(T_c^{\mathrm{homo}}) = \frac{1+\tilde{A}[\zeta(c)-1]}{1+\tilde{A}[\zeta(c-1)-1]}, \tag{31}$$

where $\tilde{A}=Ae^{(h-Ts)/RT}$ and with the notation $\zeta(s)$ for $s\leq 1$, we mean positive infinity. This expression is derived according to Eq. (25), also see Appendix, Sec. 6. So, if $c=2.115$, $\theta(T_c^{\mathrm{homo}})=0.9986>0$, i.e., the helix fraction drops discontinuously to zero at T_c^{homo} and the melting transition is termed first order. But, if $c=1.75$, $\theta(T_c^{\mathrm{homo}})=0$, i.e., the helix fraction

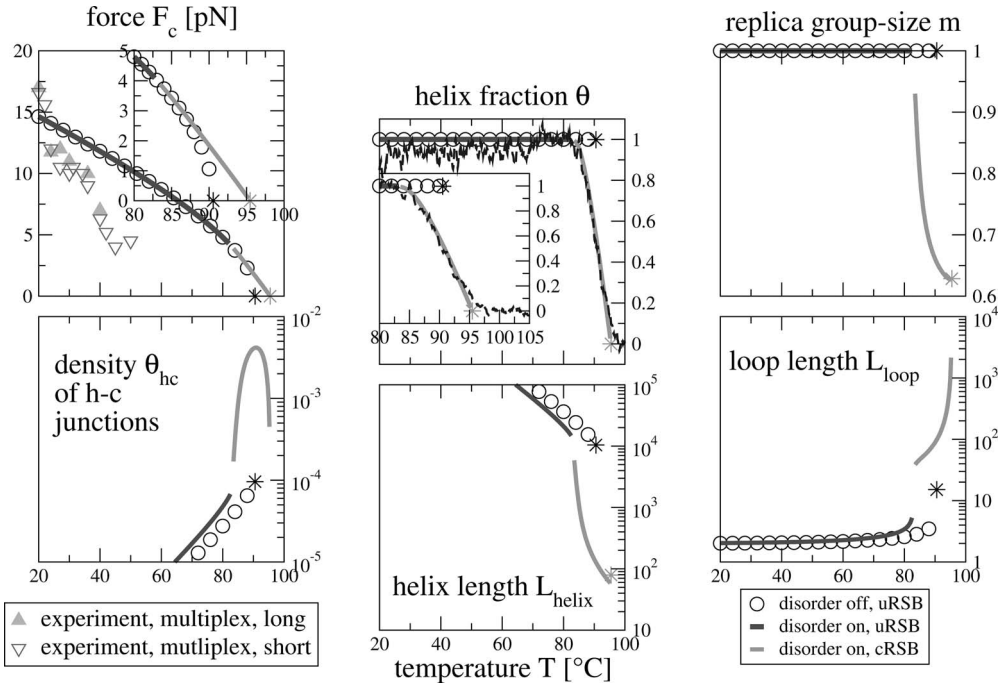

FIG. 10. The cuRSB scheme applied to the heteropolymer and homopolymer cases, for $c=2.115$. The black circles (homopolymer) and dark gray line (heteropolymer) give the uRSB (theory) results, obtained by solving $\partial_\theta \beta f_0|_{m=1}=0$ for θ and checking $\partial_m \beta f_0|_{m=1^-}>0$, where $1^-=1-10^{-6}$. The black star gives the solution T_c^{homo} of Eq. (30). We obtain $T_c^{\text{homo}}=90.55$ °C for $c=2.115$. The light gray line (heteropolymer) give the cRSB (theory) results, obtained by solving $\partial_\theta \beta f_0=0, \partial_m \beta f_0=0$ for θ, m, and checking that the stationary point is a local maximum in the m direction. The light gray star is the cRSB numerical proxy for T_c, obtained by solving $\partial_\theta \beta f_0=0, \partial_m \beta f_0=0$ for T, m, eliminating θ using Eq. (A17) for $\hat{x}=1-10^{-12}$ ($\theta \cong 0$); we obtain $T_c=95.33$ °C. Despite $c \geq 2$, the helix fraction falls continuously to zero, i.e., the melting transition is higher-order. In the top row, the left two plots show experimental results for comparison. We show the data from single-molecule unzipping via magnetic tweezers using the one-by-one protocol discussed in Sec. IV C 3. In the plot of the helix fraction, we show an experimental circular dichroism (CD) melting curve (dashed line), which has been rescaled so that the low-T values fluctuate about 1. In the plots of the density of h-c junctions θ_{hc} and the helix length L_{helix}, the exponential dependence of the uRSB curve (homopolymer and heteropolymer) continues down to $T=20$ °C.

drops continuously to zero at T_c^{homo} and the melting transition is termed higher order.

These transition orders are expected on the grounds of previous theory [6,34], where it was demonstrated that the order of the melting transition of the homogeneous perfect-matching model of dsDNA depends upon physical estimates of the parameter c as follows:

$$2 < c: \quad \text{first-order phase transition},$$

$$1 < c \leq 2: \quad \text{higher-order (continuous) phase transition},$$

$$c \leq 1: \quad \text{no true phase transition}. \tag{32}$$

If it were the case that $c \leq 1$, the helix fraction would decrease as a function of increasing temperature, but never fall to zero at any finite temperature.

Thus, our replica stationarity equations are consistent with the results of existing homogeneous PM theory. But if we look at Figs. 10 and 11, we see only a small difference between the helix-fraction curve for $c=2.115$ versus $c=1.75$. Due to the smallness of A, θ drops from unity to $\theta(T_c^{\text{homo}})$ over an interval of temperatures that is too small to see in these plots. Thus, on the temperature scale of these plots, and on the temperature scale of experiments, the homogeneous-sequence melting transition appears first order due to the cooperativity provided by the parameter A. For this reason, it is likely that neither experimental melting curves, nor unzipping experiments, of long homogeneous DNA can determine the true value of the loop exponent c. The smallness of A hides the value of c from experimental determination.

C. Random-sequence perfect-matching model treated with cuRSB and experimental comparison

1. Theoretical results

We find it interesting to apply the uRSB scheme to the random heteropolymer (random-sequence) case $g^{(1)}=g_\lambda^{(1)}(T), g^{(2)}=g_\lambda^{(2)}(T)$. For both the Kafri and Fisher values of the loop exponent c, Figs. 10 and 11 shows that the uRSB solutions cease above about 83 °C. In terms of the 2-replica picture, this means that below 83 °C the true overlap $q_0 \cong q_u$, whereas above 83 °C, $q_0 > q_u$ and so uRSB is a poor approximation.

We now apply the cuRSB scheme to the random heteropolymer case. Specifically, we restrict the cuRSB equations to $m<1$, calling the procedure cRSB. Here, we make βf_0 stationarity with respect to θ and m, then we check that the resulting stationary point is a local maximum in the m

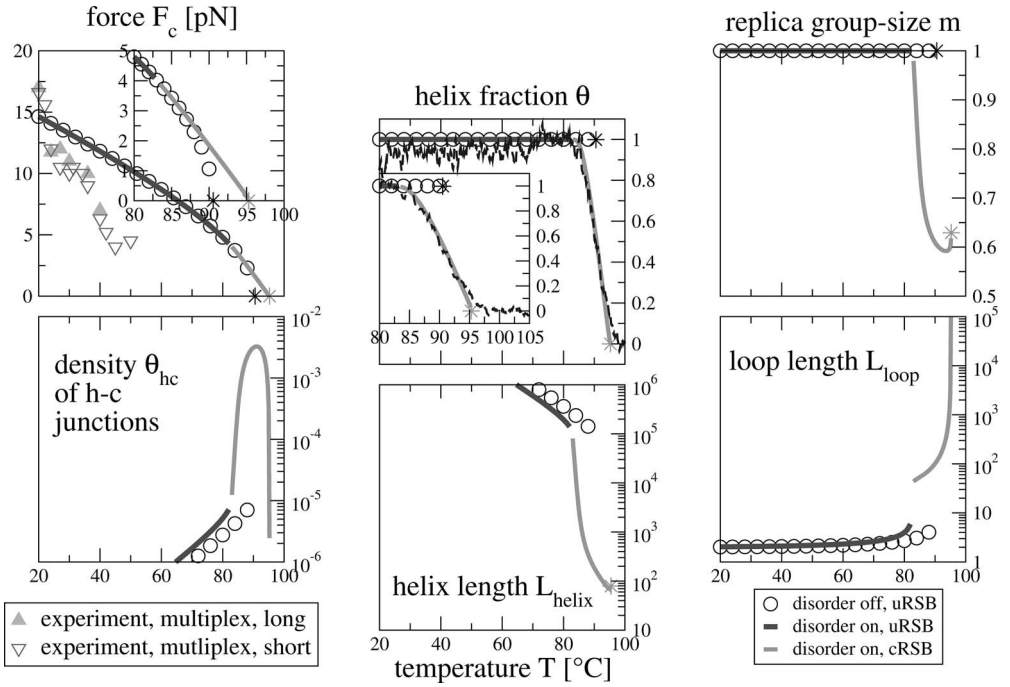

FIG. 11. The cuRSB scheme applied to the heteropolymer and homopolymer cases, for $c=1.75$. The data is the labeled and obtained in the same way as Fig. 10. For these parameters, the random-sequence model has $T_c=95.14$ °C, whereas the homogeneous model has $T_c^{\mathrm{homo}}=90.55$ °C, as for $c=2.115$.

direction. For $c=2.115$ or $c=1.75$, the results are shown in Figs. 10 or 11. We see that cRSB succeeds where uRSB fails. Solution of the cRSB scheme indicates that cuRSB selects $m<1$, thus the approximate distribution $P_{\mathrm{cu}}(q)$ in Eq. (29) has weight in the peak at $q=q_c$. We interpret this as meaning that for the true overlap distribution $P(q)$, $q_0>q_u$, and the monomers of the 1D polymer are localizing at sequence positions where δg_t is low.

For both $c=2.115$ and $c=1.75$ (Figs. 10 and 11), as T is increased, the helix fraction of the random heteropolymer departs from unity at a lower temperature than for the homopolymer. In other words, the sequence fluctuations of the random sequence stabilize loops at temperatures where they are suppressed in the homogeneous sequence. Moreover, loops emerge in the random-sequence model well below T_c, whereas in the homogeneous model loops are suppressed until extremely close to T_c; these results are consistent with existing numerical results for random-sequence models of DNA [7]. For $c=2.115$ $T_c=95.33$ °C, and for $c=1.75$ $T_c=95.14$ °C; the computation of T_c is explained in Fig. 10. The model predicts that loops emerge at 83 °C, which is about 12 °C below T_c. This is consistent with classic spectroscopic measurements on the λ genome, showing that the width of the melting transition is 10–15 °C [2,3]; note that these measurements were obtained at a variety of salt conditions.

Our theory allows us to compute six properties of the system as shown in Figs. 10 and 11, but spectroscopic and unzipping experiments measure only two of these properties—helix fraction and critical force—which allow direct comparison of our theory with experiment.

2. Experimental melting curves

To compare our calculations directly with current spectroscopic experiments, we measured circular dichroism (CD) melting curves of λ-phage genome in phosphate-buffered saline (PBS) (see Figs. 10 and 11). To obtain this data, the temperature was ramped at 5 °C/min over the interval 15–105 °C, we averaged over four runs, the concentration of the DNA was 100 $\mu g/\mathrm{mL}$, and the device is a Jasco J-710 Spectropolarimeter with a PTC-378 W Jasco temperature controller. We compared two molecules: (1) the molecular construction used in unzipping studies, that contains a hairpinned λ genome [4,35] and (2) the naked λ genome (New England Biolabs). We found that the temperature at which the reading decreased from the low-T value (emergence of loops) and the temperature at which the reading met the high-T value (separation of strands) were very similar between the two molecules (data not shown). In Figs. 10 and 11, we show only the data for the naked λ genome (dashed line in top-middle plot). The agreement between theory and experiment is extremely good. Not only does our theory predict the temperature at which loops emerge, but it predicts the melting temperature T_c to within 0.8% on the Kelvin scale. The predicted T_c is 95 °C and the observed is approximately 98 °C (see Figs. 10 and 11).

3. Experimental unzipping of the entire genome and perfect-matching models

To support the predictions of our theory, we now compare the random PM model to a recent set of unzipping experiments. The apparatus, molecular construction, and unzipping geometry are identical to that used in Danilowicz *et al.*

[4,35]; we refer the reader to this work for details. But, the procedure for changing the applied force and observing the extension are different. Briefly, the experimentally controlled force is applied in a manner that tends to separate the bp at the AT-rich end of the double helix; this geometry is cartoonized in Fig. 1. In the method of the previous work, many copies of the λ genome were simultaneously subjected to force—here we call this parallelized method "multiplex"—and the critical force was reported for the unzipping of the first 1500 bps. There were two variants of this experimental method, which we call "long" or "short," indicating that the molecular construction was left at the target temperature for 15 h or 20 min, respectively, before the unzipping force was applied.

In the experimental method used to obtain the results we present, the extension of an individual copy of the earlier described molecular construction [4,35] is recorded as the applied force is first increased, and second decreased, at a constant rate. We call this set of experiments "one-by-one" to contrast with the "multiplex" experimental technique [4]. First, the molecular construction is allowed to equilibrate at the target temperature, with zero applied force, for 5 min. Then, the applied force is increased by $0.5pN$, then held constant for 2 s, etc., until the maximal extension is observed. Next, the force trajectory is reversed. The extension at the maximal value of the force confirms the unzipping of the entire 48 502 bp genome, and the extension at the minimal force suggests the complete rezipping. As a function of temperature, the force at which unzipping occurs, and the force at which rezipping occurs, is plotted in Figs. 10 and 11 (triangles in the top-left plot). As expected from the landscape picture of unzipping present in the uncorrelated random zipper model [30], the unzipping and rezipping forces bracket the equilibrium critical force F_c predicted by the perfect-matching RHP model. Note that at the temperatures of the experiment, the perfect-matching RHP reduces to the all-or-none model because no loops are present. It is difficult to comment further on the relationship between our model and this experiment because the forces reported in the experiment will depend upon the sequence-dependent kinetic barriers to unzipping and rezipping, and the rate at which the applied force is changed. Due to the barriers and nonzero rate of change of the applied force, the extension of the molecule is probably not governed by thermal equilibrium. But, because the theoretical critical force is less than the unzipping force, but greater than the rezipping force, these unzipping experiments provide a successful consistency check for our theory.

4. Failures of other models for the phase diagram of the phage-λ genome

We now compare the infinite-chain random-sequence PM model to other candidate models for the unzipping phase diagram of the phage-λ genome. The all-or-none model, trained on the entire genome, will fail to model the entire genome because the width of the melting transition of the model was computed (by us) to be 0.01 °C (data not shown), much smaller than the width of the experimental CD melting curve. The all-or-none model is a two-state model so the helix fraction at zero force is $\theta_{\mathrm{all}}(T)=[1+u_h^{-N}]^{-1}$, where $u_h=\exp(-g/RT)$ and g is the free energy per bp of the all-helix microstate. It is the dependence on N that makes the melting transition very narrow. Thus, both the finite-chain and the infinite-chain all-or-none models are poor models for the unzipping phase diagram of the entire genome.

We may also rule out the infinite-chain zipper model as a model for the entire unzipping phase diagram of the genome of phage λ. It was shown in [30] that the infinite-chain zipper model (for both homogeneous and random sequence) has the same critical-force curve $F_c(T)$ as the all-or-none model, for which $F_c=0$ at the melting temperature $T_{\mathrm{all}}=90.5478$ °C (see Sec. IV B). This prediction disagrees with experimental CD melting curves, as shown in Figs. 10 and 11. We cannot rule out a finite-chain zipper model, but the literature on DNA melting [7] suggests that loops are necessary to model the entire λ-phage genome (see Appendix, Sec. 7).

To include loops, one may propose a homogeneous-sequence perfect-matching (PM) model. In Sec. IV B, we showed that the width of the melting transition of the infinite-chain homogeneous PM model is too small to model the experiment. Similarly, the finite-chain homogeneous PM has been shown, via matrix calculations, to have a narrow melting transition; for chain-length $N=5000$ bps and cooperativity parameter $A=10^{-3}$, the width was about 1 °C [6]. Since λ-DNA has higher N and lower A, the melting transition of the homogeneous PM model for λ-DNA will have an even smaller width. Thus, both loops and sequence heterogeneity are necessary to model the width of the melting transition—and thus to model the unzipping phase diagram—of the entire λ-phage genome.

V. DISCUSSION

A. Effective loop exponent and the order of the melting transition

In this section, we will focus on the following feature of the results: the presence of sequence heterogeneity smoothes the fall of the helix fraction from unity to zero. This smoothing effect is so strong that for the parameter $c=2.115$, the random heteropolymer has a higher-order melting transition whereas the homopolymer has a true first-order melting transition. For $c=1.75$, the homopolymer predicts a true higher-order melting transition that appears first order because T_{all} is extremely close to T_c^{homo}, and this apparent first-order transition is converted into a higher-order transition by the sequence heterogeneity. Very recently, a large-deviation-theoretic argument demonstrated that for any $c>1$, the melting transition of the random heteropolymer is higher order, i.e., continuous [36] (note that this result, while important, makes no predictions for the location of the melting temperature, or the temperature dependence of the helix fraction or critical force). Thus, the results of our replica calculation can be understood in terms of a rigorous mathematical argument. In this paper, we show that for both values of c considered, the presence of sequence heterogeneity raises the apparent order of the melting transition, and this effect shows up in an effective value of c that controls the behavior of the

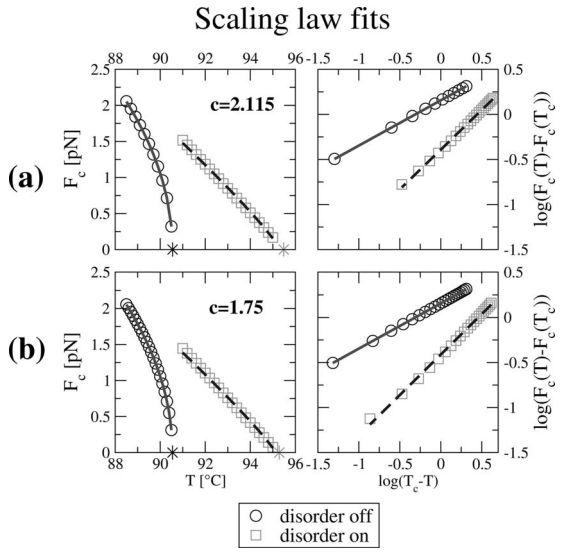

FIG. 12. Scaling law fit obtains an effective loop exponent c^{fit}, for (a) $c=2.115$ and (b) $c=1.75$. The critical-force curves are obtained from the cuRSB approximation scheme. We fit the curves to the scaling law in Eq. (34), with $\alpha=1/2\eta$, with $\eta=\min(1,c_{\mathrm{fit}}-1)$ in order to obtain the effective loop exponent c^{fit} for the random heteropolymer and homopolymer models. These exponents are obtained from the slopes of the straight-line fits in log-log scale, and the quality of the fit is measured in terms of r^2, the square of the Pearson correlation coefficient. (a) For $c=2.115$, the heteropolymer is fit on $91\leq T\leq 95$ °C giving $\alpha=0.881$ with $r^2=0.998$, thus $c^{\mathrm{ef}}=1.57$. (b) For $c=1.75$, the heteropolymer is fit on $91\leq T\leq 95$ °C giving $\alpha=0.899$ with $r^2=0.997$, thus $c^{\mathrm{ef}}=1.56$. For both $c=2.115,1.75$, the homopolymer is fit on $88.5\leq T\leq 90.5$ °C and has exponent 0.502 with $r^2=0.999\,997$ that is consistent with Eq. (33). In this figure, the base of the log is 10.

helix fraction and the critical force in the neighborhood of T_c.

1. Critical-force scaling for the homogeneous-sequence model

As temperature is increased, the homopolymer remains a complete helix up to approximately $T_{\mathrm{all}}=T_c^{\mathrm{homo}}-O(A)$. Given that the resolution of Figs. 10 and 11 is about 1 °C, the experimentally relevant shape of the critical-force curve near T_c^{homo} is determined by the all-or-none model. If we set $g^{(1)}(T)=g_F(T,F_c)$, and expand about T_{all}, we obtain

$$F_c=\frac{k_BT}{b}[e^{-(b/a)\beta g^{(1)}(T)}-1]\cong\frac{k_BT}{(ab)^{1/2}}\left(\frac{|s|}{R}\right)^{1/2}(T_{\mathrm{all}}-T)^{1/2}, \tag{33}$$

which gives the shape of the critical-force curve in the neighborhood of T_c^{homo}. From Eq. (33), we see that the critical force scales like

$$F_c\sim|T-T_c|^{\alpha} \tag{34}$$

with $\alpha=1/2$ exactly, to the extent that $T_{\mathrm{all}}\cong T_c^{\mathrm{homo}}$, and setting $T_c=T_c^{\mathrm{homo}}$. This scaling is confirmed by a numerical fit (see Fig. 12).

But, if we were to plot F_c for $T_{\mathrm{all}}<T<T_c^{\mathrm{homo}}$, theory predicts that the shape of F_c is determined by the free energetics of the loops. Specifically, in a homogeneous-sequence PM model, in which volume exclusion effects in the loop sequences are incorporated through the value of c [17], Mukamel and Shakhnovich showed that above T_{all} and very near T_c^{homo}, the scaling law Eq. (34) holds but with $\alpha=1/2\eta$, where $\eta=\min(1,c-1)$ and $T_c=T_c^{\mathrm{homo}}$. Notice that if $c\geq 2$, then both the loop-dominated scaling law, with $\alpha=1/2\eta$, and the loopless scaling law, with $\alpha=1/2$ [see Eq. (33)], predict that the critical force grows with the square root of the distance below T_c^{homo}, at least to the extent $T_{\mathrm{all}}\cong T_c^{\mathrm{homo}}$. But if $c<2$, then the critical-force curve will exhibit a crossover as T is increased passed T_{all} because the exponent of the loop-full scaling law falls below the exponent of the loopless law [Eq. (33)]. But, even if one could find a sequence of dsDNA that is well approximated by the homogeneous PM model, it is unlikely that an experiment would be able to resolve this crossover because $T_{\mathrm{all}}\cong T_c^{\mathrm{homo}}$. It is unlikely that experiments on homogeneous sequences, either spectroscopic melting curves or forced unzipping, can discriminate between different values of the loop exponent.

2. Critical-force scaling for the random-sequence model

We encode a single parameter with the manner in which sequence heterogeneity raises the order of the melting transition. We take the critical-force curve near T_c of the random-sequence model, and fit it to the homogeneous-sequence model, i.e., $\alpha=1/2\eta$, and extract the value of the loop exponent c^{fit} that solves $\eta=\min(1,c^{\mathrm{fit}}-1)$ (see Fig. 12). For the random PM model at $c=2.115$, the fitted value of the loop exponent is $c^{\mathrm{fit}}=1.57$; because $c^{\mathrm{fit}}<2$, the melting transition is higher order, according to Poland-Scheraga theory [6] [see Eq. (32)]. For the random PM model at $c=1.75$, $c^{\mathrm{fit}}=1.56$. Thus, for both values of c, $c^{\mathrm{fit}}<c$ quantitatively encodes the result that sequence heterogeneity raises the order of the melting transition. Relating microscopic structural features to the loop exponent is an approach taken in Poland-Scheraga theory [6] as well.

3. Order of the melting transition for the random-sequence model

We corroborate the above interpretation of c^{fit} with an analytical result for the distribution of loop lengths. For the infinite-chain homogeneous PM model, of all the loops that are present in the chain, the number of loops n_ℓ that have length ℓ is a fluctuating variable with mean value $\langle n_\ell\rangle$; the mean total number of loops is $\langle N_{\mathrm{loop}}\rangle=\Sigma_{\ell=2}^{N}\langle n_\ell\rangle$. We find that $\langle n_\ell\rangle/\langle N_{\mathrm{loop}}\rangle=x^\ell\ell^{-c}/[\phi(x,c)-1]$, where $\ell\geq 2$, x is a fugacity that sets the helix fraction, and $\phi(x,c)$ is the polylogarithm function [see Eq. (A15)]. For the infinite-chain random-sequence PM model, we can obtain the analogous sequence-ensemble averages $\overline{\langle n_\ell\rangle}$ and $\overline{\langle N_{\mathrm{loop}}\rangle}$, finding

$$\overline{\langle n_\ell\rangle}/\overline{\langle N_{\mathrm{loop}}\rangle}=\frac{x^\ell\ell^{-mc}}{\phi(x,mc)-1}, \tag{35}$$

which is a proxy for the probability distribution of loop lengths for a typical sequence in the sequence ensemble. Equation (35) can be obtained from the replica free energy [Eq. (23)] by differentiation with respect to a force that tends

to increase the number of loops of length ℓ. We call $c^{\mathrm{ef}}=mc$ the effective loop exponent, as it governs the distribution of loop lengths for a typical random sequence. For example, we checked that the average loop length obtained from distribution $\langle n_\ell \rangle$ agrees with L_{loop} obtained in Sec. III C. From Figs. 10 and 11, we can read off that at T_c, $m=0.629$ for either $c=2.115$ or 1.75, and thus $c^{\mathrm{ef}}=1.330$ or 1.10, respectively. These values are in broad agreement with the values $c^{\mathrm{fit}}=1.57$ or 1.56, respectively, where the quantitative discrepancy may be due to the fact that the fitting interval is 4 °C wide, whereas we have only given c^{ef} at T_c.

Since this $c^{\mathrm{ef}}=mc$ appears in our equations for the helix fraction and the free energy (per bp) [see Eq. (A17) and Eq. (23)], we conclude that c^{ef} determines the order of the melting transition as higher order (not first order), and determines the shape of the critical-force curve F_c immediately near T_c. From homogeneous PM theory, if $1<c\leq 2$, the higher-order transition [see Eq. (32)] may be classified more specifically in terms of the exponent of the scaling relation between the helix fraction and temperature near T_c,

$$\theta \sim (T_c - T)^{\beta}, \tag{36}$$

where $\beta=(2-c)/(c-1)$ as derived by Fisher [7]. Thus we can say that the transition is continuous (higher order), but we can also give the order parameter's scaling exponent $\beta=2.04$ or 8.90, which correspond to $c^{\mathrm{ef}}=1.33$ or 1.10, for $c=2.115$ or 1.75, respectively. Consequently, for both for $c=2.115$ and $c=1.75$, the second derivative of the free-energy density $\partial\theta/\partial T=\partial^2\overline{f_g}/\partial T\,\partial h^{\mathrm{hb}}$, refer to Eq. (25), is continuous at the melting transition. In the extended Ehrenfest classification [37], this phase transition would be classified as third or greater order. We may generalize by giving the lowest-order derivative at which a discontinuity appears for arbitrary c^{ef},

$$\begin{aligned}
2 < c^{\mathrm{ef}}&: \quad \text{1st derivative has jump discontinuity,}\\
3/2 < c^{\mathrm{ef}} \leq 2&: \text{2nd derivative has divergent discontinuity,}\\
c^{\mathrm{ef}} = 3/2&: \quad \text{2nd derivative has jump discontinuity,}\\
c^{\mathrm{ef}} < 3/2&: \text{3rd or greater derivative has discontinuity.}
\end{aligned} \tag{37}$$

Importantly, homogeneous-sequence theory makes the prediction that for $c=2.115$, the melting transition is first order, having a jump discontinuity in the helix fraction. For $c=1.75$, homogeneous PM theory predicts that $\beta\approx 1/3$ so that $\partial\theta/\partial T$ diverges at T_c. Our random-sequence PM theory contradicts both these statements, predicting that both θ and $\partial\theta/\partial T$ are continuous across the transition. Thus, our random-sequence theory is consistent with experimental melting curves of phage λ, see [2,3] and Figs. 10 and 11, while the homogeneous-sequence theory is not.

B. Previous theoretical efforts on the random-sequence perfect-matching model

We acknowledge that other calculations have been devised to study the zero-force melting of random-sequence models of long natural dsDNA. Poland and Scheraga [38] treated the infinite random-sequence perfect-matching (PM) model with an annealing-sequence approximation developed by Lifson and Allegra [9,10], in which the helix-propagation free energies g_t are considered as mechanical variables that anneal hand in hand with the conformational microstate of the molecule [7]. While such methods are analytical, these methods are less favorable because they do not, in general, result in the thermodynamics of a typical random sequence. Additionally, the annealing-sequence methods have been shown to give poor approximations to the exact results (Lehman and McTague [8]) in the case of the nearest-neighbor Ising model [7,11]. Our calculation is advantageous because the replica method [27] offers an analytical route to perform the sequence-ensemble averaging operation exactly, at least in principle. In practice, approximation schemes involved in the replica method reduce the exactness of the results, but we argue that (a) the results are "more exact" than annealing-sequence methods; (b) the replica method offers the possibility of reducing the severity of the approximation. Regarding specific results of annealing-sequence methods, Poland and Scheraga [38] found the dubious result that, for $c=1.75$, the typical length of a helix sequence was the same for the homogeneous-sequence and random-sequence models. Contrariwise, we find that near T_c^{homo}, the helix length of the random-sequence model is orders of magnitude shorter (see Fig. 11). But the most significant advantage of our calculation is that our theory outputs an effective loop exponent c^{ef} that analytically controls the shape of F_c near T_c, and thus cleanly classifies the melting transition as a type of phase transition.

In another theoretical effort, Poland and Scheraga have numerically studied the finite-chain random-sequence PM model. They found transition widths consistent with our results [7,38]. Since this model is finite, it does not yield a classification of the phase transition that results in the infinite-chain limit. As an alternative to an infinite-chain calculation, one may study how the properties of the finite model change as the system size is increased, but to our knowledge, this has not been done.

C. Previous theoretical efforts on the random-sequence Peyrard-Bishop model

We now compare our results to those for the Peyrard-Bishop (PB) model, in which loops are modeled as Gaussian chains. The original PB model [39] corresponds to the PM model with $c=3/2$; the homogeneous-sequence PB model has a higher-order melting transition, consistent with Poland-Scheraga theory. Recently, Cule and Hwa studied the variable-stiffness PB model [40], in which the Gaussian-chain stiffness coefficient depends on the loop configuration. The variable-stiffness feature does not make the transition first order, but only sharpens the transition [41]; it appears that the variable-stiffness parameter is analogous to the cooperativity parameter A of the PM model. To the variable-stiffness PB model, Cule and Hwa added sequence randomness, finding that both variable-stiffness and sequence randomness where necessary to generate multistep melting

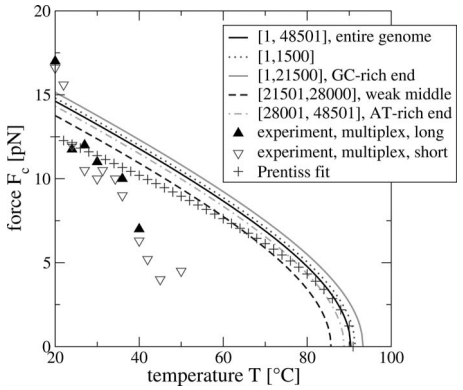

FIG. 13. Unzipping subsequences of the λ genome: the all-or-none model compared with experiment. For each curve labeled with a subsequence of the genome of λ phage (sequence position is labeled from the GC-rich end), we computed the critical force of an all-or-none model, with the free energy per bp of the all-helix state trained on the indicated subsequence. For comparison, we show the published results (triangles) of the multiplexed single-molecule experiment on the unzipping of the subsequence [1,1500] (see Sec. V D); this data is reproduced with permission [4]. We also show the critical force (plus signs) [4].

curves for sequences of approximately 3000 bps [41]. Despite the multistep behavior found for intermediate-length sequences (3000 bps), the authors speculated that the melting curve would become smooth as the length of the random sequence becomes large; this speculation is consistent with our results. Thus, the speculated behavior of the random-sequence variable-stiffness PB model is consistent with our results for infinite sequences: there is a phase transition of the higher-order (continuous) type.

D. Comparison with unzipping studies on the first 1500 bps

The theoretical and experimental results presented in Sec. IV show mild discrepancies with previous theoretical and experimental findings. We now comment on this disagreement. In a multiplex single-molecule experiment to determine the temperature-force phase diagram for unzipping the first 1500 bps of the 48 502 bp genome of phage λ, Danilowicz *et al.* presented an experimental critical-force curve and described it with a finite-chain all-or-none model (chain may exist only in the all-helix or all-coil microstates) [4]. The free energy per bp of the all-helix microstate was $g_{\rm fit}^{(1)}(T)=h_{\rm fit}-Ts_{\rm fit}$, with $s_{\rm fit}=-20.6$ cal/mol·K obtained from averaging over the first 1500 bps at the left end of the l strand, and $h_{\rm fit}=T_{1/2}s_{\rm fit}=-7.5$ kcal/mol, where $T_{1/2}=90.9$ °C was determined as the midtransition point of the circular dichroism (CD) melting curve of the entire genome of phage λ. This $T_{1/2}$ is perhaps a good approximation for the melting temperature of a molecule composed of the first 1500 bps because the width of the transition of the entire genome is ≈12 K while $T_{1/2}\approx 364$ K, so the error is about 3%.

In Fig. 13 we show the theoretical (plus signs) and experimental (triangles) results presented by Danilowicz *et al.* We reproduce the critical-force curve that results from their parameters; we have ignored the elastic term of the model for the free energy of the unbound section used in that work [4]. At each temperature T, the critical force F_c is computed by solving $g_F(T,F_c)=g_{\rm fit}^{(1)}(T)$ [see Eq. (6)]. The data point at $F_c=0$ is the melting temperature $h_{\rm fit}/s_{\rm fit}$, which has value $T_{1/2}$. At each point on the $F_c(T)$ curve, the model is evenly distributed between the all-helix and all-coil state.

On the temperature range spanning their experimental data, $20\le T\le 50$ °C, the λ genome has no loops (see Fig. 10) so the critical force of our random-sequence PM model is identical to the critical force of the all-or-none model with comparable base-pairing parameters. So, in Fig. 13 we show the results of the all-or-none model (lines) with $g_\lambda^{(1)}(t,t';T)=h_\lambda(t,t')-Ts_\lambda(t,t')$, which is computed with Eq. (13) in Sec. II B, but we restrict the average to the portion of the genome between sequence positions t and t'. We consider different choices of $[t,t']$. As above, the critical force is obtained from $g_F(T,F_c)=g_\lambda^{(1)}(t,t';T)$, and the melting temperature is computed as $T_{\rm all}(t,t')=h_\lambda(t,t')/s_\lambda(t,t')$.

For the sequence [1,1500] and the entire genome, our computed critical-force curves are very similar, and both are roughly 2 pN higher than both the experimental and theoretical critical forces presented by Danilowicz *et al.* [4] for $24\le T\le 35$ °C. The disagreement between the present and previous theoretical results is understood because we obtain $h_\lambda(1,1500)=-9.46$ kcal/mol and $s_\lambda(1,1500)=-26.0$ cal/mol·K, whereas Danilowicz *et al.* obtained different values because they used a different nearest-neighbor parameter set and neglected the corrections for sodium concentration.

But the disagreement between our experiment and the previous experiment [4] is not understood (see Fig. 14). Specifically, since our all-or-none model shows that sequence [1,1500] is more stable to unzipping than the sequence [1,48 502] (Fig. 13), we would expect that the experimental F_c from the multiplex experiment [4] should be above the "rezip" data points in Fig. 14. The "rezip" and "unzip" data points bracket the true experimental F_c for the λ genome. This observation implies that the experimental F_c determined by Danilowicz *et al.* underestimates the true value, or that our experimental F_c overestimates the true value. This is the first discrepancy between the previous and present experiment.

A second discrepancy between our experiment and the previous experiment is that the previous experiment shows a drop in the experimentally reported critical force F_c, occurring near 40 °C [4], whereas this drop does not appear in the present results. We speculate that both the first and the second discrepancies may be due to kinetic barriers that prevent experimental sampling of equilibrium on the experimental time scale (these barriers could be, e.g., of the type embodied in the random zipper model [30]).

These experimental discrepancies imply a discrepancy between our theory and the previous experiment. Despite these discrepancies, our results of the all-or-none model follow expected trends in stability (see Fig. 13). At a given temperature T, the subsequence labeled "weak middle" has the least stable helix-state free energy, the AT-rich subsequence is more stable, and the GC-rich subsequence is even more

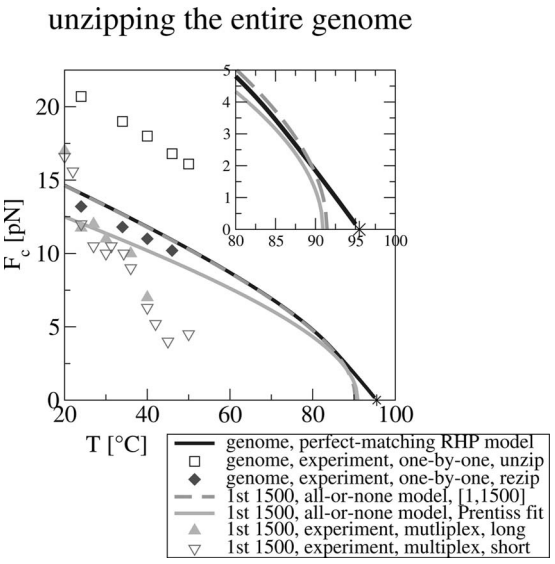

FIG. 14. Unzipping the entire genome: the perfect-matching RHP model compared with experiment. In the first three data sets (black line, open squares, solid diamonds), we compare our perfect-matching RHP model to the results of our "one-by-one" experiment in which an individual copy of the phage-λ genome is unzipped and then re-zipped (see Sec. IV C 3). The second four data sets (dashed line, gray line, up triangle, down triangle) show how the all-or-none model compares with the "multiplex" experiment on first 1500 bps, in which many copies of the genome are unzipped at the same time (see Sec. V D). The designations "long" and "short" refer to the duration of time at which the sample is allowed to thermalize at the target temperature.

stable. At a given temperature T, the critical force F_c to unzip the model parameterized on these subsequences follows this same trend, providing a consistency check for our calculation. The values of the melting temperatures T_{all} also follows this trend.

VI. CONCLUSION

We have calculated the phase diagram (critical-force curve) of the random-sequence infinite-chain perfect-matching (PM) model for the single-molecule forced unzipping of the 48 502 bp genome of phage lambda, see Figs. 10, 11, and 14. In the random-sequence approximation for DNA, we assume that the thermodynamics of the phage-λ genome is similar to that of a fictitious genome that is obtained from the λ genome by randomly rearranging the base-pairing interactions along in the linear genome structure. This approximation is supported by the accuracy of the melting curve predicted by our model. In the infinite-chain approximation for dsDNA, we assume that near the melting temperature, the length of the chain is many times the length of a typical helical sequence. We find this approximation to be self-consistent; see Appendix, Sec. 7. In a perfect-matching model for dsDNA, we assume that the only bases that share a pairing interaction are those that are paired in the native state. An additional feature of a perfect-matching model is that the impact of the structure of single-stranded dsDNA upon the free energy of a loop (a contiguous sequence of separated base pairs) is represented with the scaling law A/ℓ^c for the entropic portion of the partition function of a loop [see Eq. (4)]. The loop exponent c depends on the model for how monomers in a given loop interact with each other and with other loops. In this work, we consider $c=2.115$ (or $c=1.75$), corresponding to the model in which monomers in a loop interact with each other, and do (or do not) interact with other loops.

Our calculation of the unzipping phase diagram of a ds-DNA model with loops and sequence heterogeneity was performed analytically with the replica method. We use a heuristic argument (see Appendix, Sec. 3) against ergodicity breaking to interpret the results of our replica-theoretic calculation. These methods can be applied to a variety of uncorrelated random-sequence models parameterized by $g^{(1)}(T)$ and $g^{(2)}(T)$ as well as for a variety of models for loop structure, embodied in the loop exponent c. In the phase diagram, the critical-force curve defines a line of first-order phase transitions. The line ends at zero force where there is a higher-order melting transition. The critical-force curve near T_c, see Fig. 12, has an experimentally observable shape that gives evidence that the melting transition is higher-order (continuous). Specifically, our theory predicts that the transition is a third- or greater-order phase transition; see Sec. V A 3 and the discussion of the loop exponent c^{ef}, which is an output of our replica-theoretic calculation that governs the helix fraction and critical-force curve near the melting temperature.

We contrast the results for the random-sequence model against the results for the homogeneous-sequence model. In the homogeneous model, at zero force, loops are suppressed until less that 0.01 °C below the critical temperature T_c. Consequently, for experimentally observable temperatures near T_c, the homogeneous sequence F_c has a shape that is determined by the all-or-none model, giving a scaling law $F_c|T-T_c|^{\alpha}$ with exponent $\alpha=1/2$ almost exactly, see Sec. V A 1 (in the all-or-none model, the molecule exists in only two microstates: all bps paired or all bps unpaired). Thus, the true first-order phase transition associated with the Kafri *et al.* [16] value of the loop exponent ($c=2.115$) might be present but it is not experimentally observable. Contrariwise, the random-sequence model, at zero force, shows loops 12 °C below T_c. The presence of loops cause the critical-force curve, see Figs. 10 and 11, to deviate substantially from that of the homogeneous-sequence model. The resulting T_c is ≈5 °C higher, and the critical-force scaling law has exponent $\alpha=0.881$ or 0.899 for $c=2.115$ or 1.75, respectively, indicating a slightly sublinear temperature dependence (see Fig. 12). Motivated by the presence of loops, we fit the random-sequence critical-force curve to the scaling law $F_c|T-T_c|^{\alpha}$, with $\alpha=1/2\eta$ and $\eta=\min(1,c^{\mathrm{fit}}-1)$, obtaining a fitted value c^{fit} of the loop exponent, see Sec. V A 2. Since $c^{\mathrm{fit}}=1.57$ or 1.56 for $c=2.115$ or 1.75, respectively, Poland-Scheraga theory states that the melting transition is higher order. These exponents are experimentally accessible observables that can be determined for the 48 502 bp λ-phage genome. The experimental validation would support our claim that the effective loop c^{ef} predicts that the melting transition is third or greater order.

The experimental measurement of F_c near T_c is hindered by the denaturation of proteins involved in the molecular attachment of the genome to the plastic bead that responds to the magnetic tweezers. Due to this protein, the unzipping studies conducted in phosphate-buffered saline (PBS) are limited to temperatures below about 55 °C. But the melting temperature of dsDNA depends strongly on salt conditions. For example, the genome melts at about 75 °C in a TRIS solution (data not shown). Thus, it is likely possible to find a solution with ionic conditions that lower the melting transition of the genome to temperatures where the molecular construction is stable.

In this work, the replica method has generated results with semiquantitative agreement with experiment (see Figs. 10–12 which show critical-force curves and melting curves). Our calculation contains no adjustable parameters, refer to Table I, and our predicted melting temperature agrees with the experimentally determined value to 0.8% in the Kelvin scale.

We have shown how the determination of the experimental critical-force scaling exponent α can give confirmation of the third or greater order of the melting transition of the genome of bacteriophage λ. That the melting transition is third or greater order means that the helix fraction and its temperature derivative (differential melting profile) are continuous functions of temperature. We can briefly speculate on the generalization of our claim to other sequences of dsDNA. Our model of the λ genome includes three key features of the sequence: (1) the sequence is very long, viz., the sequence is many multiples of the length of a typical helix section near the melting temperature (see Appendix, Sec. 7), (2) the standard deviation of the base-pairing free energies remains nonzero in the neighborhood of the melting temperature predicted by the homopolymeric perfect-matching model (see Figs. 3 and 10), and (3) the base-pairing free-energy is uncorrelated from one sequence position to the next. With these three features, the model melts so that the temperature derivative of the helix fraction is nondivergent: for each small increase of temperature, a finite fraction of the molecule's bps transit from the helix to coil states. We expect our characterization of the melting transition to apply to any DNA sequence consistent with these three features.

ACKNOWLEDGMENTS

C.B.R. acknowledges support from the Ford Foundation and from Harvard. We thank David Nelson for helpful comments on this work. We thank Dima Lukatsky for comments on the manuscript.

APPENDIX

1. Classification of phase transitions

In the statistical mechanical theory of macroscopic systems, an extended version [37] of the Ehrenfest classification scheme [42] can be used to describe the manner in which a macroscopic observable changes abruptly as a macroscopically controllable parameter is varied. The partition function $Z(T)$ of the system is a function of the macroparameters (among them is temperature T) and thus so is the free-energy density $f(T)=-RT \ln Z(T)/N$; R is the universal gas constant and N is proportional to the number of particles in the system. In the extended Ehrenfest scheme, we say that there is a kth-order phase transition at temperature T_{tr} if a kth derivative of the free-energy density with respect to any combination of the macroparameters is discontinuous (has a jump, which is a discontinuity of the simple kind, or a divergence, which is a discontinuity of the nonsimple kind [43]) but all derivatives of lower order are continuous (smooth). In this article, the term higher order means that the phase transition is not first order. Additionally, we might specify the kind of discontinuity with the terms jump kth order or divergent kth order. In general, first-order transitions are jump and higher-order transitions are divergent, so unless otherwise indicated, this is to be understood.

2. Derivation of H_n

To derive Eqs. (17)–(19), we first rearrange the terms in the Hamiltonian H_g [see Eq. (7)]. First, we set $u_g(\rho_i,\rho_{i-1}) = v_g^+(\rho_i)+v_g^-(\rho_i,\rho_{i-1})$, where

$$v_g^+(\rho_i) = \begin{cases} h - Ts & \rho_i > 0 \\ 0 & \text{otherwise}, \end{cases} \tag{A1}$$

$$v_g^-(\rho_i,\rho_{i-1}) = \begin{cases} -(h - Ts) & \rho_i > 0 \text{ and } |\rho_i - \rho_{i-1}| \geq 2 \\ 0 & \text{otherwise}. \end{cases} \tag{A2}$$

This decomposition of u_g introduces an error that makes a nonthermodynamic, i.e., negligible, contribution to H_g. We can then write

$$H_g[\rho] = \sum_{i=1}^{N} v_A(\rho_i,\rho_{i-1}) + \sum_{i=1}^{N} w_g(\rho_i) + g_F \cdot (N - \rho_N) \tag{A3}$$

where $v_A(\rho_i,\rho_{i-1}) = u_A(\rho_i,\rho_{i-1}) + v_g^-(\rho_i,\rho_{i-1})$ and $w_g(\rho_i) = v_g^+(\rho_i) + h_{\rho_i}^{\mathrm{hb}} + \delta g_{\rho_i}$ and

$$h_{\rho_i}^{\mathrm{hb}} = \begin{cases} h^{\mathrm{hb}} & \rho_i > 0 \\ 0 & \text{otherwise}, \end{cases} \tag{A4}$$

as defined in Sec. II A 2. Note that v_A is a sequence-position-independent two-body potential, whereas w_g is a sequence-position-dependent one-body potential. These w's are the pre-course-grained no-loop versions of the w's in Sec. II C. We can then write the n-replica partition function as

$$Z_g^n = \underset{\rho^1}{Tr} \cdots \underset{\rho^n}{Tr} \exp\left\{ -\sum_{\alpha=1}^{n} \beta H_A[\rho^\alpha] - \sum_{t=1}^{N} \beta w_g(t) \cdot \Theta^{1,\ldots,n}(t) \right\}, \tag{A5}$$

where $\alpha \in \{1, \ldots, n\}$ labels each replica, and

$$H_A[\rho^\alpha] = \sum_{i=1}^{N} v_A(\rho_i^\alpha, \rho_{i-1}^\alpha) + g_F \cdot (N - \rho_N^\alpha), \tag{A6}$$

and $\Theta^{1,\ldots,n}(t) = \Sigma_{\alpha=1}^{n} \theta^\alpha(t)$, where $\theta^\alpha(t)$ is the number of monomers at sequence position t in replica α. In

$$\overline{Z_g^n} = \underset{\rho^1}{Tr} \cdots \underset{\rho^n}{Tr} \exp\left\{ -\sum_{\alpha=1}^{n} \beta H_A[\rho^\alpha] \right\} \overline{\exp\left\{ -\sum_{t=1}^{N} \beta w_g(t) \cdot \Theta^{1,\ldots,n}(t) \right\}}, \tag{A7}$$

we average each δg_t over Gaussian fluctuations with mean 0 and standard deviation $g^{(2)}$ (see Sec. II B 1), obtaining Eq. (17). In the present notation, Eq. (20) may be formalized as

$$\langle \rho^\alpha | \rho^\gamma \rangle = \frac{1}{N} \sum_{t=1}^{N} \theta_t^\alpha \theta_t^\gamma. \tag{A8}$$

3. No ergodicity-breaking

a. Introduction

In the present work, we employ the replica method in order to compute the sequence-ensemble average of the free energy. The method requires an ansatz for the overlap matrix order parameter, e.g., see Fig. 7. Below, we make remarks in order to interpret our particular choice for this ansatz. In particular, we demonstrate that ergodicity-breaking is not to be expected on regions of the phase diagram where the replica method is capable of detecting it.

b. Heuristic argument

Consider the 1D polymer, with the Hamiltonian $H_g[\rho]$ in Eq. (7) at fixed finite N, some T, and $F=0$, with some particular sequence which is randomly generated at $g^{(1)} = g_\lambda^{(1)}(T), g^{(2)} = g_\lambda^{(2)}(T)$. We define a dynamics for the system as a Monte Carlo (MC) evolution. In this dynamics, we allow two types of "moves." In the first type, we attempt to move a randomly chosen monomer to an adjacent position. In the second type, we chose a position at random. If there is a monomer there, we attempt to remove it. If there is not a monomer there, we attempt to insert a monomer. Moves are accepted with a probability that is the ratio of Boltzmann weights for the system before and after the move. We assume that these microstate transition rules respect broken ergodicity if it exists.

We now define ergodicity breaking with respect to this dynamics using the language of Palmer [44]. We call the microstate space of the model Ω and define a component as a subset Ω^I of Ω such that the time scale for escape from region Ω^I is much longer than the time scale on which the model reaches equilibrium within Ω^I. We say that the model has components with absolute confinement if at each N we can identify a subset of components such that as N grows, the escape time scale for members of this subset increases unboundedly while the time-scale for reaching equilibrium within components in this set is bounded as N grows. We say that the model breaks ergodicity at parameters (T,F) if there is at least one absolutely confined component Ω^I that is a proper subset of Ω. Additionally, we say that the model exhibits thermodynamic ergodicity breaking if the absolutely confined component makes a thermodynamically important contribution to the partition function. Finally, we say that the model exhibits degenerate thermodynamic ergodicity breaking if there are multiple thermodynamically important and absolutely confined components. We imagine boundary regions between components as constituting large free-energy barriers. In what follows, we will only be concerned with thermodynamically important and absolutely confined components, which we will simply call components. We will only be concerned with ergodicity breaking of the thermodynamic type, which we will simply call ergodicity breaking. We focus our discussion on the number of components; if there are multiple components, the model has degenerate ergodicity breaking.

We now present expectations about the presence or absence of degenerate ergodicity-breaking at different temperatures. At high T, we expect each strand of the double-stranded model to have free coil statistics. We base this expectation on the existence of a finite temperature phase transition for a thermally melted homogeneous-sequence model with loop exponent $c>1$ [6]. Specifically, if we model a long random sequence of DNA with extremely high GC content as a homogeneous-sequence perfect-matching model, theory predicts a phase transition at some finite T_{GC}. Consequently, our model for the sequence of λ-DNA must melt at some $T<T_{GC}$. Thus, at high T, both the free energy per bp and the bound-state fraction should be zero, up to finite N fluctuations. If we call a particular arrangement of 1D monomers a loop structure, then a single loop structure will dominate. Thus we expect there to be a single component, and thus there is no degenerate ergodicity-breaking.

As we lower T, we expect there to be a value, call it T_c such that for $T<T_c$, the Gibbs equilibrium state [44] is dominated by one or many components that have a free-energy density (i.e., per bp) that is less than zero. Again, this is suggested by the behavior of homogeneous-sequence mod-

els; at low enough T our model should exist as a single helix, here the homogeneous-sequence model is accurate, and will have a negative free-energy density (see Fig. 5). In each component Ω^I we can say the following. The free-energy density is negative (and of similar magnitude) and so the distribution of loop lengths has exponential decay for large loops [see Eq. (4)]. Thus, there is some cutoff, ℓ_{cut}, above which there are essentially no loops. Also, the distribution of the length of the unzipped section has exponential decay. Thus the helix fraction must be nonzero. So we can identify T_c as the largest temperature below which the helix fraction of the Gibbs state is nonzero. The value T_c is the temperature above which the strands are separated, except perhaps for a thermodynamically unimportant number of helix-state bps. We can call T_c a critical temperature, in the sense of a divergent correlation length (typical loop size), if the melting transition is higher order (continuous).

We would like to know the number of these components $\{\Omega^I\}$ at low T. We hypothesize that there are many, i.e., that the model exhibits degenerate ergodicity breaking, and consider the consequences.

We divide the 1D polymer into subsystems by partitioning sequence-position space into intervals of size ℓ_{cut} and labeling them $b \in \{1,\ldots,M\}$. The microstate of subsystem b, γ_b, is a particular arrangement of monomers on interval b. We remove the interaction between subsystems and consider the nature of Ω under a Hamiltonian of the form $\Sigma_b h_b^{\text{ef}}(\gamma_b)$. Under the hypothesis that the whole system exhibits degenerate ergodicity breaking, each independent subsystem should also exhibit degenerate ergodicity breaking; we assume that free or bound boundary conditions of the 1D polymer does not affect the number of components. In the microstate space of subsystem b, call it ω_b, there are multiple components ω_b^I, each labeled by index I and having an escape time scale that grows with ℓ_{cut}. The thermodynamics of the collection of subsystems is dominated by $\cup_{\{I_b\}} \otimes \Pi_b \omega_b^{I_b} = \Omega_{\text{thermo}} \subset \Omega$.

We now consider a restricted version of model $\Sigma_b h_b^{\text{ef}}(\gamma_b)$ in which the microstate of the whole system, Γ, is confined to a subregion $\otimes \Pi_b(\omega_b^{I_b} \cup \omega_b^{J_b})$ such that each subsystem is restricted to the union of two particular components, $\omega_b^{I_b}$ and $\omega_b^{J_b}$, that are connected by the dynamics. We map this system to an Ising model by, for each subsystem b, assigning a spin state to each of the two components I_b and J_b; we consider all such assignments. For each subsystem b, the difference in the constrained free energy on $\omega_b^{I_b}$ and $\omega_b^{J_b}$, under $\Sigma_b h_b^{\text{ef}}(\gamma_b)$, maps to a random magnetic field in the Ising description. We now add ferromagnetic coupling j between neighboring spins with magnitude of order $\log \ell_{\text{cut}}$, obtaining a one-dimensional random-field Ising model (RFIM). By adding the ferromagnetic interaction, we have constructed a model ($d=1$ RFIM) that is more cooperative than the original model H_g restricted to $\otimes \Pi_b(\omega_b^{I_b} \cup \omega_b^{J_b})$. But, the $d=1$ homogeneous Ising model is more cooperative than the $d=1$ RFIM. Since the Landau argument [12] tells us that the $d=1$ homogeneous Ising model does not have an ordering transition at $T>0$, neither does the RFIM, and thus neither does H_g restricted to $\otimes \Pi_b(\omega_b^{I_b} \cup \omega_b^{J_b})$.

At very high T, there clearly cannot be multiple components. Because H_g restricted to $\otimes \Pi_b(\omega_b^{I_b} \cup \omega_b^{J_b})$ does not have an ordering transition at finite T, the component structure of the high-T macrostate persists to all $T>0$. Namely, for all $T>0$, there cannot be multiple components, i.e., there cannot be multiple regions in the space of system microstates $\otimes \Pi_b(\omega_b^{I_b} \cup \omega_b^{J_b})$ with an escape times that grow unboundedly with N. So, H_g on $\otimes \Pi_b(\omega_b^{I_b} \cup \omega_b^{J_b})$ does not exhibit degenerate ergodicity breaking at $T>0$.

We now consider another restricted version of model $\Sigma_b h_b^{\text{ef}}(\gamma_b)$ in which Γ is confined to the subregion $\otimes \Pi_b(\omega_b^{I_b} \cup \omega_b^{J_b}) \cup \omega_b^{K_b}$, where $\omega_b^{K_b}$ is dynamically connected to $\omega_b^{I_b}$ or $\omega_b^{J_b}$. Again, this restricted model is equivalent to a $d=1$ RFIM after we have assigned, for each subsystem b, two of the three components to one spin state, and the third component to the second spin state. Similar arguments as above show that H_g does not exhibit degenerate ergodicity breaking when restricted to $\otimes \Pi_b(\omega_b^{I_b} \cup \omega_b^{J_b}) \cup \omega_b^{K_b}$.

We continue this procedure until we arrive at the statement that H_g does not exhibit degenerate ergodicity breaking when restricted to $\otimes \Pi_b(\cup_{I_b} \omega_b^{I_b})$, which is Ω_{thermo}. This contradicts the hypothesis of many components in Ω under H_g. We conclude that for $T<T_c$, Ω contains only one component. Thus, there can be no degenerate ergodicity breaking for $T<T_c$.

c. Overlap distribution has a single peak

In replica theory [27], if a model does not exhibit degenerate ergodicity breaking, then the corresponding distribution of the 2-replica overlap exhibits a single peak. The converse is also true. This relationship can be understood with the following argument. For a typical sequence in the sequence ensemble, if one replica with this sequence has a single component, then two noninteracting replicas of this sequence form a combined system that has a single component [see Eq. (27)]. Consequently, the 2-replica overlap $\langle \rho^1 | \rho^2 \rangle$, which is an intensive property of the two replica system, has contributions from $O(N)$ independent random numbers (i.e., has the clustering property [27]) so that fluctuations are limited to the scale $1/N^{1/2}$ (see Note 1 of Chapter 1 of [45]).

4. Derivation of Z_{group}

a. Explicit expression for Z_{group}

We first obtain a formal expression for the entropy term of the cuRSB free energy [see Eq. (23)] at $F=0$, i.e., $g_F=0$. At large integer n and $m \in \{1,\ldots,n\}$, we apply the cuRSB constraints to the n-replica system H_n and turn off the term that couples the replicas by setting $g^{(2)}=0$. We obtain a system of n/m independent groups, each group composed of m perfectly correlated chains. We obtain the macrostate of this system as a function of θ and m by computing the right-hand side of

$$e^{S_n[Q^c]} \cong [Z_{\text{group}}(N,p,m,F)]^{n/m}, \tag{A9}$$

where $p=\lceil N\theta \rceil$, where $\lceil \cdots \rceil$ is the ceiling function, and the approximate equality becomes exact in the thermodynamic limit. For $\theta>0$, we have

$$Z_{\text{group}}|_{F=0} = \sum_{\ell_1=1}^{\infty} \cdots \sum_{\ell_p=1}^{\infty} \delta\left(N, \sum_{i=1}^{p} \ell_i\right) \prod_{i=1}^{p} u(\ell_i)^m, \quad \text{(A10)}$$

where $\delta(a,b)$ is the Kronecker delta, and

$$u(\ell) = \begin{cases} 1, & \ell = 1 \\ \dfrac{A}{\ell^c} e^{(h-Ts)/RT}, & \ell \geq 2 \end{cases} \quad \text{(A11)}$$

is the partition sum of a bound sequence constrained to length ℓ, where the free energy of the reference state has been shifted by $g^{(1)}$ units per loop so that the stacking energy and main-chain entropy of the helix-propagation free energy can be put in the energetic term of $E_n[\langle\rho|\rho\rangle]$ [see Eq. (19)]. We ignore the contribution to Z_{group} due to the unbound sequence, see the $N-\rho_N$ term in Eq. (7), because it the contribution is nonthermodynamic in magnitude.

For the computation of the fixed N partition function Z_{group}, we choose the Leontovich method, applied, e.g., by Lifshitz to the computation of the entropy of a Gaussian chain constrained to a particular realization of its density field [23]. In the first step, we allow N to fluctuate against a force represented by fugacity x. Thus, we construct the grand-canonical partition function

$$\Xi_{\text{group}}(x,p,m) = \sum_{N=p}^{\infty} x^N Z_{\text{group}}(N,p,m,F)\Big|_{F=0}. \quad \text{(A12)}$$

In the grand-canonical picture, there are p bound sequence groups—each group is a set of m bound sequences with lengths that fluctuate in unison. Each of the p sequence groups fluctuate in length independently of one another against force x. Thus, we have

$$\Xi_{\text{group}}(x,p,m) = [U(x,m,0)]^p, \quad \text{(A13)}$$

where $U(x,m,0)=\Sigma_{\ell=1}^{\infty} x^\ell u(\ell)$ is the grand partition function of a bound sequence.

In the Leontovich method, we obtain the fixed N free energy of Z_{group} by taking the free energy of Ξ_{group} and subtracting the energy of interaction with the force represented by x, obtaining

$$Z_{\text{group}}|_{F=0} = \frac{[U(\hat{x},m,0)]^p}{\hat{x}^N}, \quad \text{(A14)}$$

where the fugacity $\hat{x}$ is chosen so that, for each replica, the thermal average of the total sequence length is N. This is done by making $\log Z_{\text{group}}$ stationary with respect to $\hat{x}$.

We generalize $U(x,m,0)$ to $U(x,m,a) = x + \tilde{A}^m \Sigma_{\ell=2}^{\infty} x^\ell / \ell^{mc-a}$, and sum to obtain

$$U(x,m,a) = x + \tilde{A}^m[\phi(x, m\cdot c - a) - x], \quad \text{(A15)}$$

where $\tilde{A} = Ae^{(h-Ts)/RT}$ and $\phi(x,s) = \Sigma_{\ell=1}^{\infty} x^\ell / \ell^s$ is a special function known both as the polylogarithm function in applied mathematics, and as the ring function in polymer physics [46]. Useful properties of this function may be found in a paper by Truesdell [47].

b. Phase behavior and the intergroup overlap

We now consider n/m independent copies of the m-replica group discussed above, were n is a large integer and $m \in \{1,\ldots,n\}$. The macrostate of the $\frac{n}{m}$-group system is determined by Z_{group} and associated thermal averages. While preserving the visual image of the structure of this system, we continue m to the unit interval in all analytical expression for Z_{group}, the fugacity $\hat{x}$, and thermal averages. We thus obtain the "continued" macrostate of the $\frac{n}{m}$-group system.

We now calculate the overlap between replicas in distinct groups when the macrostate of the $\frac{n}{m}$-group system has been continued. For a given (θ,m,T), and $F=0$, the unbound sequence has finite (of order N^0) extent and does not contribute to the thermodynamics of the system; the unbound fraction is zero. Ignoring boundary effects, we assume that for each replica, the fraction of the time that a sequence position is occupied by a monomer is independent of sequence position. This assumption results in a thermal-average overlap of $q_u(\theta,m,T,F)|_{F=0} = \theta^2$.

5. "Minimizing" the variational free energy in the replica method

The stationarity equations result from zeroing the partial derivatives of βf_0, i.e.,

$$t_\theta(\theta,m,T) \equiv m\frac{\partial \beta f_0}{\partial \theta} = \beta g^{(1)} m - \frac{[\beta g^{(2)}]^2}{2} m^2[1-2\theta] - \log[U(\hat{x},m,0)],$$

$$t_m(\theta,m,T) \equiv \frac{m^2}{\theta}\frac{\partial \beta f_0}{\partial m} = -\frac{[\beta g^{(2)}]^2}{2} m^2[1-\theta] - \left\{\frac{mU'(\hat{x},m,0)}{U(\hat{x},m,0)} - \log U(\hat{x},m,0)\right\} - \frac{1}{\theta}\log \hat{x}, \quad \text{(A16)}$$

where $\tilde{A} = Ae^{g^{(1)}/RT}$, $U'(x,m,a) = \partial U(x,m,a)/\partial m$, and

$$\theta = \frac{U(\hat{x},m,0)}{U(\hat{x},m,1)}, \quad \text{(A17)}$$

which results from applying the derivative in Eq. (25) to the cuRSB free energy in Eq. (23). At large integer n, $m \in [1,n]$, so that in the limit $n \to 0$, the interval $[1,n]$ inverts to $m \in (0,1]$ and so minimization with respect to m becomes maximization. At each value of T, the numerical search for the values of θ and m that zeros t_θ and t_m is performed in terms of $\hat{x}$ and m. If we cannot find a stationary point, then we look for a nonstationary optimal point where t_θ is zero, t_m is positive, and $m=1$.

If we apply the uRSB scheme, see Sec. IV B, to the homopolymer case $g^{(2)}=0$ we obtain the single stationarity equation

$$t_\theta(\theta,m,T)\big|_{g^{(2)}=0}^{m=1} = \beta g^{(1)} - \log U(\hat{x},1,0) = 0. \quad \text{(A18)}$$

6. Minimizing the variational free energy of the homogeneous-sequence model

For the DNA model described in Sec. II A, we may obtain the thermodynamics of the homopolymer, $\delta g_t = 0$, by a route alternative to taking the special limits of the replica procedure. By the maximum-term method, we can obtain an equation for the thermodynamically dominant value of the helix fraction. We start with the constrained partition function

$$Z_{\text{homo}}(p,T,F)\big|_{F=0} = e^{-pg^{(1)}/RT} \sum_{\ell_1=1}^{\infty} \cdots \sum_{\ell_p=1}^{\infty} \delta\left(N, \sum_{i=1}^{p} \ell_i\right) \prod_{i=1}^{p} u(\ell_i), \quad \text{(A19)}$$

where the system is constrained to have p helix-state bps. As in Appendix, Sec. 4, the partition function of a bound sequence of length ℓ is

$$u(\ell) = \begin{cases} 1, & \ell = 1 \\ \dfrac{A}{\ell^c} e^{(h-Ts)/RT}, & \ell \geq 2, \end{cases} \quad \text{(A20)}$$

where the zero of energy is set to the free energy of the helix state of a bp. In the maximum-term method [7,48] the unconstrained partition function

$$Z_{\text{homo}}(T) = \sum_{p=0}^{N} Z_{\text{homo}}(p,T) \quad \text{(A21)}$$

is approximated by the dominant value of p; this approximation becomes exact in the thermodynamic limit N going to infinity.

The dominant value of p is determined by the minimization equations

$$\frac{\partial}{\partial \theta} f_{\text{homo}}(\theta,T) = 0,$$

$$\frac{\partial^2}{\partial \theta^2} f_{\text{homo}}(\theta,T) > 0, \quad \text{(A22)}$$

where

$$f_{\text{homo}}(\theta,T) = -\frac{RT}{N} \log Z_{\text{homo}}(p,T) \quad \text{(A23)}$$

and $p = \lceil N\theta \rceil$, where $\lceil \cdots \rceil$ is the ceiling function. To obtain an analytical expression for $f_{\text{homo}}(\theta,T)$, we proceed as in Appendix, Sec. 4, and apply the Leontovich method, obtaining

$$Z_{\text{homo}}(\theta,T) = e^{-pg^{(1)}/RT} \frac{[U(\hat{x},1,0)]^p}{\hat{x}^N}, \quad \text{(A24)}$$

where $U(x,1,0)$, is the partition function of a bound sequence in contact with a bp reservoir at fugacity x, and is defined in Appendix, Sec. 4. After Eq. (A24) is plugged into Eq. (A23), the stationarity equation in Eq. (A22) becomes

$$\frac{\partial}{\partial \theta} f_{\text{homo}}(\theta,T) = \beta g^{(1)} - \log U(\hat{x},1,0) = 0, \quad \text{(A25)}$$

where $\hat{x}$ satisfies $\theta = \partial f_{\text{homo}} / \partial h^{\text{hb}} = U(\hat{x},1,0)/U(\hat{x},1,1)$. This definition of the helix fraction is consistent with that used for the random-sequence model [see Eq. (25)]. Equation (A25) is the same as Eq. (A18) in Appendix, Sec. 5.

7. Self-consistency of the long-chain and perfect-matching approximations

According to Poland and Scheraga [38], the long-chain approximation is self-consistent, at zero force, when most of the denaturation occurs through the formation of internal loops rather the lengthening of the unbound sequence. This occurs when $N/L_{\text{helix}} > \sim 100$ (meaning greater than or of the order of). In our calculation, near T_c, $L_{\text{helix}} < 100$ (see Fig. 10), so for the λ-genome N/L_{helix} is at least 480. Thus, the infinite-chain approximation is valid.

Regarding the self-consistency of the perfect-matching approximation, we first note that near the melting temperature of our PM model, the typical length of a helix is $L_{\text{helix}} \approx 50$ at its lowest value (see Fig. 10). To claim self-consistency, we must evaluate the possibility of imperfectly matched helices of this length scale. We define an imperfect match as a base-pairing interaction between bps from opposite strands that are not complementary in the native (ground) state. Since we are describing the λ genome as random and uncorrelated, the probability that a section of length $L_{\text{helix}} \approx 50$ from one strand has an imperfect match, at a given position in the other strand, is $1/4^{50} \approx 10^{-30}$. Since the genome is 5×10^4 bps, and if we assume that the genome is composed of 1000 helical sections of length $L_{\text{helix}} \approx 50$, then the expected number of imperfect matches is $1 \times 10^{-30} \cdot 1000 \cdot 1000 \approx 10^{-24}$, so it is unlikely that a single imperfect match of length L_{helix} will be found in the entire genome. Furthermore, short helices and incompletely matched helices—containing a non-Watson-Crick bp—are free energetically penalized by the loop termination factor A. Thus, the perfect-matching approximation is self-consistent for uncorrelated random sequences.

[1] J. D. Watson and F. H. C. Crick, Nature (London) **171**, 737 (1953).
[2] O. Gotoh, Y. Husimi, S. Yabuki *et al.*, Biopolymers **15**, 655 (1976).
[3] R. D. Blake and P. V. Haydock, Biopolymers **18**, 3089 (1979).
[4] C. Danilowicz, Y. Kafri, R. S. Conroy, V. W. Coljee, J. Weeks, and M. Prentiss, Phys. Rev. Lett. **93**, 078101 (2004).
[5] J. Liphardt, B. Onoa, S. B. Smith *et al.*, Science **292**, 733

(2001).
[6] D. Poland and H. A. Scheraga, J. Chem. Phys. **45**, 1456 (1966).
[7] D. Poland and H. A. Scheraga, *Theory of Helix-Coil Transitions in Biopolymers* (Academic, New York, 1970).
[8] G. W. Lehman and J. P. McTague, J. Chem. Phys. **49**, 3170 (1968).
[9] S. Lifson, Biopolymers **1**, 25 (1963).
[10] S. Lifson and G. Allegra, Biopolymers **2**, 65 (1964).
[11] D. Poland and H. A. Scheraga, Biopolymers **7**, 887 (1969).
[12] L. D. Landau and E. M. Lifshitz, *Statistical Physics, Part I* (Pergamon, Oxford, 1980).
[13] J. H. Gibbs and E. A. DiMarzio, J. Chem. Phys. **30**, 271 (1959).
[14] T. L. Hill, J. Chem. Phys. **30**, 383 (1959).
[15] P. Yakovchuk, E. Protozanova, and M. D. Frank-Kamenetskii, Nucleic Acids Res. **34**, 564 (2006).
[16] Y. Kafri, D. Mukamel, and L. Peliti, Phys. Rev. Lett. **85**, 4988 (2000).
[17] E. A. Mukamel and E. I. Shakhnovich, Phys. Rev. E **66**, 032901 (2002).
[18] B. H. Zimm and J. K. Bragg, J. Chem. Phys. **31**, 526 (1959).
[19] R. D. Blake and S. G. Delcourt, Nucleic Acids Res. **26**, 3323 (1998).
[20] R. Blossey and E. Carlon, Phys. Rev. E **68**, 061911 (2003).
[21] M. E. Fisher, J. Chem. Phys. **45**, 1469 (1966).
[22] R. D. Blake, J. W. Bizzaro, J. D. Blake *et al.*, Bioinformatics **15**, 370 (1999).
[23] A. Y. Grosberg and A. R. Khokhlov, *Statistical Physics of Macromolecules* (American Institute of Physics, New York, 1994).
[24] D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University Press, New York, 1987).
[25] D. L. Daniels, F. Sanger, and A. R. Coulson, Cold Spring Harbor Symp. Quant. Biol. **47**, 1009 (1982).
[26] P. N. Borer, B. Dengler, J. Ignacio Tinoco *et al.*, J. Mol. Biol. **86**, 843 (1974).
[27] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[28] J. Applequist and V. Damle, J. Chem. Phys. **39**, 2719 (1963).
[29] J. Applequist and V. Damle, J. Am. Chem. Soc. **87**, 1450 (1965).
[30] D. K. Lubensky and D. R. Nelson, Phys. Rev. E **65**, 031917 (2002).
[31] Y. Kafri and A. Polkovnikov, Phys. Rev. Lett. **97**, 208104 (2006).
[32] S. Srebnik, A. K. Chakraborty, and E. I. Shakhnovich, Phys. Rev. Lett. **77**, 3157 (1996).
[33] J. A. Schellman, C. R. Trav. Lab. Carlsburg **29**, 230 (1955).
[34] Y. Kafri, D. Mukamel, and L. Peliti, Eur. Phys. J. B **27**, 135 (2002).
[35] C. Danilowicz, V. W. Coljee, C. Bouzigues *et al.*, Proc. Natl. Acad. Sci. U.S.A. **100**, 1694 (2003).
[36] G. Giacomin and F. L. Toninelli, Phys. Rev. Lett. **96**, 070602 (2006).
[37] A. B. Pippard, *Elements of Classical Thermodynamics* (Cambridge University Press, Cambridge, 1957).
[38] D. Poland and H. A. Scheraga, Physiol. Chem. Phys. **1**, 389 (1969).
[39] M. Peyrard and A. R. Bishop, Phys. Rev. Lett. **62**, 2755 (1989).
[40] T. Dauxois, M. Peyrard, and A. R. Bishop, Phys. Rev. E **47**, R44 (1993).
[41] D. Cule and T. Hwa, Phys. Rev. Lett. **79**, 2375 (1997).
[42] G. Jaeger, Arch. Hist. Exact Sci. **53**, 51 (1998).
[43] W. Rudin, *Principles of Mathematical Analysis* (McGraw-Hill, New York, 1976).
[44] R. G. Palmer, Adv. Phys. **31**, 669 (1982).
[45] G. Parisi, *Statistical Field Theory* (Perseus Books, Reading, Massachusettes, 1998).
[46] H. Jacobson and W. H. Stockmayer, J. Chem. Phys. **18**, 1600 (1950).
[47] C. Truesdell, Ann. Math. **46**, 144 (1945).
[48] T. L. Hill, *An Introduction to Statistical Thermodynamics* (Dover, New York, 1986).
[49] S. B. Smith, Y. Cui, and C. Bustamante, Science **271**, 795 (1996).