Looping probability of random heteropolymers helps to understand the scaling properties of biopolymers

Y. Zhan and L. Giorgetti[*]
Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, CH-4058 Basel, Switzerland

G. Tiana[†]
Center for Complexity and Biosystems and Department of Physics, Università degli Studi di Milano and INFN,
via Celoria 16, 20133 Milano, Italy

(Received 18 April 2016; published 2 September 2016)

Random heteropolymers are a minimal description of biopolymers and can provide a theoretical framework to the investigate the formation of loops in biophysical experiments. The looping probability as a function of polymer length was observed to display in some biopolymers, like chromosomes in cell nuclei or long RNA chains, anomalous scaling exponents. Combining a two-state model with self-adjusting simulated-tempering calculations, we calculate numerically the looping properties of several realizations of the random interactions within the chain. We find a continuous set of exponents upon varying the temperature, which arises from finite-size effects and is amplified by the disorder of the interactions. We suggest that this could provide a simple explanation for the anomalous scaling exponents found in experiments. In addition, our results have important implications notably for the study of chromosome folding as they show that scaling exponents cannot be the sole criteria for testing hypothesis-driven models of chromosome architecture.

DOI: 10.1103/PhysRevE.94.032402

I. INTRODUCTION

Most biological molecules are polymers, and the formation of contacts between monomers which are not close along the chain often plays an important biological role. For example, in the nucleus of mammalian cells, the encounter of an enhancer and a gene promoter that can be millions of base pairs away along the chromatin fiber is often necessary for the expression of the gene [1]. In the case of proteins, the formation of noncovalent interactions between distant amino acids is, in many cases, among the first steps in the folding process [2].

There are several experimental techniques to study, either directly or indirectly, the formation of contacts between pairs of monomers as a function of their distance N along the polymeric chain. Arguably, when N is large enough, the detailed chemistry of the system loses importance and one can highlight its more general physical properties. The looping probability of peptides with repeated AGQ sequence, measured by Förster resonance energy transfer, displays a power law with exponent 1.55 in water and 1.7 in urea and guanidine [3]. The folding rate of proteins, measured by stopped-flow experiments, was shown to correlate with the (rescaled) average value of N of pairs of amino acids which are in contact in the native state [4]. In long RNA chains the contact probability displays an exponent $\beta \approx 1$ [5]. In the case of chromosome folding, a class of biochemical techniques collectively known as chromosome conformation capture (3C) makes it possible to measure contact probabilities along the chromatin fiber following chemical cross linking of nuclei [6]. In human and mouse chromosomes, these techniques revealed that the looping probability between chromosomal loci depends on N as a power law $N^{-\beta}$ with exponent $\beta \approx 1$ above the 10^6-base-pairs scale [7] and even lower at a smaller scale [8]. Importantly, these scaling exponents have been used to derive and test models regarding the mechanisms that could give rise to the peculiar folding patterns observed in the genome (see Sec. VIII below). It is therefore important to understand if anomalous scaling exponents necessarily arise from specific model-specific mechanisms or can rather emerge as general properties of biopolymers.

The simplest theoretical framework to describe the contact formation in a biopolymer at equilibrium as a function of N is that of two interacting monomers linked by a homopolymer. One can employ a two-state description of the system, assuming that the formation of the contact between the two ends does not change the density of the polymer. In this case, if $\epsilon < 0$ is the energy gain of the system upon formation of the contact, the associated probability can be approximated as

$$c(N) = \frac{\exp(-\epsilon/T)}{g(N) + \exp(-\epsilon/T)}, \tag{1}$$

where $g(N)$ is the density of state of the system displaying the contact with respect to the unbound state. Its shape depends on the properties of the linking homopolymer. If this can be regarded as an ideal chain, then $g(N) = N^{3/2}$; if it is a random coil due to the repulsion between its elements, $g(N) = N^{9/5}$, while it is constant in a globule [9]. In the limit of large N one then expects a scaling law of the type $c \propto N^{-\beta}$, with $\beta = 0, 1/2$, or $9/5$, as discussed above. The scaling exponents found for repeat peptides [3] lie between those expected for an ideal chain and a random coil. In the case of chromatin, the anomalous exponent $\beta \leqslant 1$ found in experiments is not compatible with the above model and several mechanisms have been evoked to explain this finding: nonequilibrium effects similar to what observed in the "crumpled globule" state [10,11], looping interactions mediated by soluble DNA-binding molecules [12] or energy-driven mechanisms

[*] luca.giorgetti@fmi.ch
[†] guido.tiana@unimi.it

such as loop extrusion by DNA-bound protein complexes [8,13].

However, in most cases, the monomers which build polymers of biological interest are chemically heterogeneous, and the homopolymeric assumption is questionable. The problem we would like to address in the present work is the role of heterogeneous interactions in determining the scaling properties of the contact probability between monomers. Specifically, we study the looping probability of random heteropolymers [14], regarding them as a minimal model for biomolecules.

To investigate this problem, we use a simple model, in which the polymer is described as a chain of beads connected by rigid links. Pairs of beads interact through a spherical-well potential with a hard core of radius r_H, a width r, and a depth B_{ij} which depends on the specific pair. For the sake of generality, we considered the energies B_{ij} as quenched stochastic variables, defined by a Gaussian distribution. In this way we are not focusing on a particular kind of biopolymer, but we are looking for the general properties which arise only because of the heterogeneity of the interactions.

Operatively, we investigated the equilibrium contact probability of heteropolymeric chains by means of numerical simulations. The stochasticity of the interaction energies was modeled by generating several realizations of the set of Gaussian variables, and for each of them carrying out a conformational sampling. This approach poses the problem of averaging the results of the samplings over the quenched energies. The contact probability itself does not result to be a self-averaging quantity, and consequently its average over the realizations of the quenched variables B_{ij} is poorly informative [15]. In Sec. IV we discuss under which conditions the average of quantities associated with the contact probability are informative.

Another problem one has to face is that the conformational sampling of disordered systems is computationally cumbersome, due to the roughness of the associated energy landscape. There are several computational techniques based on the multicanonical ensemble which, sampling the system simultaneously at different temperatures, facilitate conformational sampling [16,17]. However, they rely on the choice of a set of temperatures that are optimized to enhance diffusion in the temperature space. This set is not self-averaging, and consequently requires a manual fine tuning for each realization of the quenched variables. This is impractical if one wants to collect results from enough replicas to calculate reliable averages. To solve this problem in an automatic way, we made use of an adaptive simulated-tempering scheme developed in Ref. [18].

In Sec. II we describe a consistent theoretical framework which is necessary to study quantitatively the looping probability in heteropolymers. This framework is applied to a simple model of random heteropolymers, described in Sec. III. In Sec. IV we analyze the main obstacle one finds in a naive derivation of the scaling properties of the looping probability in polymers with a disordered interaction. We then analyze in a consistent way the looping properties (in Sec. V) and the related compactness (in Sec. VI) of a set of heteropolymers as a function of their length. Then, in Sec. VII, we perform a similar analysis on the scaling properties describing the formation of loops in the different segments of a fixed-length polymer, a case which is relevant for recent experiments on chromosome systems [6–10,12,13,19], as described in Sec. VIII. We then discuss the consequences of the model in Sec. IX and draw some conclusions in Sec. X.

II. THE THEORETICAL FRAMEWORK

In order to find the most appropriate way of calculating the scaling properties of the looping probability of a random heteropolymer, one can use a two-state model. One can assume that the bound and unbound states display, respectively, energies $E_1 + \epsilon$ and E_2, where E_1 and E_2 are quenched random variables regarded as the sum of the internal contact energies of the chain, while ϵ is the interaction energy between the ends of the chain. Further assuming that E_1 and E_2 are uncorrelated and that the two states have the same density, the central-limit theorem suggests that

$$p(E_1) = p(E_2) = \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left[-\frac{(E_{1,2} - N\epsilon_0)^2}{2N\sigma^2}\right], \quad (2)$$

where N is the length of the chain, ϵ_0 the average interaction between the monomers, and σ their standard deviation. We define $\Delta E \equiv E_1 - E_2$ and assume a density of states of the unbound state with respect to the looped state in the form of a power law of the kind N^β. Thus, the entropy difference is $\beta \ln N$ and the free-energy difference between the two states is given by

$$\Delta F = \Delta E + \epsilon + T\beta \ln N, \quad (3)$$

where ΔE is a stochastic variable with distribution

$$p(\Delta E) = \frac{1}{\sqrt{4\pi N\sigma^2}} \exp\left(-\frac{\Delta E^2}{4N\sigma^2}\right). \quad (4)$$

According to this model, the variability of the looping free energy, and consequently of the looping probability, at a given value of N is due to the variability of the internal energy difference ΔE. In other words, ΔE plays the role of the quenched disorder affecting the looping free energy defined as a function of N. The associated probability can be obtained by inverting Eq. (3) and substituting it into Eq. (4), that is,

$$p(\Delta F) = \frac{1}{\sqrt{4\pi N\sigma^2}} \exp\left[-\frac{(\Delta F - T\beta \ln N - \epsilon)^2}{4N\sigma^2}\right]. \quad (5)$$

This probability can be maximized with respect to β and ϵ according to a maximum-likelihood principle, obtaining

$$\beta = -\frac{1}{T}\frac{\sum_N \frac{1}{N}\sum_N \frac{\ln(N)\Delta F}{N} - \sum_N \frac{\ln(N)}{N}\sum_N \frac{\Delta F}{N}}{\sum_N \frac{1}{N}\sum_N \frac{\ln^2(N)}{N} - \left[\sum_N \frac{\ln(N)}{N}\right]^2}, \quad (6)$$

formally identical to the expression of a weighted linear regression.

From the simulations (or from a set of experiments) one can calculate the free-energy difference ΔF from the contact probability

$$\Delta F = -T \ln\left[\frac{c}{1-c}\right] \quad (7)$$

and use Eq. (6) to obtain β from a linear regression of F versus $\ln N$ with weights N^{-1}. This weighting is a consequence of the extensivity of the energy of the chain and has as consequence that larger-N points contribute less to the determination of β.

III. THE COMPUTATIONAL MODEL AND ALGORITHM

In the present work heteropolymers are described as chains of beads linked by rigid links of length $a = 1$ (which sets the length scale of the system). Beads interact with a two-body potential $U = \sum_{i<j} u_{ij}$, where the two-body terms are defined as

$$u_{ij} = \begin{cases} +\infty & \text{if } |r_i - r_j| < r_H, \\ B_{ij} & \text{if } r_H \leqslant |r_i - r_j| < r, \\ 0 & \text{if } |r_i - r_j| \geqslant r. \end{cases} \tag{8}$$

The B_{ij} are quenched stochastic energies, distributed according to a Gaussian function with zero mean and standard deviation $\sigma_B = 1$ (which sets the energy scale of the system). In the calculations, we chose [20] $r_H = 0.6$ and $r = 1.5$ (in units of a). The equilibrium properties of random heteropolymers are studied generating 500 realizations of the random interactions B_{ij}, sampling the conformational space of each of them, and performing averages over the realizations of the random interactions as described in Sec. IV below.

Conformational samplings are carried out with an iterative simulated-tempering algorithm [18]. It is based on a Metropolis scheme in which elementary moves are flips of single beads and pivot moves (see Ref. [21], where the code used for the simulations is described). A simulated tempering is then applied in which the temperatures $\{T_i\}$ and the free-energy factors $\{g_i\}$ which define the simulated tempering [17] are adjusted during the simulation to optimize the diffusion of the temperature, from scratch in each realization of the interaction matrix. Specifically, the simulation starts with a plain Metropolis at high temperature $T_0 = 2$ (in units of σ, setting Boltzmann constant to 1). From the distribution of energies calculated from this sampling, the ideal values of T_1 and g_1 to have a temperature-exchange rate of 0.1 are estimated and a simulated tempering over these two temperatures is carried out. A weighted-histogram algorithm is then applied to obtain the distribution of energies from the energy distributions obtained so far, and a further pair T_2 and g_2 is added to the tempering. This procedure is iterated until the target temperature T is reached. A set of rules is also applied in the case where actual exchange rates depart from the predicted ones, as described in detail in Ref. [18]. An example of this procedure results in a sampling of different temperatures as that displayed in Fig. 1, which makes it possible to calculate equilibrium averages of polymers up to ~10^2 monomers.

IV. THE SELF-AVERAGING ISSUE

The average $\overline{x}$ of a conformational property x of the random heteropolymer over the quenched stochastic energies provides valuable information only if the associated standard error σ_x is small, namely if the quantity is self-averaging [15]. In the thermodynamic limit, this corresponds to the condition

$$\xi_x \equiv \frac{\sigma_x}{|\overline{x}|} \to 0. \tag{9}$$

Usually extensive properties are self-averaging [22], while intensive properties, probability distributions, and partition

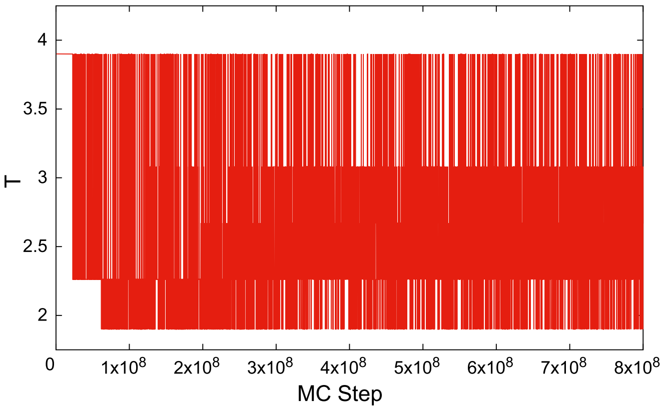

FIG. 1. An example of evolution of the temperatures in the self-adjusting simulated-tempering simulation.

functions are not. Thus, we do not expect $c(N)$ to be self-averaging, and in fact ξ_c is quite large, increasing above 1 quite fast as a function of N at low temperatures [cf. Fig. 2(a)]. This is the reason why in the context of disordered systems one focuses the attention on free energies. However, in the present case we are considering a free-energy difference between two states of the system, which is expected to scale as $\ln N$ according to Eq. (3). The associated self-averaging parameter thus scales as $\chi_{\Delta F} \sim N^{1/2}/\ln N$, which has a nonmonotonic behavior as a function of N, eventually diverging in the thermodynamic limit, although not very fast [cf. Fig. 2(b)].

Thus, strictly speaking, ΔF is not self-averaging. Nor it is any quantity which can be derived by the contact probability c. However, if one is interested in finite systems of the typical size of biopolymers, a sufficient request is that the variability of ΔF associated with the disorder is smaller than its average; that is, $\xi_{\Delta F} \ll 1$ in a specified interval of N.

Equation (5) suggests that the variability of ΔF over the quenched disorder should follow

$$\xi_{\Delta F} = \frac{2\sigma N^{1/2}}{|\epsilon + T\beta \ln N|} \tag{10}$$

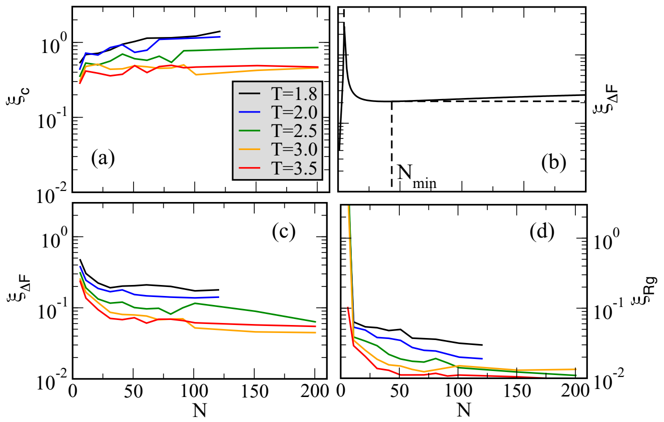

FIG. 2. (a) The relative error ξ_c associated with c; (b) a sketch of the theoretical behavior of $\xi_{\Delta F}$ according to Eq. (10); (c),(d) the relative error ξ calculated for ΔF and for the gyration radius R_g, respectively.

and consequently display a divergence at $N_{\text{div}} = \exp[-\epsilon/T\beta]$ and a minimum at $N_{\text{min}} = \exp[2 - \epsilon/T\beta]$, diverging at large N [cf. Fig. 2(b)]. Thus, we can expect ΔF to be representative of a typical realization of the disordered interactions if $N > N_{\text{div}}$ and $N \sim N_{\text{min}}$.

In Fig. 2(c) is plotted the value of $\xi_{\Delta F}$ at different temperatures as a function of the length N of the chain in semilog scale, calculated over 500 realizations of the random interactions. For each temperature we show the points up to the largest value of N for which we can guarantee the correct equilibration of the simulated-tempering algorithm. In the studied range of N, the calculated $\xi_{\Delta F}$ is decreasing, thus suggesting that $N_{\text{div}} < N < N_{\text{min}}$. Moreover, already for $N > 10$ the $\xi_{\Delta F}$ assumes small values, indicating that the standard error on ΔF is of the order of a few percent of the mean. That is, except for very short chains, the average of ΔF over the stochastic interactions are representative of their typical values. A similar behavior is observed for the gyration radius R_g of the polymer [see Fig. 2(d)].

V. SCALING OF THE FREE ENERGY ASSOCIATED WITH THE LOOPING PROBABILITY

From the same simulations used to estimate the degree of self-averageness, we calculated the values of $\overline{\Delta F}$ as a function of N, in order to estimate its scaling properties.

The linear fit of ΔF as a function of $\ln N$ is displayed in Fig. 3 for simulations carried out at different temperatures. The linear fit appears good at $T > 2.0$ and seems to worsen at lower temperatures. In particular, at $T \leqslant 2.0$ a power-law behavior applies up to $N{\approx}60$, while $\overline{\Delta F}$ appears weakly dependent on N above ≈ 60, similarly to the behavior of a collapsed globule in a homopolymer.

Interpreting Eq. (5) as the likelihood of observing a value of ΔF in a chain of specified length, the quality of the linear fit can be expressed in terms of the average log-likelihood, that is nothing else but

$$\chi^2 = \frac{1}{Z_N}\sum_n^N \frac{(\overline{\Delta F}(n) - \epsilon - T\beta \ln n)^2}{n\sigma^2}, \qquad (11)$$

where N is the length of the longest chain considered in the fit and $Z_N = \sum_n^N (n\sigma^2)^{-1}$. The values of χ^2 as a function of N are reported in the inset of Fig. 3. The fits of the points at $T > 2.0$ display a constant or decreasing χ^2 of the order of 10^{-2}, while at lower temperatures it increases with N. However, even at low temperatures the value of χ^2 remains lower than 1 for all the N studied, indicating that the fitting line matches the points within their error bars.

This is a result of the fact that both the estimation of β and the quantification χ^2 of the error of the fit emphasize smaller polymers because for them the variability of ΔF due to the disordered interactions is smaller. In the case of longer polymers, $\overline{\Delta F}$ seems to become independent on N, but at the same time it becomes less and less representative of a typical heteropolymer. In fact, even if $\overline{\Delta F}$ were constant at large N, the leading term of Eq. (11) would be $\chi^2 \sim N^{-1}\sum_n \ln^2 n/n$; approximating the sum with an integral gives $\chi^2 \sim \ln^3 N/N$, which vanishes at large N. In other words, it is the small-N slope that determines β, because at large N the free energy is dominated by the disorder. If the small-N scaling properties are due to finite-size effects, these will thus dominate the results even when considering longer chains.

The values of the parameter β obtained from the fits at different temperatures are reported as solid circles in Fig. 4. At high temperature ($T = 3.5$) the scaling exponent β converges to 2.06, which is comparable with the value 2.10 ± 0.15 obtained numerically for self-avoiding walks in three dimensions [23] and somewhat larger than the theoretical

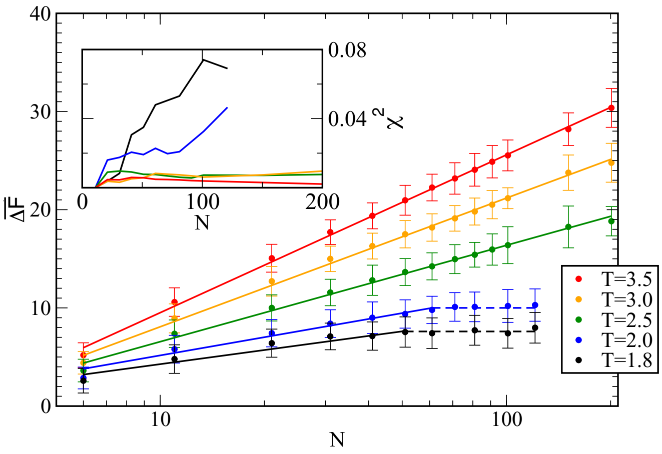

FIG. 3. The average value of ΔF as a function of N, the latter displayed in a logarithmic scale. For each value of N, 500 realizations of the disordered interaction are simulated. The points are fitted according to Eq. (6), and the corresponding line is drawn in the figure. (Inset) The χ^2 associated with the fits calculated up to length N.

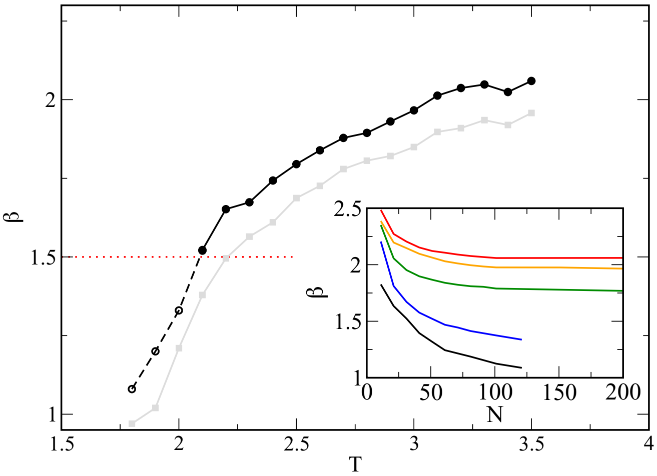

FIG. 4. The exponents β obtained using Eq. (6) at different temperatures from the fits of the simulated data up to the largest polymer we could equilibrate (circles). As a reference, the dotted curve indicates the exponent 3/2 expected for an ideal chain. Empty circles indicate the exponents below the θ point, strongly affected by finite-size effects. The gray squares indicate the exponents found in a fit of $\ln c$ versus $\ln N$. (Inset) The exponent calculated from fits up to length N.

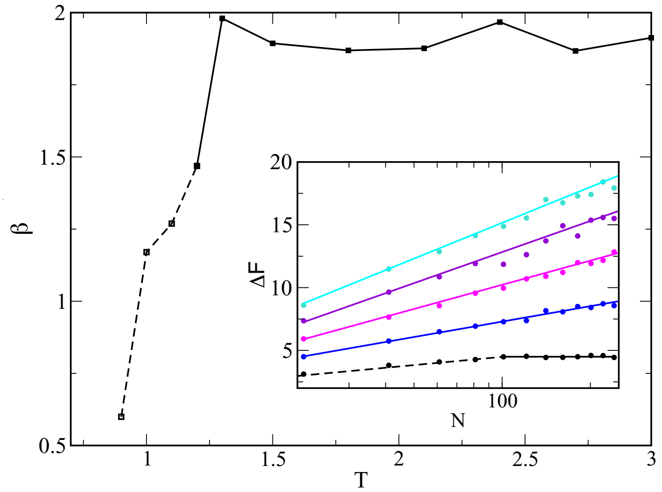

FIG. 5. The scaling exponent β calculated for a homopolymer (i.e., $\epsilon_0 = -0.1, \sigma = 0$) as a function of temperature T. Open symbols indicate the exponents associated with finite-size behavior (cf. dashed line in the inset). (Inset) The binding free energies whose fits were used to obtain the scaling exponents (the different sets correspond, starting from above, to $T = 2.1$, $T = 1.8$, $T = 1.5$, $T = 1.2$, and $T = 0.9$).

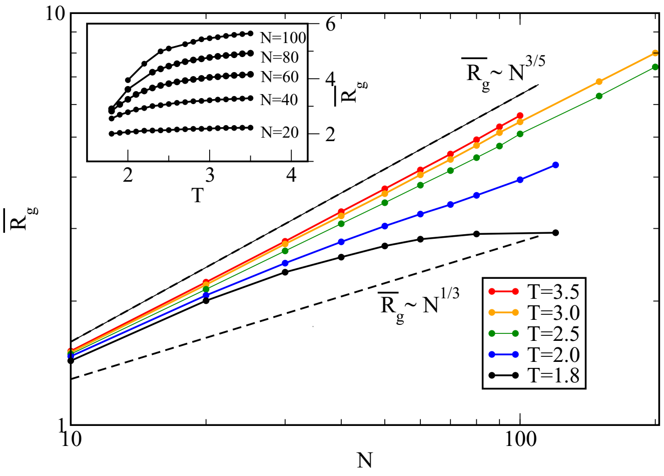

FIG. 6. The average gyration radius $\overline{R_g}$ at different temperatures as a function of the length of the chain plotted in log-log scale. As a reference, we indicate with dashed lines the $N^{3/5}$ curve expected for a random coil and the $N^{1/3}$ curve expected for a globule. (Inset) The value of R_g as a function of temperature for different lengths N.

result 9/5 obtained by de Gennes solving a zero-dimensional Ising model [9].

As the temperature is decreased, β decreases continuously to the value $\beta = 3/2$ typical of the θ point at $T \approx 2.0$. This plot is markedly different from that of a homopolymer, in which case only two kinds of exponents are expected, associated with the coil state and the ideal behavior at the θ point. In fact, the exponents found from numerical simulations of homopolymers of comparable size are displayed in Fig. 5. Moreover, even a random heteropolymer in the coil or θ state in the limit of short interaction range is expected to display the same exponents of the homopolymer, superposed to an exponential cutoff [24].

Below the θ point the fit gives exponents $1 \lesssim \beta \lesssim 1.5$ (cf. empty circles in Fig. 4). Since the small-N contribution dominates due to the dependence on N of the denominator at the exponent of Eq. (5), the exponents β seem to converge to a N-independent value, different from zero, even below the θ point (cf. inset of Fig. 4).

The scaling of $\overline{\Delta F}$ below the θ point with exponents lower than 3/2 is a finite–size effect, also present in homopolymers (cf. Fig. 5). This is a consequence of the fact that if the polymer is too short, it is not able to define a bulk volume, necessary for the looping entropy to lose its dependence on N, but its volume essentially coincides with its surface. The order of magnitude of N below which this effect takes place is found by $4\pi R^2 2r_H = 4/3\pi R^3$, with $R = r_H N^{1/3}$ in a globule, that is, $N = 6^3 \approx 10^2$, in agreement with what shown in Fig. 5.

Often a simple regression of $\ln c$ versus $\ln N$ was applied to the analysis of the scaling properties of the contact probability [3] of biopolymers. This is more difficult to justify theoretically than the fit described in Sec. II. Anyway, the results of such a fit are displayed with gray squares in Fig. 4. The resulting exponents are slightly smaller than those obtained with the two-state model described above, but in this case the (unweighted) χ^2 of the fit ranges from 0.2 at high temperature to ≈1.8 at low temperature. At variance with the the weighted fit described above, in this case the χ^2 of the fit, as well as the value of the exponents, depend on the specific range of N employed in the simulations.

VI. COMPACTNESS OF THE POLYMER

In order to compare the exponents β found for the random heteropolymer with those known from the theory of homopolymers, it is interesting to understand whether the polymer is, at the different temperatures studied above, in a globular or in a coil state. This problem is well-defined because the resulting thermal average R_g of the gyration radius is self-averaging (see Sec. IV), and consequently we can study its average $\overline{R_g}$ over the realizations of the disordered interaction. On the other hand, it is complicated by the small size of the system, while a globule-coil phase transition is defined, strictly speaking, only for an infinitely long polymer.

The average value of $\overline{R_g}$ as a function of N is displayed in log-log scale in Fig. 6 at different temperatures. For $T \geqslant 3.0$ the curves overlap almost perfectly to each other, with a slope of ≈3/5, that of a random coil in the case of a homopolymer. This is not unexpected, since at high temperature the heterogeneity in the interactions within the chain becomes negligible with respect to T, and the heteropolymer behaves effectively as a homopolymer.

For temperatures $T < 3.0$ the slope of $\ln \overline{R_g}$ versus $\ln N$ decreases and reaches 1/2, the value that homopolymers display at the θ point, at $T \approx 2.1$. If one decreases the temperature further, the curve is no longer linear in the range of N under consideration. This is likely to be a finite-size effect, since the gyration radius has to grow at least as $N^{1/3}$, corresponding to a fully compact structure.

The decrease of $\overline{R_g}$ as a function of T can also be visualized directly in the inset of Fig. 6 for each value of

N. A clear transition in $\overline{R_g}$ cannot be seen at any value of N. At large values of N, where transitions are expected to be sharper, we are not able to equilibrate the lowest temperatures, corresponding to the compact phase. Consequently, we are not able to highlight clearly a globule-coil transition, similar to that of homopolymers.

The clearest set of data is that calculated for $N = 60$. At $T = 1.8$ the mean gyration radius is 2.7, not far from that of a maximally compact globule, that is $N^{1/3}r_H = 2.4$. At $T = 2.0$ the value of $\overline{R_g}$ is 3.2, close to that associated with that of an ideal chain, that is, $0.41N^{1/2} = 3.18$. Anyway, the curve increases smoothly from the more compact to the more elongated conformations.

Summing up, the random heteropolymer displays at high temperature properties of the radius of gyration similar to those of homopolymers, including a θ point at which the size of the heteropolymer scales as that of an ideal chain. At lower temperatures, in the range of lengths we could equilibrate, the size is dominated by finite-size effects.

VII. SCALING PROPERTIES WITHIN A FIXED-LENGTH CHAIN

Sometimes the experimental data to analyze are not the looping probability of polymers of different lengths, but the looping probabilities of the various segments, of different lengths, within a given polymer. This is, for example, the case of chromosome conformation capture experiments on the chromatin fiber [7]. The standard way of extracting the scaling exponent is a linear regression of $\ln c(i,j)$ versus $\ln|i-j|$ of the whole set of data, where $|i-j| \leqslant N$ is the length of the segment starting at monomer i and ending at monomer j of the N-bead polymer. It was also suggested that fitting c versus n is a better strategy [25]; this is, however, unwise in the case of heteropolymers, because of the lack of self-averaging of c (cf. Sec. IV).

In any case, if the heterogeneity in the looping probability at fixed intermonomer linear distance is due to the variability of the interactions, the correct way of extracting the scaling behavior is similar to that described in Sec. II. As in the case of heteropolymers of different lengths, one can define a looping free energy ΔF [cf. Eq. (7)] and develop calculations similar to those which lead to Eq. (6). However, now Eq. (3) depends on $|i-j|$ instead of N; that is,

$$\Delta F(i,j) = \Delta E + \epsilon + T\beta' \ln|i-j|, \qquad (12)$$

where we define the scaling exponent as β' to distinguish it from that of varying-size polymers. Now Eq. (2) is still valid, but N is fixed. The result is that, according to this model, β' should be obtained by an unweighted linear regression of $\Delta F(i,j)$ versus $\ln|j-i|$. Here, the main difference with Eq. (6) is the lack of weights in the sum.

As one is usually interested in the scaling properties of any two monomers as a function of their distance n along the chain, and not of two specific monomers i and j (which is, anyway, hardly self-averaging), a more convenient quantity to study is $\Delta F(n) = (N-n+1)^{-1}\sum_j \Delta F(j,j+n)$. From the properties of convolutions of Gaussian distributions and

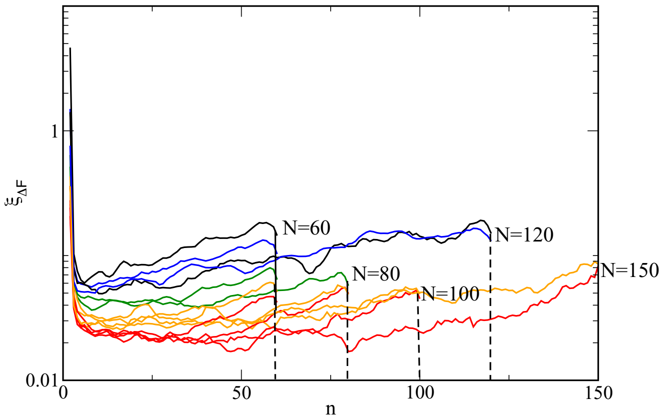

FIG. 7. The degree of self-averaging of $\Delta F(n)$ calculated at different values of N and of the temperature. The color code indicates the temperature and is the same as in Fig. 2.

Eq. (5) one obtains

$$p[\Delta F(n)] = \sqrt{\frac{(N-n+1)}{4\pi N\sigma^2}} \exp\left[-\frac{(\Delta F - T\beta\ln n + \epsilon)^2}{4N(N-n+1)^{-1}\sigma^2}\right]. \qquad (13)$$

Consequently, β' can be found, in analogy with Eq. (5), from a linear fit of $\Delta F(n)$ versus $\ln n$, weighted by $(N-n+1)/N$. Operatively, this is not different from a linear regression of $\Delta F(i,j)$ versus $\ln|i-j|$, since $(N-n+1)$ is just the multiplicity of pairs of monomers at linear distance n.

The parameter $\xi^2_{\Delta F(n)}$ which describes the degree of self-averaging of $F(n)$ is displayed in Fig. 7. For each T and N it displays a nonmonotonic behavior as a function of n. At low n, $\xi^2_{\Delta F(n)}$ is large as in the case of fixed-length heteropolymer (cf. Fig. 2); then it drops because each value of $\Delta F(n)$ is the average not only on the realizations of the disorder, but also on the $N-n+1$ segments of length n, and each of them can be regarded as a realization of the disorder as well (see the discussion in Ref. [24]). As n increases, this effect diminishes, and $\xi^2_{\Delta F(n)}$ increases. For fixed n, $\xi^2_{\Delta F(n)}$ displays at each temperature in the region $n \sim N$ a decreasing behavior, which suggests the self-averaging character of this quantity.

The behavior of $\overline{\Delta F(n)}$ as a function of $\ln n$ is displayed in Fig. 8, obtained from polymers with $N = 60, 80, 100, 120$ at different temperatures. The χ^2, weighted according to Eq. (13), associated with the fit from $n = 6$ (below which self-averaging is absent, cf. Fig. 7) to varying n is displayed in the inset of Fig. 8. At $T > 2.0$, corresponding to the elongated phase of the polymer (cf. previous section), the linear fit is very good except when $n \approx N$. At lower temperatures, only the central region is linear ($6 \lesssim n \lesssim 60$), while for $n \sim N$ the curve bends down similarly to that expected for a homopolymeric globule. However, in all cases the associated χ^2 remains lower than 1, due to the larger weight of small n to the fit.

The values of β' obtained from the fits is displayed in Fig. 9. Overall, the values of β' are smaller than those of β corresponding to the same temperature. At the highest temperature it displays the value ≈9/5 predicted for self-avoiding walks. At low temperatures, β' can reach values as low as 0.92. The reason is again that finite-size effects are

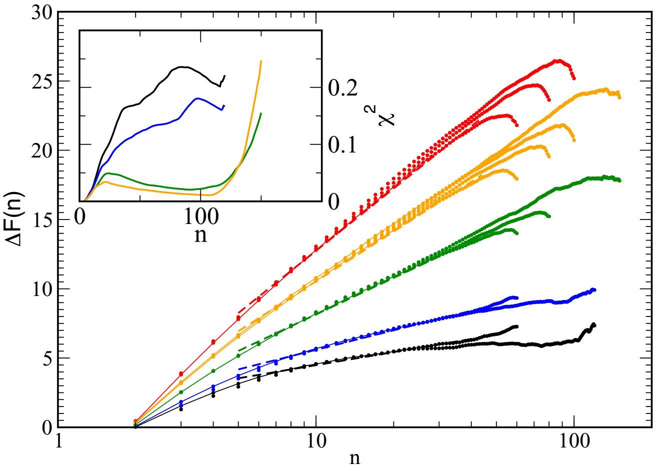

FIG. 8. The scaling of $\overline{\Delta F(n)}$ as a function of $\ln n$ at different temperatures (color code of Fig. 2) for different values of N. The fit, done between $N = 6$ and $n = 60$, is displayed with a dashed line. (Inset) The χ^2 associated with the fit up to length n.

amplified by the larger weight of small fragments of the chain, which is anyway unavoidable because fragments with $n \sim N$ are dominated by disorder.

VIII. IMPLICATIONS FOR CHROMOSOME CONFORMATION CAPTURE EXPERIMENTS

These results have important implications in the context of studies of chromosome conformation based on chromosome conformation capture (3C) experiments. In 3C-based methods, digestion and successive religation of formaldehyde-cross-linked chromatin in cell nuclei allows the detection of spatial proximity between DNA sequences (Fig. 10). In recent versions of 3C methods such as Hi-C, 4C, and 5C (reviewed in Ref. [6]), high-throughput sequencing is used to detect 3C DNA ligation products, making it possible to extract actual interaction frequencies. 3C-based experiments have allowed fundamental discoveries, notably that the folding of mammalian chromosomes is highly hierarchical. Each chromosome displays large-scale patterns of preferential associations into two so-called compartments, spanning several million base pairs of either active or inactive chromatin [7]. Compartments are further subdivided into smaller blocks of preferential interactions, referred to as topological associating domains (TADs) [26,27]. TADs are further characterized by the presence of smaller structures that occasionally define smaller domains dubbed loops domains [28].

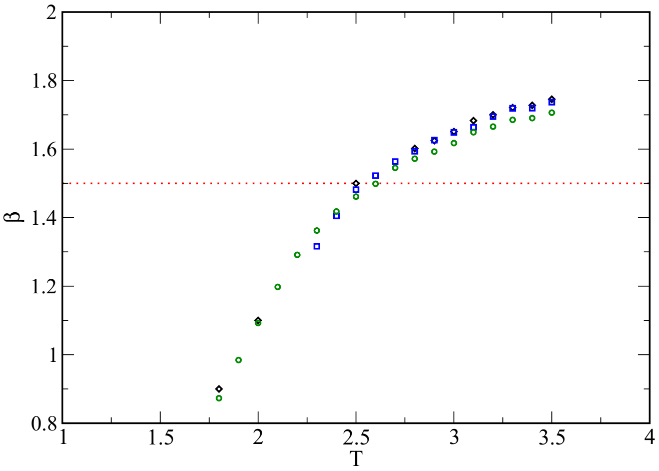

FIG. 9. The exponents β' associated with the fits of $\Delta F(n)$ versus n (solid black symbols), for the cases $N = 60$ (green circles), $N = 80$ (blue squares), and $N = 120$ (black diamonds).

In addition, 3C-based experiments make it possible to access the scaling behavior of chromosomes. Linear fitting of the logarithm of the experimentally determined contact probability versus the logarithm of the linear distance along the chain gives scaling exponents that are lower than those that are typical for homopolymers. Hi-C-based measurements led to scaling exponents of 1 over large genomic distances (between 10^6 and 10^7 base pairs) [7] and even smaller (∼0.75) at shorter genomic distances [8]. Importantly, these scaling behaviors have been often used to test alternative models for how chromosomes are folded in the three-dimensional space, and what mechanisms give rise to the observed hierarchical structure, at various genomic length scales [7,8,12,29,30]. In the earliest application of this strategy [7] it was shown that the $\beta \sim 1$ behavior observed on human chromosomes in the megabase range can be explained in terms of fractal globule (or crumpled globule, according to the original nomenclature [11]). A fractal globule is the out-of-equilibrium structure obtained by a rapid collapsed of a swollen coil; not having the time to explore the associated conformational space, in this metastable globular state the polymer partially retains the correlations it displayed in the coil state, and in particular the fact that each monomer binds preferentially to those which are close along the chain. Successive investigations suggested that other models must be invoked to explain the deviations from the $\beta \sim 1$ behavior, which are observed either when studying shorter genomic ranges [8] or when considering single chromosomes instead of their average behavior [12]. In addition, scaling exponents were recently used to support the validity of models based on energy-driven mechanisms such as loop extrusion by DNA-associated protein complexes [8,13], which could explain how specific chromosome structures such as TADs and loop domains emerge. Finally, mitotic chromosomes have been shown to display a peculiar double-decay regime, which was used to infer a model where loop extrusion leads to chromosome condensation [29].

Importantly, our calculations suggest that finite-size effects, combined with the heterogeneity of the interactions in the chain, are sufficient to account for the observed range of scaling exponents. Of course the model we described does not provide a mechanistic interpretation of the observed exponents. Nevertheless, it suggests that scaling exponents cannot be the only quantitative observable used to construct and validate a model for chromosome folding. Other properties of the chain, in particular, distance distributions between pairs of loci, correlations between them, or even their dynamic properties, which can all be measured experimentally should be also used to distinguish between alternative models.

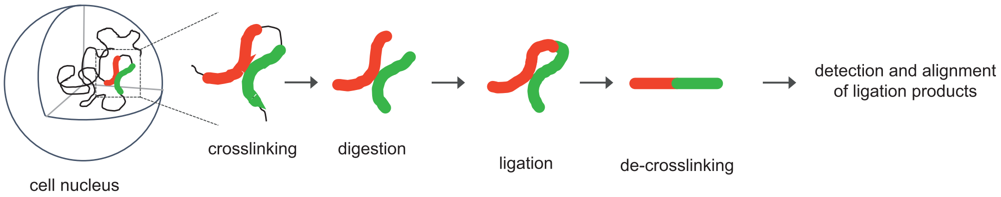

FIG. 10. Schematics of 3C-based techniques. Chromatin is cross linked with formaldehyde in the nuclei of a population of cells, digested with a restriction enzyme, and religated to favor the formation of hybrid DNA molecules that represent physical interaction events. After cremation decrosslinking, ligation products are purified, detected by DNA sequencing, and aligned to the reference genome.

IX. DISCUSSION

A. The polymer two-state model

The free-energy difference between looped and unlooped states within a two-state model provides a consistent way of studying the scaling properties associated with the looping mechanism with respect to the length of the random heteropolymer. From a theoretical argument and from numerical simulations, based on a self-adjusting simulated-tempering technique, the fluctuations about the average over the realizations of the random interaction within the heteropolymer are small, in the range of length of the order of 10^2 monomers but not in the thermodynamic limit.

Polymers of $\sim 10^2$ monomers are the longest systems for which we could guarantee equilibration, although with a consistent computational effort. Fortunately, this is the typical size of biological polymers. In fact, protein domains have an average length of 150 residues [31]. Topological associating domains in mammalian chromatin display a typical length of 10^6 bases, corresponding to 10^2 Kuhn lengths [32].

At high temperature, where the polymer is elongated, the looping probability of random heteropolymers displays a scaling exponent which varies continuously with respect to the temperature from ≈ 2.05 to 1.5. This is different from the behavior of homopolymers, for which only two possible exponents are expected.

At lower temperatures, corresponding to a compact phase of the heteropolymer, the determination of the scaling exponent is more cumbersome. Short chains display significant finite-size effects, resulting in a scaling of the looping probability with exponents smaller than 1.5. Longer chains display large disorder-dependent variability, which down-weights the determination of the exponent and the evaluation of the associated error. This amplifies the role of finite-size effects in the determination of the exponents even of large chains.

This phenomenon operates, for different reasons, both when considering chains of different lengths and segments of different lengths in a fixed-length heteropolymer. In the former case, the looping free energy is affected by the disorder provided by the internal energy of the chain, which is an extensive quantity. In the latter case, the free energy must be averaged over all the segments of the same length to be self-averaging, and the number of such segments decreases with the overall length of the chain. Anyway, fits of self-averaging free energies at low temperatures emphasize finite-size effects, resulting in exponents smaller than $3/2$.

B. Comparison with other models

Other investigations of the role of disorder in the looping of polymers were described in the literature, especially to describe the DNA double helix. In Ref. [33] it was shown that quenched randomness in the rest angles of a Kratky-Porod model result in a persistence length and response to external forces which are self-averaging (the latter under the hypothesis of small forces) and which are simply renormalized by the disorder.

A transfer-matrix formalism was used to study the effect of quenched (nonrandom) defects in the helasticity [34], resulting in a consistent increase in the looping probability of the polymer model. The same model was the extended [35] including random defects; a strong dependence of the looping probability was observed, suggesting a non-self-averageness of this property.

However, these models are controlled by the elasticity of the polymer and were designed to describe the properties of DNA strands of length comparable with their persistence length. The present model is thought to describe polymers, like chromatin and proteins, of length much larger than their persistence length (cf. Sec. IX A), and consequently no rigidity is modeled beyond the (inextensible) distance between consecutive monomers.

A perturbative calculation describing a flexible heteropolymer with random two-body interactions [24] showed that in the limit of small interaction volume the contact probability displays the standard homopolymeric exponents, affected by an exponential cutoff (cf. Sec. IX C below).

C. Role of excluded volume

The values of β found in the variable-length segments of a fixed-length chain are smaller than those of a set of chains of different lengths. There are two differences between the two cases. The former is that considering the variable-lengths segments of the same chain leaves correlations in the contact energies, which are absent when considering different realizations of varying-length chains. Moreover, when studying the variable-lengths segments of the same chain, the "tails" of the chain (i.e., the segments 1 to $i-1$ and $j+1$ to N, when studying the looping of i with j) may play a role. As a matter of fact, also for homopolymers it was shown [36] that the length of the tail can affect considerably the looping mechanism. The reason is that the excluded volume of the tail can shield the two monomers defining the loop, decreasing their binding probability.

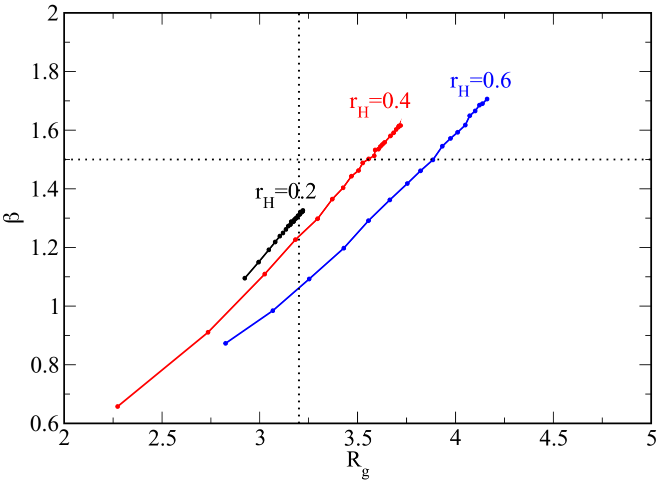

FIG. 11. The exponents β found at different temperatures, corresponding to different gyration radii R_g, using models with different length scale of the interaction potentials. The segments of a chain with $N = 60$ are used to calculate the values of β. The dotted lines indicate the expected values of R_g and of β at the θ point.

To investigate this point, we have repeated the simulations with different potentials, defined by different choices of the hardcore radius r_{HC} (and interaction radius proportional to r_{HC}), calculating the value of the exponent β for each of them. In Fig. 11 we show the result of these calculations. Since models with different r_{HC} display different temperature scales for the coil-globule transition, we use as an independent variable the gyration radius R_g. For each value of R_g, with decreasing r_{HC} the resulting β increases towards the values found with chains of different lengths, suggesting that the shielding effect plays a role in determining the difference between the two cases.

These results also suggests that the difference between the present numerical calculations and the analytical results found in Ref. [24], namely that for $T \geqslant \theta$ the exponent of a heteropolymer should not change with respect to the homopolymeric case, while only an exponential cutoff appears in the looping probability, can be associated with the hypothesis $r_{HC} \to 0$ used in the analytical calculations.

X. CONCLUSIONS

In random heteropolymers, scaling exponents relating the contact probability between monomers with their linear distance along the chain display strong finite-size effects which are amplified because of the large variability of the probability of long-range contacts, which are consequence of their lack of self-averageness. We suggest that this effect can strongly affect the interpretation of experimental data describing the scaling of contact probability in biopolymers. In the case of chromosome folding, our results suggest that one should be careful in selecting a physical model to describe the behavior of chromosome based on its scaling exponents, as a random heteropolymer can show exponents similar to those observed in experiments.

[1] F. Spitz, Semin. Cell Dev. Biol. **57**, 57 (2016).
[2] S. W. Bruun, V. Iesmantavicius, J. Danielsson, and F. M. Poulsen, Proc. Natl. Acad. Sci. USA **107**, 13306 (2010).
[3] M. Buscaglia, L. J. Lapidus, W. A. Eaton, and J. Hofrichter, Biophys. J. **91**, 276 (2006).
[4] K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).
[5] L. Liu and C. Hyeon, Biophys J. **110**, 2320 (2016).
[6] A. Denker and W. de Laat, Genes Dev. **30**, 1357 (2016).
[7] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, Science **326**, 289 (2009).
[8] A. L. Sanborn, S. S. P. Rao, S.-C. Huanga, N. C. Duranda, M. H. Huntley, A. I. Jewett, I. D. Bochkova, D. Chinnappan, A. Cutkosky, J. Li, Kristopher P. Geeting, A. Gnirkee, A. Melnikove, D. McKenna, E. K. Stamenova, E. S. Lander, and E. L. Aiden, Proc. Natl. Acad. Sci. USA **112**, E6456 (2015).
[9] P.-G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, NY, 1979).
[10] L. Mirny, Chromosome Res. **19**, 37 (2011).
[11] A. Yu. Grosberg, S. K. Nechaev and E. I. Shakhnovich, J. Phys. (France) **49**, 2095 (1988).
[12] M. Barbieri, M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo, and M. Nicodemi, Proc. Natl. Acad. Sci. USA **109**, 16173 (2011).
[13] Goloborodko, J. F. Marko, and L. A. Mirny, Cell Reports **15**, 2038 (2016).
[14] E. I. Shakhnovich and A. M. Gutin, J. Phys. (France) **50**, 1843 (1989).
[15] I. M. Lifshits, Zh. Eksp. Teor. Fiz. **12**, 117 (1942).
[16] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[17] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).
[18] G. Tiana and L. Sutto, Phys. Rev. E **84**, 061910 (2011).
[19] J. Johnson, C. A. Brackley, P. R. Cook, and D. Marenduzzo, J. Phys.: Condens. Matter **27**, 064119 (2015).
[20] L. Giorgetti, R. Galupa, E. P. Nora, T. Piolot, F. Lam, J. Dekker, G. Tiana, and E. Heard, Cell **157**, 950 (2014).
[21] G. Tiana, F. Villa, Y. Zhan, R. Capelli, C. Paissoni, P. Sormanni, E. Heard, L. Giorgetti, and R. Meloni, Comput. Phys. Commun. **186**, 93 (2014).
[22] R. Brout, Phys. Rev. **115**, 824 (1959).
[23] A. J. Guttman and M. F. Sykes, J. Phys. C **6**, 945 (1973).
[24] G. Tiana, Phys. Rev. E. **92**, 010702R (2015).
[25] A. Clauset, C. S. Shalizi and M. E. J. Newman, SIAM Rev. **51**, 661 (2009).
[26] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Bluthgen, and J. D. E. Heard, Nature (London) **485**, 381 (2012).
[27] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, Nature (London) **485**, 376 (2012).

[28] S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, Cell **159**, 1665 (2014).

[29] N. Naumova, M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie, L. A. Mirny, and J. Dekker, Science **342**, 948 (2013).

[30] F. Benedetti, J. Dorier, Y. Burnier, and A. Stasiak, Nucl. Acid Res. **42**, 2848 (2014).

[31] D. Xu and R. Nussinov, Folding Des. **3**, 11 (1998).

[32] J. Dekker, J. Biol. Chem. **283**, 34532 (2008).

[33] D. Bensimon, D. Dohmi, and M. Mézard, Erophys. Lett. **42**, 97 (1998).

[34] J. Yan, R. Kawamura, and J. F. Marko, Phys. Rev. E **71**, 061905 (2005).

[35] P. Ranjith, P. B. Sunil Kumar, and G. I. Menon, Phys. Rev. Lett. **94**, 138102 (2005).

[36] H. S. Chan and K. A. Dill, J. Chem. Phys. **90**, 492 (1989).