Calculator for conformational statistics of DNA and applications to high-curvature bending

Brian C. Ross and Paul A. Wiggins

Phys. Rev. E **87**, 032707 — Published 11 March 2013

DOI: 10.1103/PhysRevE.87.032707

Calculator for conformational statistics of DNA and applications to high-curvature bending

Brian C. Ross
Department of Physics, University of Washington, Seattle, WA, USA

Paul A. Wiggins
*Departments of Physics and Bioengineering, University of Washington, Seattle, WA, USA**

(Dated: February 22, 2013)

DNA conformation plays an important role in a host of cellular processes. Despite the central importance of DNA conformation, there is not yet a general-purpose calculator for conformational statistics that is designed for the scientific community. Here we describe a public tool we developed for calculating an important class of conformational statistics: the end-to-end probability density of finding a locus of the DNA polymer at a given displacement and orientation relative to a second locus on the same polymer. As a demonstration, we propose a novel cyclization experiment and use our calculator to show that this experiment could measure the energy of DNA bending as a direct function of bend angle in the poorly-understood high-bending regime. Our tool is available as both an online calculator and a downloadable program at: http://mtshasta.phys.washington.edu/wormulator/

PACS numbers:

Introduction

Many important cellular processes are controlled or influenced by the mechanical properties of DNA [1]. For example, stiffness plays a role in regulatory DNA looping [2] and condensation of the chromosome [3, 4], while the flexibility of DNA at very short scales affects the binding affinity of DNA-bending and DNA-bridging proteins [5]. In each of these cases, proteins bind two or more genetic loci, constraining the displacement and orientation of these bound loci relative to one another. (See Fig. 1.) These constraints increase the free energy of the polymer by decreasing the entropy (number of accessible configurations) and increasing the enthalpy (bending energy) of the DNA polymer.

For example, consider the binding of a transcription factor T with two distinct DNA binding domains to DNA sites A and B. The equilibrium concentration of the protein-DNA complex ATB is predicted to be:

$$[\mathrm{ATB}] = K_A K_B [\mathrm{A}][\mathrm{B}][\mathrm{T}], \tag{1}$$

where K_A and K_B are the equilibrium association constants for the binding of sites A and B to T at equilibrium concentrations [A], [B] and [T] respectively. If we consider the concentration of looped complexes where A and B are on the same DNA molecule, once sequence A is bound, the relevant concentration of B is not its solution concentration, but rather the concentration of B at its binding site on T *given that* A is bound to the other end of T. This concentration, $[\mathrm{B}]^*$, is called the effective concentration or J factor and is predicted by the statistical mechanics of the DNA polymer [6, 7]:

$$[\mathrm{B}]^* \equiv J = 8\pi^2 p(\Delta\mathbf{R}, \mathbf{\Omega}_\ell | \mathbf{\Omega}_0; \ell), \tag{2}$$

where J is the J factor, p is the conditional probability density (or concentration per radian cubed) of the polymer with physical displacement $\Delta\mathbf{R}$ and orientation $\mathbf{\Omega}_\ell$ at locus A relative to locus B with orientation $\mathbf{\Omega}_0$ when sites A and B are separated by contour length ℓ along the polymer chain. $\Delta\mathbf{R}$, $\mathbf{\Omega}_\ell$, and $\mathbf{\Omega}_0$ are a consequence of protein conformation and ℓ is a consequence of DNA sequence. The factor of $8\pi^2$ is a consequence of the requirement that the angular orientation of B is also constrained whereas the equilibrium constants are defined assuming an isotropic angular distribution of molecules in the solvent ($1/8\pi^2$). Therefore the concentration of looped complex is predicted to be:

$$[\mathrm{ATB}_{\mathrm{loop}}] = K_A K_B J [\mathrm{DNA}][\mathrm{T}], \tag{3}$$

where $[\mathrm{DNA}] = [A] = [B]$ is the solution concentration of each DNA sequence. We define the looping free energy:

$$G_{\mathrm{loop}} \equiv -k_B T \log J, \tag{4}$$

where k_B is the Boltzmann constant and T the absolute temperature and J is in units of Molarity.

Two special cases are worth discussing which simplify Eqn. 2. In the special case where the tangent of the DNA sequence is constrained, but not the twist, the effective concentration simplifies to

$$J = 4\pi p(\Delta\mathbf{R}, \mathbf{u}_\ell | \mathbf{u}_0; \ell), \tag{5}$$

where p is the probability density (or concentration per radian squared) of the polymer with physical displacement $\Delta\mathbf{R}$ and tangent $\mathbf{u}_\ell$ at locus A relative to locus B with tangent $\mathbf{u}_0$ when sites A and B are separated by contour length ℓ along the polymer chain. A second

*Electronic address: pwiggins@uw.edu

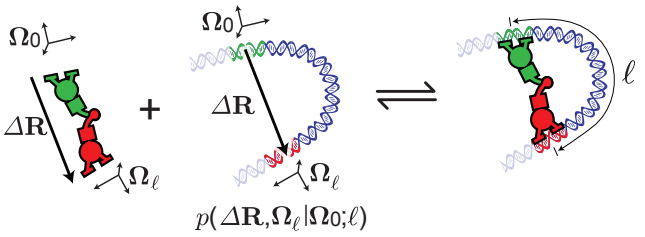

FIG. 1: (Color online) DNA looping induced by the binding of a transcription factor with two DNA binding domains. The conformation of the protein determines the displacement between binding sites ($\Delta\mathbf{R}$) and the orientations $\mathbf{\Omega}_0$ and $\mathbf{\Omega}_\ell$. The DNA sequence determines the contour length (ℓ) between the binding sites. Once the first domain has bound DNA, the conformational statistics of the polymer determine the effective concentration ($J = 8\pi^2 p$) of the second binding site at its binding domain.

special case describes the situation where there are no orientational constraints. In this case, the effective concentration further simplifies to

$$J = p(\Delta\mathbf{R}; \ell), \tag{6}$$

where p is the probability density (or concentration) of the polymer with physical displacement $\Delta\mathbf{R}$ at locus A relative to locus B when sites A and B are separated by contour length ℓ along the polymer.

The conditional probability density which gives $[\mathrm{B}]^*$ is computed using the bending energy-weighted Boltzmann distribution within a path integral over all polymer configurations satisfying the endpoint conditions (e.g. [7]):

$$p(\Delta\mathbf{R}, \mathbf{\Omega}_\ell|\mathbf{\Omega}_0; \ell) \equiv \frac{dP}{d^3\Delta\mathbf{R}d^3\mathbf{\Omega}_\ell}, \tag{7}$$

$$= Z^{-1}\int [D\mathbf{X}(s)]\; e^{-E[\vec{X}(s)]/k_BT} \times\delta^3[\Delta\mathbf{R} - (\mathbf{R}_(\ell) - \mathbf{R}(0))] \times\delta^3[\mathbf{\Omega}_\ell - \mathbf{\Omega}(\ell)], \tag{8}$$

where P is the cumulative probability, $\mathbf{X}(s)$ parametrizes the conformation, δ is the Dirac delta function which enforces the final displacement and orientation constraints, Z is the partition function (the path integral evaluated without constraint delta functions), and E is the bending free energy functional.

The end-to-end partition function depends the energetics of DNA bending and twisting. The most popular mechanical description of DNA is given by the wormlike chain model (WLC), in which the DNA axis is modeled as a continuous line in space, with a tangent vector $\mathbf{u}(s)$ and a rotating twist angle $\psi(s)$ that are functions of the contour location s. The energy of a wormlike chain polymer is [7]:

$$E = k_BT\int_0^\ell \left[\frac{l_p}{2}\left(\frac{d\mathbf{u}}{ds}\right)^2 + \frac{l_t}{2}\left(\frac{d\psi}{ds} - \omega_\psi\right)^2\right] ds. \tag{9}$$

The parameter l_p is called the (bending) persistence length [7], and sets the intrinsic length scale for DNA bending. The persistence length is roughly 50 nm for naked *in vitro* DNA [8]. In this paper we will call segments of DNA 'long' or 'short' if the ratio of the contour length to the persistence length is significantly greater or less than one. The parameter l_t, which is called the twist persistence length, sets the length scale for computing statistics of DNA twist. Various measurements of the twist persistence length of DNA give roughly 100-120 nm [8]. The unstressed twist rate ω_ψ of DNA is approximately one full helical turn per 10.5 bases.

Methods

We have developed web-based and downloadable calculators for computing the various end-to-end distributions of DNA under the phantom-chain approximation [33] whereby excluded-volume interactions are ignored. These distributions are computed under the assumption that the polymer is in equilibrium in a thermal environment. Our calculator can use three complementary methods and treats a variety of polymer models. An eigenfunction-based method, due to Spakowitz and Wang [9, 10], is best-suited for cases in which the locus separation is much greater than the bending scale of the DNA. The numerical Monte Carlo method efficiently computes statistics of shorter DNA contours, and can bias the sampling towards high-energy conformations to improve statistics. Finally, the harmonic approximation method of Zhang and Crothers [11] handles polymers that are sharply deformed due to positional and/or orientational constraints. The web calculator has a straightforward and intuitive interface but is restricted to the wormlike chain model and lacks the Zhang-Crothers method. The downloadable program uses a command-prompt interface, but has the full range of capabilities (including sequence-dependent and nonharmonic polymer models for Monte Carlo) and can be used for intensive calculations.

Gaussian chain method

In the limit of very long contours, the DNA essentially performs a random walk in which the step taken over any individual persistence length is much smaller than the total distance traversed. In this situation the end-to-end distribution approaches the Gaussian chain distribution [7]:

$$p(\Delta\mathbf{R}) \longrightarrow \left(\frac{3}{4\pi l_p L}\right)^{3/2} \cdot e^{-3(\Delta R)^2/4l_p L} \qquad (10)$$

where the orientational distributions are uniform. This long-chain limit is trivial to calculate and is included in our program.

Eigenfunction method

An exact solution to the wormlike chain end-statistics problem was obtained by Spakowitz and Wang [9, 10], which can be written as follows:

$$p(\mathbf{R}_\ell, \mathbf{\Omega}_\ell | \mathbf{R}_0, \mathbf{\Omega}_0; L) = \sum_{l_0, l_f, m, j} \mathcal{F}^{-1}\left[\mathcal{L}^{-1}\left\{f^{mj}_{l_0 l_f}(\mathbf{\Omega}_0, \mathbf{\Omega}_\ell, \hat{\mathbf{k}}) \cdot \mathcal{G}^{mj}_{l_0 l_f}(k,p)\right\}\right]. \qquad (11)$$

Here $\mathcal{F}^{-1}$ is the inverse Fourier operator that converts the variable $\mathbf{k}$ (having magnitude k and direction $\hat{\mathbf{k}}$) into $\mathbf{R}_\ell - \mathbf{R}_0$; and $\mathcal{L}^{-1}$ is the inverse Laplace operator which converts p to the chain length L. $\mathbf{R}_0$ and $\mathbf{R}_\ell$ give the positions of the ends: $\mathbf{R}_\ell = \mathbf{R}_0 + \Delta\mathbf{R}$. The functions f (a product of Wigner functions [7]) and $\mathcal{G}$ (a product of continued fractions) are given explicitly in [9, 10]. The variables l_0 and l_f range from 0 to infinity in both the sums and the continued fractions, but the higher terms tend towards zero so in practice we drop all terms above some cutoff l_{max} when we perform a calculation. The shorter the contour (relative to a persistence length), the higher l_{max} must be to achieve a given accuracy.

Evaluating Eq. (11) in the straightforward way involves eight nested iterations: the four sums over l_0, l_f, m and j; the three inverse Fourier integrals; and the inverse Laplace transform. In order to speed up the calculation, our implementation pre-computes and stores the roots of the continued-fraction polynomials that contribute to the residues of the inverse Laplace transform. Effectively, we compute the following:

1. $g^{mj}_{l_0 l_f}(k, L) = \mathcal{L}^{-1}\left\{\mathcal{G}^{mj}_{l_0 l_f}(k,p)\right\}$
2. $p(\mathbf{R}_\ell, \mathbf{\Omega}_\ell | \mathbf{R}_0, \mathbf{\Omega}_0; L) = \sum_{l_0, l_f, m, j} \mathcal{F}^{-1}\left[f^{mj}_{l_0 l_f}(\mathbf{\Omega}_0, \mathbf{\Omega}_\ell, \hat{\mathbf{k}}) \cdot g^{mj}_{l_0 l_f}(k, L)\right].$

The limiting step 2 now involves seven nested sums rather than eight, greatly speeding evaluation of the expression. The memory required to store the results of step 1 can be significant, but overall we have found this tradeoff to be worthwhile.

In certain special cases we can obtain further boosts in speed by exploiting symmetries of the perturbation series. For the full distribution (11) we can take advantage of the fact that the continued fractions are nearly symmetric with respect to m and j, up to an additive term in the Laplace variable. To obtain the reduced distribution $p(\mathbf{R}_\ell, \mathbf{u}_\ell | \mathbf{R}_0, \mathbf{u}_0; L)$ that ignores the relative twist of the two ends, we ignore the sum over j and set $j = 0$; to obtain $p(\mathbf{R}_\ell | \mathbf{R}_0; L)$ we set $m = j = 0$. To obtain the orientation-only distribution $p(\mathbf{\Omega}_\ell | \mathbf{\Omega}_0; L)$ we replace the inverse-Fourier operation with a simple evaluation at $\mathbf{k} = 0$ and restrict $l_0 = l_f$. Finally, to compute statistics for cyclization ($\mathbf{R}_\ell = \mathbf{R}_0$ and $\mathbf{\Omega}_\ell = \mathbf{\Omega}_0$), we set $l_0 = l_f$, and exploit the fact that the expression is symmetric with respect to m and j, and with m and $-m$ modulo a complex conjugation.

Our inverse-Laplace solver uses a C++ implementation of the complex Jenkins-Traub root-finding algorithm written by Henrik Vestermark (http://www.hvks.com/Numerical/ports.html). A root-polisher using Newton's method ensures that the roots are at machine precision.

Monte Carlo method

Our second method for calculating end statistics is Monte Carlo sampling, in which a large number of representative conformations are generated beginning from $(\mathbf{R}_0, \mathbf{\Omega}_0)$, and the distribution $p(\mathbf{R}_\ell, \mathbf{\Omega}_\ell)$ is estimated by counting the number of conformations whose second end lies within some finite window of the desired $(\mathbf{R}_\ell, \mathbf{\Omega}_\ell)$. In order to generate representative conformations, the algorithm must be given the end-to-end distribution for a single segment; in other words, Monte Carlo constructs $p(\mathbf{R}_\ell, \mathbf{\Omega}_0 | \mathbf{R}_0, \mathbf{\Omega}_\ell; L)$ from a polymer model $p(\mathbf{R}_{\ell_s}, \mathbf{\Omega}_{\ell_s} | \mathbf{R}_0, \mathbf{\Omega}_0; \ell_s)$ where ℓ_s is the segment length. Monte Carlo is fastest at short contour lengths because short polymers are quick to construct. It is therefore complementary to the eigenfunction technique of Spakowitz, which works best at long contour lengths.

We write our polymer model as $p(\mathbf{X}; \ell_s)$ where $\mathbf{X} = \{\Delta r_\parallel, \Delta r_\perp, \Delta r_\times, \theta, \phi, \psi\}$ describes translations and rotations using axes affixed to the polymer. In our Monte Carlo implementation the user specifies $E(\mathbf{X}; \ell_s)$, and a discretized probability function is computed using the Boltzmann factor $p \propto J(\mathbf{X}) e^{-E(\mathbf{X})}$ where $J(\mathbf{X})$ is the volume factor appropriate to the system. Due to our use of numerical interpolation tables both $E(\mathbf{X})$ and $J(\mathbf{X})$ can be specified arbitrarily (although using independent distributions for the various X_i saves significant memory). Our Monte Carlo method can accommodate nonharmonic energy functions, extensible polymers, coupled degrees of freedom, 1-3 dimensional polymers, and sequence-dependent models such as the DNA model of [12] which is included in our Monte Carlo calculator. In order to evolve the polymer once the trajectory vector $\mathbf{X}$ has been sampled at each segment, we use the method outlined in ref. [13].

Our Monte Carlo implementation can either sample endpoint statistics as the chains are generated, or store

the endpoints of all chains and sample their statistics in a separate step. The first method saves memory as little storage is required. The second method is much faster when the distribution is being measured at multiple points since most of the computational expense is in generating the chains.

Thermal sampling on molecular scales is much faster than numerical sampling on a computer, which leads to the problem that sparse regions of the end-to-end distribution may be relevant to biology despite being hard to access numerically. To sample these regions, our Monte Carlo code has the ability to bias its sampling towards certain conformations by constructing the chains from a different single-segment distribution than the distribution that defines the model. The sampling bias is then corrected for by post-weighting. For example, if a given bend angle of a polymer is drawn from a point where the sampling distribution has twice the value of the true distribution then that joint will contribute a factor of $1/2$ to the weight. The total weight is the product of weighting factors at all evaluations of the interpolation tables. Measurement of the end-to-end distribution $p()$ then involves summing weighted conformations. The generic name for this technique is 'importance sampling'.

Our Monte Carlo method estimates the sampling error of a general biased sample by binning the weighting factors and estimating the counting error in each bin b:

$$N^2V^2\left\langle \delta p^2\right\rangle = \sum_b w_b^2\left(\left\langle n_b^2\right\rangle - \left\langle n_b\right\rangle^2\right)$$

If we take these bins to be very small so that $\langle n_b\rangle \ll 1$, then n_b is almost certain to be zero or one, in which case $n_b^2 = n_b$. Then

$$\begin{aligned} N^2V^2\left\langle \delta p^2\right\rangle &\approx \sum_b w_b^2\left(\left\langle n_b\right\rangle - \left\langle n_b\right\rangle^2\right) \\ &\approx \sum_b w_b^2\left\langle n_b\right\rangle \end{aligned}$$

We can estimate the error using the expression $\delta p \approx \sqrt{\sum_i w_i^2}/NV$ using the sample set w_i. In the special case of an unweighted sample set this reduces to $\delta p \approx p/\sqrt{n_{hits}}$, although for unweighted samples we explicitly use $\delta p \approx p/\sqrt{n_{hits}-1}$ to remove the bias of having estimated the mean from the same sample set.

In addition to the various end-to-end distributions, we include routines for measuring the various moments of the distribution: the mean end-to-end *distance* function $\left\langle R^{2n}\right\rangle$, and the mean of $\langle(\mathbf{R}\cdot\mathbf{u}_0)^n\rangle$ for any n, where $\mathbf{R}$ is the end-to-end displacement and $\mathbf{u}_0$ is the initial tangent vector. These functions complement analytical results of these same quantities [7], as those can be difficult to evaluate. To estimate the error in the moments, our program divides the set of N conformations into m disjoint subsets, computes the moment separately using each subset, and then estimates the error based on the variance in the moments of the subsets.

Harmonic approximation method

The eigenfunction and Monte Carlo methods described above are most accurate when there are low-energy polymer conformations that satisfy the end-to-end constraints. To complement these, we have also implemented the 'harmonic approximation' (HA) method of Zhang and Crothers [11] which works best in the regime of high-energy, sharply-bent conformations. The HA method estimates the probability function by integrating Eq. 8 about the minimum-energy configuration of the polymer that satisfies the given constraints. When the polymer is sharply bent, the energy trough tends to be steep, fluctuations are small and approximations made in the perturbative integral become ignorable. As in the case of Monte Carlo, the HA method can incorporate any number of positional and/or orientational constraints along the length of the polymer.

Our HA calculator is essentially an extension of our Monte Carlo calculator, so it can be applied to the same variety of DNA models as Monte Carlo deals with. Sequence-dependence, extensibility, coupled degrees of freedom and non-harmonic energy functions can all be accounted for using HA. The HA method itself is described in detail in ref. [11]; to summarize, their method first finds the minimum-energy conformation of the polymer, then computes the distribution by summing over fluctuations about this conformation (assumed small-amplitude if the polymer is sharply constrained). The main differences in our implementation are 1) that we do not assume a wormlike chain model and therefore the answer is couched in terms of derivatives of the energy function; 2) the constraints are arbitrary whereas ref. [11] concentrates on cyclization; 3) the unconstrained partition function is found by numerically integrating our interpolation tables [14]. To find the minimum-energy conformation, we use a general-purpose multi-dimensional root finder from the GNU Scientific Library (GSL) [15].

For some purposes we would like to know the normal modes of a constrained polymer. For example, our calculator allows one to perform biased Monte Carlo where the normal modes of the constrained polymer comprise the bias function. Unfortunately, real-valued normal modes do not come directly out of the Zhang and Crothers analysis, because the delta-function constraints are made tractable for integration by Fourier-transforming them into complex space. However, if we replace the singular delta-functions with a narrow Gaussian, then both the energy and the constraint appear straightforwardly within an exponential, which we can expand to second order and convert to real-valued normal modes.

Formally, to obtain normal modes we write the distribution as $p = Z_c/Z$ where Z_c is the constrained partition function and Z is the unconstrained partition function.

Z is just a summation over the interpolation tables. To evaluate Z_c, we compute:

$$Z_c = \int dx_1 dx_2 \dots dx_N \sqrt{\frac{k_1}{2\pi}} \sqrt{\frac{k_2}{2\pi}} \cdots \sqrt{\frac{k_m}{2\pi}} \times \left(e^{-E_0} e^{-\frac{k_1}{2} f_1^2} e^{-\frac{k_2}{2} f_2^2} \dots e^{-\frac{k_m}{2} f_m^2} \right) \quad (12)$$

where f_i denote the constraint functions and E_0 is the energy of the minimum-energy conformation. Performing the integration, we obtain:

$$Z_c = (2\pi)^{(N-m)/2} \sqrt{\frac{|\mathbf{K}_c|}{|\mathbf{M}|}} e^{-E_0'} \quad (13)$$

where $E_0' = E_0 + \frac{1}{2}\mathbf{f}_0^T \mathbf{K}_c \mathbf{f}_0$, $\mathbf{K}_c$ is a generalized stiffness matrix, and $\mathbf{M}$ is defined by:

$$M_{ij} = \frac{\partial}{\partial_{x_i}} \frac{\partial}{\partial_{x_j}} (E + \lambda \cdot \mathbf{f}) + \left(\frac{\partial \mathbf{f}}{\partial_{x_i}} \right)^T \mathbf{K}_c \left(\frac{\partial \mathbf{f}}{\partial_{x_j}} \right). \quad (14)$$

For good sampling statistics, we choose $\mathbf{K}_c$ so that the variances in the f_i will be of the same order as the respective sampling window sizes. Specifically, we dial in $\mathbf{K}_c$ such that the projection of $\mathbf{M}$ into $\mathbf{f}$-space is diagonal with entries $(\mathbf{PMP}^T)_{ii} = T_{ii} = 1/\sigma_i^2$, where σ_i is the window size of constraint i. Projections between $\Delta\mathbf{X}$ and $\mathbf{f}$ are effected by $\mathbf{f} = \boldsymbol{\beta}\Delta\mathbf{X}$ and $\Delta\mathbf{X} = \mathbf{P}\mathbf{f}$, where $\beta_{ij} = \partial_{x_i} f_j$ and $\mathbf{P} = \left(\boldsymbol{\beta}^T\boldsymbol{\beta}\right)\boldsymbol{\beta}^T$. Writing $\mathbf{M}$ in the form $\mathbf{M} = \mathbf{M}_0 + \boldsymbol{\beta}\mathbf{K}_c\boldsymbol{\beta}^T$, we obtain the constraint stiffness matrix $\mathbf{K}_c = \mathbf{T} - (\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^T\mathbf{M}_0\boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}$.

Web interface

The online calculator allows the user to measure the end-to-end distribution $p(\mathbf{R}_\ell, \mathbf{\Omega}_\ell | \mathbf{R}_0, \mathbf{\Omega}_0; \ell)$ for single values of $\{\Delta\mathbf{R}, \mathbf{\Omega}_0, \mathbf{\Omega}_\ell, \ell\}$ by providing a polymer length ℓ, endpoint tangents and relative displacement and twist. The online calculator only supports the wormlike chain model; however, one can control the material parameters (bending/twist persistence lengths and intrinsic twist) so to model polymers other than double-stranded DNA. Checkbox options to sum over $\mathbf{R}$, tangents and/or twists allow the various reduced distributions to be computed. One convenience, not present in the command-line tool, is that the program supports several length units (nanometers, persistence lengths, base pairs of DNA, etc.). Online computations may be performed using the eigenfunction and Monte Carlo methods.

The output of the basic computation outlined above is a single number: a probability density (probability per unit volume and/or unit angular volume) for the polymer's second end to be in the given position and/or orientation relative to the first end. Frequently, the user would like to map this distribution over a range of values in some parameter – for example, to predict the efficiency of cyclization over a range of polymer lengths. It would be tedious to do these multiple evaluations manually, so the online calculator incorporates a 'counter' for accomplishing this automatically: given a range of values of the counter variable, the program will evaluate multiple times, once for each value of the counter. To set up unique conditions for every run the user writes the counter variable '`c`' into the input fields: for example the twist field might read `sin((pi/8)*c)`, or the length may be `e^-c`. If the counter is used, then the output of the calculator will be a table, where the values of the counter are displayed alongside the outputs of each evaluation of the distribution.

It is easy for the user to request a calculation that will either require too much memory or run practically forever, especially when using the eigenfunction method which involves many nested loops. In order to avoid overtaxing the server, the online calculator restricts the permitted ranges of parameters that affect memory usage and computation time: the maximum l-value and the number of integration steps in the eigenfunction calculation, the number of samples and discrete segments that Monte Carlo generates, and the range of the counter variable. Because of these restrictions, intensive calculations can only be done using the command-line tool.

Downloadable tool

To perform a calculation using the downloadable command-line tool, the user may enter commands directly into the interactive prompt, or else place those commands into a file and have the program execute them all at once. A basic computation requires two or three instructions. The program has capabilities for generating and saving tables, inspecting intermediate stages of the calculation, controlling the random sequences, and measuring computation time and memory usage. A help file that is included documents all of the commands and gives examples for each type of calculation.

The Monte Carlo component of the command-line tool has several capabilities that are not available from the web site. One is the ability to use very general 2D or 3D polymer models, including those with non-harmonic energy functions, coupled degrees of freedom, sequence dependence and extensible segments, along with the trick of biased sampling. Additionally, only the command-line tool has the perturbative method of Zhang and Crothers. The command-line tool can perform very lengthy calculations that are forbidden online. Finally, the ability to export tables is useful for storing results, making plots, and troubleshooting.

Validation and Results

In order to validate our program, we performed a number of calculations that could be checked either explicitly or against a different method. For the perturbative calculation, we compared selected computations of Euler angles and Wigner functions with hand-derived results, verified that the distribution asymptotically approaches the expected Gaussian for long chains, and reproduced the cyclization plot given in ref. [10]. We also compared probability densities given by our implementation of the perturbative method with equivalent calculations performed using a symbolic calculator (Mathematica [16]), drawing test cases from the full distribution and from the orientation-only and cyclization distributions. Tests of the Monte Carlo method included explicit checks on the propagation and rotation of individual segments, on the sampling of the bending/twisting energy functions, and evaluations of $\langle \mathbf{R} \cdot \mathbf{u}_0 \rangle$ which should approach $l_p \left(1 - e^{-L/l_p}\right)$. In all cases the results agreed with the predictions within numerical precision, as long as the parameters controlling accuracy (l_{max}, K_{max}, etc. for the perturbative method; segment length and number of runs for Monte Carlo) were made stringent enough.

When a polymer's length is on the order of a couple of persistence lengths, both the perturbative calculation and Monte Carlo can give good answers with reasonable computational cost. We generated end-to-end distributions of a 3-persistence-length stretch of DNA using these two methods and found the expected agreement.

Figure 2 Panel B compares the various computational methods for computing J factors for the special case of DNA cyclization, in which the configurational constraint is zero displacement between the loci and aligned orientation:

$$J_{\text{cycl}} = 8\pi^2 p(0, \mathbf{\Omega} | \mathbf{\Omega}; \ell). \tag{15}$$

Cyclization provides an important connection between theory and experiment, as the J factor is simply the equilibrium ratio of cyclized to dimerized DNA fragments when the species are allowed to reversibly anneal (Fig. 2, Panel A).

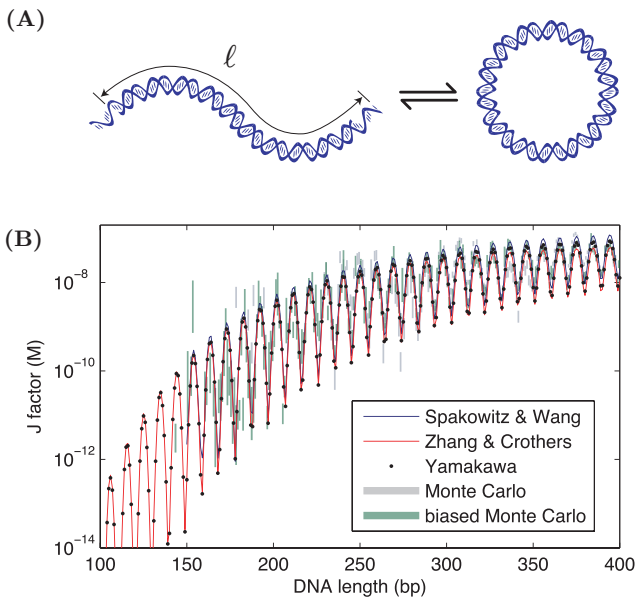

FIG. 2: (Color online) **Panel A:** The DNA cyclization reaction is a convenient tool for exploring the effective concentration. In cyclization, linear sequences are ligated to form DNA circles. The constraints for cyclization are: zero displacement between ends ($\Delta\mathbf{R} = 0$), and the orientations of the two polymer ends are aligned ($\mathbf{\Omega}_0 = \mathbf{\Omega}_\ell$). **Panel B:** DNA cyclization J factor as a function of length using the wormlike chain model [7], as computed by five different computational techniques. Both entropic and energetic contributions play an important role in determining the J factor. The J factor declines steeply at short contour length since the bending energy scales like the inverse contour length. The J factor is also modulated by the helical repeat of DNA (10.5 bp) since the ends must be in helical registry for ligation. The perturbative methods are accurate for short (highly-stressed) contours but fail at longer lengths where fluctuations about the minimum-energy conformation become large. Monte Carlo and the method of Spakowitz and Wang both work well for long contours, but at short lengths Monte Carlo becomes inaccurate due to sparse sampling of highly-stressed conformations while the Spakowitz-Wang method fails to converge with the given tolerances. Our biased Monte Carlo runs lowered the energy barrier by a factor of 1.5 (and corrected for it by post-weighting). The perturbative results were obtained using discrete chains with 500 segments; the Monte Carlo results used 20-segment chains to obtain good statistics with the biasing method. We have included a perturbative result of Yamakawa [7] that is not computed by our program, to compare against the perturbative result of Zhang and Crothers.

Application: Double-Gap Cyclization

Cyclization measurements, using either ligation or FRET, are an extremely sensitive measure of DNA bending energy and provide compelling evidence that DNA is more flexible at high bending angles than predicted by the wormlike chain model. Many theories of high flexibility of tightly bent DNA predict the localization of bending to a kink, similar to that observed when a metal tape measure is bent [17–22]. Such kinking could lead to DNA bending at a well-defined large kink angle [17], or if the mechanism leading to kinking was localized DNA melting, a flexible joint [19]. Whether such structures exist and what their characteristics are remain important and outstanding questions.

Unfortunately, traditional cyclization measurements do not provide a straightforward approach for differentiating between various DNA kinking models since the dependence of the cyclization J factor on the bending free energy is very complicated. To overcome this problem, we propose a modified DNA cyclization experi-

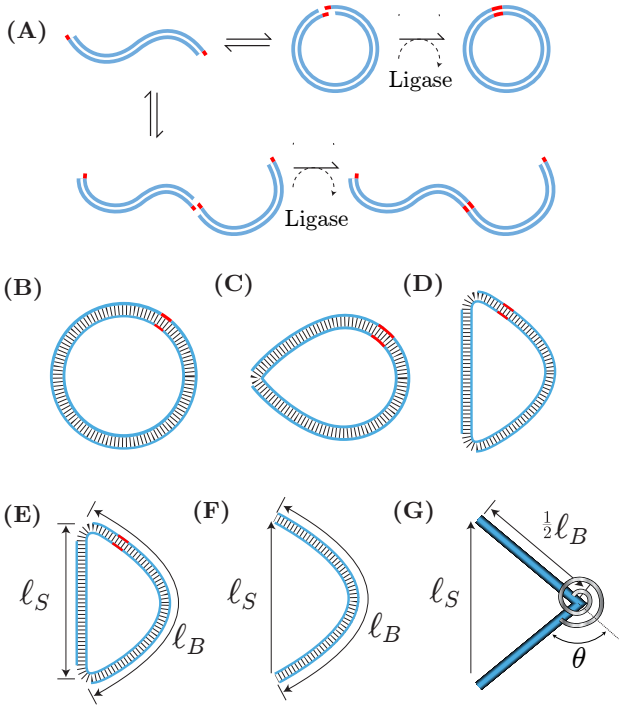

FIG. 3: (Color online) **Panel A: Cyclization Reaction.** The single-stranded complementary sequence (red) at the ends of a linear monomer can anneal to form either a linear dimer or a monomer circle. These intermediates are ligated irreversibly at limiting ligase concentration. **Panel B-D:** DNA cycles with no (circle), one (tear drop) and two (bow) single stranded gaps. Panel D corresponds to Double-Gap Cyclization. **Panel E:** The bow configuration is parameterized by a string length, ℓ_S and a bow length ℓ_B. The bending of the longer bow segment puts the string under tension. **Panel G:** For our analytic treatment, we approximate the bow configuration as two rigid segments (length $\frac{1}{2}\ell_B$) bending a torsional spring to angular displacement $\theta(\ell_B, \ell_S)$.

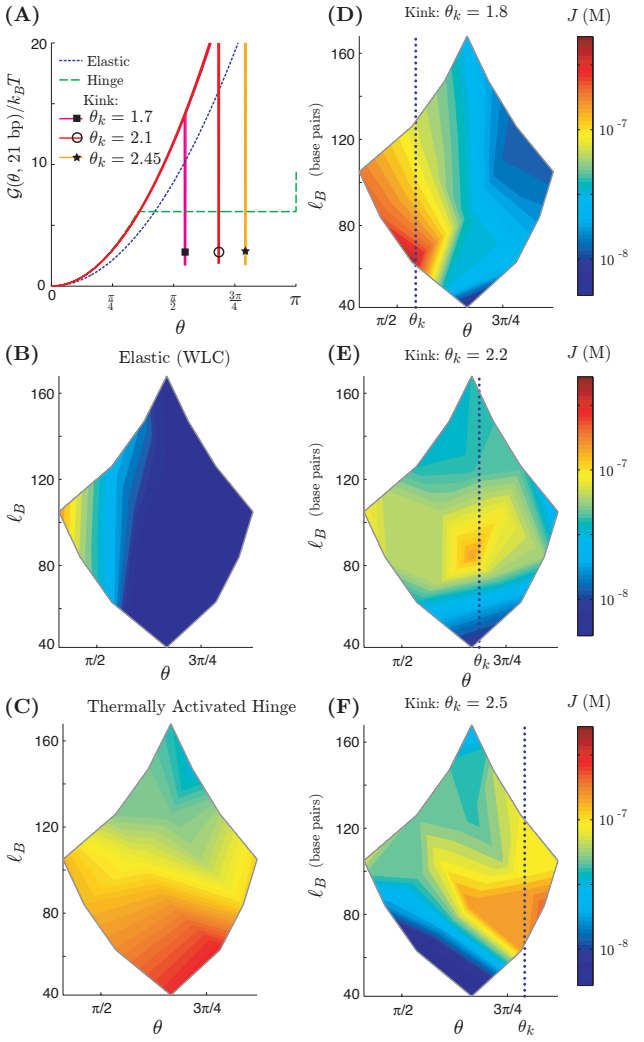

FIG. 4: (Color online) **Panel A:** DNA bending free energy as a function of bend angle for 21 bp DNA segments. All theories have the same bend persistence length, but Models II and III kink when tightly bent. Model II has no preferred kick angle and the free energy of kinking is independent of bend angle above 1 rad. Model III considers three different well defined kink angles. **Panel B-F:** J factor for models plotted as a function of bow length and bend angle. WLC (Model I) has an extremely low J factor for all large bend angles compared to the kinking models. Model II, which has no preferred kick angle, increases in J factor towards short contour length due to entropy [20]. In Model III, with well defined kink angles, the J factor is peaked in the vicinity of the kink angle, a clear signature of kinking.

ment, termed 'Double-Gap Cyclization', that facilitates the measurement of the DNA conformational energy for a wide range of curvatures at fixed contour length. In brief, Double-Gap Cyclization uses double-stranded DNA that contains two nicks which serve as flexible joints (see the Appendix for details). Due to the stiffness of the DNA at short length scales, the cyclized conformation will consist of a short 'string' under tension, and a longer 'bow' whose bending is mostly localized near the midpoint, as illustrated in Figure 3.

To investigate whether Double-Gap Cyclization could detect DNA kinks, we constructed a number of DNA theories with physiological dsDNA persistence length, but with significantly different bending free energies [34]. In addition to the elastic rod theory (WLC), we considered three examples of models with well defined kink angles, and an example with a thermally-activated flexible hinge. The bending energies are shown in Fig. 4, Panel A.

For each DNA model we considered, we computed the Double-Gap Cyclization J factor using our calculator's Monte Carlo method. We discretized the DNA backbone using a segment length of 21 base pairs, corresponding to two complete helical turns of the DNA, which allowed us to treat the DNA as an isotropic rod and ignore helical phasing between segments. To improve statistics, we biased the sampling distribution by reducing the bending energy by a factor of 1.5. For each DNA bending model we measured J factors over a wide range of bow

and string lengths using 10 million simulated conformations. The results are shown in Panels B-F of Fig. 4.

In DNA models that soften at high bend angles, a significant enhancement in cyclization (large J factor) was observed. In theories with well defined kink angles, this enhancement is strongest around the kink angle, as predicted by an approximate solution of the J factor that we derived (see Appendix). The additional bending of the DNA chain outside the kink broadens the kink angle peak. These calculations predict that if strong DNA kinking does occur then it should be observable by Double-Gap Cyclization measurements, and that any preferred kink angle can be directly measured by comparing the cyclization rates between different lengths of the bow and string.

The minimum bow segment length in Double-Gap Cyclization sets the contour length resolution of the bending free energy measurement. Are there significant limits to the resolution of this technique? One important limitation to a technique that incorporates ssDNA gaps in the substrate is the melting of DNA at the gap leading to the conversion of dsDNA bow to ssDNA string. We can set an approximate limit to our resolution by setting the DNA melting energy per base pair to the force generated by the bow. We estimate that Double-Gap Cyclization could be used to probe the mechanics of segments as short as 50 bp.

Conclusion

Our goal in writing our end-statistics calculator was to provide a comprehensive and user-friendly tool for computing end-to-end statistics of DNA. Using our calculator, one can compute a variety of statistics of biological importance, such as DNA-looping free energies for transcription factors, and their dependence on sequence [23], the presence of kinking proteins, etc. Using our calculator, we have proposed a modification to the canonical DNA cyclization experiment that directly measures the bending free energy as a function of bend angle. We demonstrate that these results could be used to bootstrap to a new theory of DNA bending describing DNA polymer statistics on the short length scales required for understanding protein-mediated DNA bending.

In closing, we want to emphasize that although our program was designed with DNA in mind, it can equally be applied to endpoint statistics of any polymer whatsoever, as long as the phantom chain approximation is valid.

Acknowledgments

PAW would like to thank the late Jon Widom for many stimulating conversations about DNA mechanics. The authors are especially grateful to Henrik Vestermark for the use of his complex root solver. This work was supported by the National Science Foundation under Grant PHY-084845.

Appendix: Cyclization experiments and J factors

Gapless Cyclization J factor: To measure the cyclization J Factor, a DNA fragment (monomer) with complementary single-stranded ends is engineered. The ends of monomers can anneal to form two species: a monomer circle (cyclization) and a linear dimer. (We will consistently ignore higher-order species such as trimers as they do not affect the J Factor.) To measure the equilibrium monomer, circle, and dimer concentrations, the molecules are ligated at limiting ligase concentrations [24]. (T4 DNA ligase forms two covalent phosphodiester bonds between the 3' hydroxyl end of one nucleotide and the 5' phosphate end of another.) Under this rapid pre-equilibrium kinetic scheme, the product of annealed circle and dimer is proportional to their equilibrium concentrations [24]. See Fig. 3 for a schematic picture of the cyclization experiment. The cyclization J factor is

$$J(\ell) = 8\pi^2 \left[p(\Delta\mathbf{R}, \vec{\Omega}_\ell; \vec{\Omega}_0, \ell) \right]_{\Delta\mathbf{R}=0}^{\Omega_\ell=\Omega_0} \tag{16}$$

Single-Gap Cyclization J factor: A second class of experimentally tractable cyclization reactions has long been recognized [24–26]: A linear monomer can form a cyclic molecule without enforcing the alignment of the end orientation while maintaining the end-to-end closure constraint. Experimentally, this can be realized by linking the ends of a dsDNA segment by a short single stranded gap since ssDNA has a short persistence length [27–29] [35].

The lowest free energy configuration for an elastic rod closed at a flexible hinge is a tear-drop shaped conformation shown schematically in Panel C, Fig. 3. The cyclization J factor for such a reaction can be calculated in terms of the position-only distribution functions:

$$J(\ell) = \left[p(\Delta\mathbf{R}; \ell) \right]_{\Delta\mathbf{R}=0} . \tag{17}$$

Together the DNA circle and tear-drop provide two distinct amplitudes that are highly sensitive to DNA elasticity [36], but these two numbers are not sufficient to determine the DNA bending free energy alone [30].

Double-Gap Cyclization J factor: Inspired by the inclusion of a single single-stranded gap in the DNA backbone, we investigated the cyclization J factor with substrates with two internal single-stranded gaps to form a bow conformation, as shown in Fig. 3, Panel D. In the cyclized molecule, we will denote the shorter double region as the string, with length ℓ_S, and the longer region as the bow, with length ℓ_B. (Please see Fig. 3, Panel E.) In the short-contour-length limit, the dominant bow

molecular conformation will have a straight string. We can therefore approximate the bow free energy by replacing the string with a rigid rod length ℓ_S. (Please see Fig. 3, Panel F.) In this approximation, the Double-Gap Cyclization J Factor has an extremely simple form:

$$J(\ell_S, \ell_B) = [p(\Delta\mathbf{R}; \ell_B)]_{|\Delta\mathbf{R}|=\ell_S}\,, \tag{18}$$

which depends only on the magnitude of the displacement $\Delta\mathbf{R}$. The J factor is therefore a direct measure of the end-to-end spatial distribution function in the moderate to high bending regime ($\ell_B < \ell_p$). The bend angle is controlled by the length of the bow string ℓ_S. See Fig. 3, Panel G.

The spatial evolution of an inextensible polymer such as DNA depends entirely on the more fundamental angular distribution function, which in our case reduces to $p(\theta)$. We therefore need to reinterpret Eqn. 18 in terms of the angular distribution in order to build a theory of DNA statistics (e.g. [30]), which is possible at short length scales where there is a simple one-to-one mapping between points in the spatial and angular distributions. We have performed this type of analysis in another context: Determining the bending energy of DNA immobilized to a mica surface [31].

The short-contour-length limit guarantees that the DNA string will be nearly straight, and the bending will be concentrated near the middle of the DNA bow. We can therefore bootstrap a theory of chain statistics from the spatial distribution function by discretizing a given DNA chain into links ℓ_B in length, each bent in the middle by a torsional spring. Note that this is a coarser approximation than was used in the Monte Carlo results shown in Figure 4. We interpret the J factor in Eqn. 18 as the spatial distribution function for two adjacent half links, as illustrated in Fig. 3, Panel E. We can compute the spatial distribution function and J factor exactly:

$$J = \frac{4p(\theta; \ell_B)\sin\frac{1}{2}\theta}{\ell_B^3 \sin\theta}, \tag{19}$$

where $p(\theta; \ell_B)$ is the single-segment bending distribution for the theory of discrete links, and θ is the angle between the chain tangents. By equating the link model J factor (Eqn. 19) to the measured J factor we can infer an effective bootstrapped angular distribution function:

$$p^*(\theta; \ell_B) \equiv \frac{\ell_B^3 J(\ell_S, \ell_B)\sin\theta}{4\sin\frac{1}{2}\theta}, \tag{20}$$

where the bend angle is defined: $\theta \equiv 2\cos^{-1}(\ell_S/\ell_B)$ as shown in Fig. 3, Panel G. We define the bending free energy at contour length scale ℓ_B (e.g. [30]):

$$\mathcal{G}^*(\theta; \ell_B) \equiv -k_B T \log p^*(\theta; \ell_B), \tag{21}$$

where k_B is the Boltzmann constant. Although measurements of the bending free energy using Double-Gap Cyclization are better for large angles, there is an important constraint that determines the angular distribution function for small deflections: the persistence length (e.g. [30]). In summary, measurement of Double-Gap Cyclization determines the bending free energy and can be used to define a theory of DNA statistics at longer contour length.

[1] Hernan G Garcia, Paul Grayson, Lin Han, Mandar Inamdar, Jané Kondev, Philip C Nelson, Rob Phillips, Jonathan Widom, and Paul A Wiggins. Biological consequences of tightly bent dna: the other life of a macromolecular celebrity. *Biopolymers*, 85(2):115–30, Feb 2007.

[2] L. Saiz and J.M.G. Vilar. DNA looping: the consequences and its control. *Current opinion in structural biology*, 16(3):344–350, 2006.

[3] R. Kavenoff and B.C. Bowen. Electron microscopy of membrane-free folded chromosomes from escherichia coli. *Chromosoma*, 59(2):89–101, 1976.

[4] P.J.J. Robinson, L. Fairall, V.A.T. Huynh, and D. Rhodes. Em measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. *PNAS*, 103(17):6506–6511, 2006.

[5] M.S. Luijsterburg, M.C. Noom, G.J.L. Wuite, and R.T. Dame. The architectural role of nucleoid-associated proteins in the organization of bacterial chromatin: a molecular perspective. *Journal of structural biology*, 156(2):262–272, 2006.

[6] H. Jacobson and W.H. Stockmayer. Intramolecular reaction in polycondensations. i. the theory of linear systems. *The Journal of Chemical Physics*, 18:1600, 1950.

[7] H. Yamakawa. *Helical wormlike chains in polymer solutions*. Springer Berlin, 1997.

[8] Hagerman,P. (1988) Flexibility of DNA *Annu. Rev. Biophys. Biophys. Chem.*, **17**, 265–286.

[9] Spakowitz,A.J. and Wang,Z.G.. End-to-end distance vector distribution with fixed end orientations for the wormlike chain model. *Physical Review E*, 72(4):041802, 2005.

[10] A.J. Spakowitz. Wormlike chain statistics with twist and fixed ends. *Europhysics Letters*, 73(5):684–690, 2006.

[11] Y. Zhang and D.M. Crothers. Statistical mechanics of sequence-dependent circular DNA and its application for DNA cyclization. *Biophysical Journal*, 84(1):136–153, 2003.

[12] Olson,W., Gorin,A., Lu,X., Hock,L., and Zhurkin,V. (1998) DNA sequence-dependent deformability deduced from protein–DNA crystal complexes *Proc. Natl. Acad. Sci. USA*, **95**, 11163-11168.

[13] El Hassan,M. and Calladine,C. (1995) The assessment of the geometry of dinucleotide steps in double-helical DNA; a new local calculation scheme *J. Mol. Biol.*, **251**, 648–664.

[14] See Supplemental Material at [URL will be inserted by publisher] for the formulae used by our harmonic approximation implementation to calculate probability densities from very general constraints on the polymer conformation.

[15] Galassi,M. et al. (2009) *GNU Scientific Library Reference Manual* Network Theory Ltd, UK.

[16] I. Wolfram Research. (2004) *Mathematica* Wolfram Research, Inc., Champaign, IL, USA.

[17] F. H. Crick and A. Klug, Nature **255**, 530 (1975).

[18] R. L. Fosdick and R. D. James, J. Elast. **11**, 165 (1981).

[19] J. Yan and J. F. Marko, Phys Rev Lett **93**, 108108 (2004).

[20] P.A. Wiggins, R. Phillips, and P.C. Nelson, Physical Review E **71**, 021909 (2005).

[21] Q. Du, C. Smith, N. Shiffeldrim, M. Vologodskaia, and A. Vologodskii, Proc Natl Acad Sci U S A **102**, 5397 (2005).

[22] Q. Du, A. Kotlyar, and A. Vologodskii, Nucleic Acids Res **36**, 1120 (2008).

[23] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A.C. Thåstr "om, Y. Field, I.K. Moore, J.P.Z. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, 2006.

[24] D. Shore, J. Langowski, and R. L. Baldwin, Proc Natl Acad Sci U S A **78**, 4833 (1981).

[25] J. Shimada and H. Yamakawa, Macromolecules **17**, 689 (1984).

[26] H. Qu, C. Tseng, Y. Wang, A. Levine, and G. Zocchi, EPL (Europhysics Letters) **90**, 18003 (2010).

[27] S. B. Smith, Y. J. Cui, and C. Bustamante, Science **271**, 795 (1996).

[28] C. Rivetti, C. Walker, and C. Bustamante, J Mol Biol **280**, 41 (1998).

[29] Q. Du, M. Vologodskaia, H. Kuhn, M. Frank-Kamenetskii, and A. Vologodskii, Biophys J **88**, 4137 (2005).

[30] P. A. Wiggins and P. C. Nelson, Phys Rev E Stat Nonlin Soft Matter Phys **73**, 031906 (2006).

[31] P. A. Wiggins, T. van der Heijden, F. Moreno-Herrero, A. Spakowitz, R. Phillips, J. Widom, C. Dekker, and P. C. Nelson, Nat Nanotechnol **1**, 137 (2006).

[32] M. Doi and S.F. Edwards. *The theory of polymer dynamics*. Oxford University Press, USA, 1988.

[33] The two arms to either side of the constrained loci are irrelevant if one neglects nonlocal forces such as those due to excluded volume. We will consistently make these approximations here, partly because neglecting excluded volume makes calculation of statistics much more tractable, and allows the user to ignore the DNA outside the two loci of interest. There are several situations where we can justifiably neglect excluded volume. One such situation occurs when the two loci are closer together than a couple of the persistence lengths (bending scale), so that self-overlap between the intervening DNA is unlikely, and when forces by the outside DNA are either ignorable (due to a low density) or isotropic. Surprisingly, in the opposite case of a very high DNA density, excluded volume is also believed to be ignorable [32], because the effective external pressure is isotropic and equal over all regions of the polymer. The twin strands of supercoiled DNA are supported by contact with one another which precludes the direct use of a phantom chain model, but the effective polymer composed of the forward and backward strands together may be treated by our methods. Supercoiled DNA may also branch, but our calculator does not treat branched polymers.

[34] Due to the action of the Renormalization Group, the DNA bending energy must be defined at a contour length scale [30]. We defined the bending free energy at the two helical repeat length scale.

[35] Even though it is convenient to discuss the reaction as if ligation occurs at this single stranded gap, it is clear from the free energy interpretation of the J Factor that a monomer with complementary single-stranded ends, but including an internal single-stranded gap, can be used as the substrate in a standard cyclization reaction [29].

[36] At first glance it may appear that these two measure-

ments would probe somewhat degenerate DNA physics, but we have previously demonstrated that DNA kinking tends to lead to a greater enhancement of the tear drop than the DNA circle.