

SCIENTIFIC REPORTS



OPEN

Evaluating the role of coherent delocalized phonon-like modes in DNA cyclization

Ludmil B. Alexandrov^{1,2}, Kim Ø. Rasmussen¹, Alan R. Bishop¹ & Boian S. Alexandrov^{1,2}

The innate flexibility of a DNA sequence is quantified by the Jacobson-Stockmayer's J-factor, which measures the propensity for DNA loop formation. Recent studies of ultra-short DNA sequences revealed a discrepancy of up to six orders of magnitude between experimentally measured and theoretically predicted J-factors. These large differences suggest that, in addition to the elastic moduli of the double helix, other factors contribute to loop formation. Here, we develop a new theoretical model that explores how coherent delocalized phonon-like modes in DNA provide single-stranded "flexible hinges" to assist in loop formation. We combine the Czapla-Swigon-Olson structural model of DNA with our extended Peyrard-Bishop-Dauxois model and, without changing any of the parameters of the two models, apply this new computational framework to 86 experimentally characterized DNA sequences. Our results demonstrate that the new computational framework can predict J-factors within an order of magnitude of experimental measurements for most ultra-short DNA sequences, while continuing to accurately describe the J-factors of longer sequences. Further, we demonstrate that our computational framework can be used to describe the cyclization of DNA sequences that contain a base pair mismatch. Overall, our results support the conclusion that coherent delocalized phonon-like modes play an important role in DNA cyclization.

The flexibility of the DNA molecule plays an important role in a multitude of biological functions as well as in the compact storage of the genetic material of cells¹. For example, sharply bending DNA around nucleosomes is used to tightly pack the genome in each eukaryotic cell², while DNA bending by transcription factors is a common cellular mechanism that participates in regulation of gene expression³⁻⁶. There have been numerous experimental and theoretical studies examining the thermodynamics of DNA bending⁷⁻²⁶. In the majority of cases, bending of DNA results in loop formations with lengths shorter than DNA persistence length²⁷. Such loops contradict the classical polymer physics view, in which double-stranded DNA (dsDNA) is virtually unbendable at scales below its persistence length.

Due to its long persistence length of approximately 150 base pairs (bp), dsDNA has typically been modeled as an elastic rod with mechanical properties well described by the wormlike chain model (WLC)²⁸. Within the WLC model, the conformational properties of a sequence depend entirely on the notion of a persistence length, and hence, DNA loops with lengths shorter than 150 bp are energetically extremely costly and the probability for their spontaneous thermodynamic creation is vanishingly small²⁹. As such, the WLC model's prediction of Jacobson-Stockmayer's J-factors (which measures the propensity for DNA loop formation, see ref. 30) for short DNA sequences is in a sharp contrast to several *in vivo* and *in vitro* experimental observations⁹⁻¹⁶. Thus, the basic WLC model is unable to accurately characterize the cyclization of short DNA fragments, although it consistently describes the cyclization of long DNA segments.

The basic WLC model ignores both the nucleotide sequence and the three-dimensional (3D) helical structure of DNA, both of which are known to affect cyclization rates. For example, ignoring the 3D nature of DNA results in disregarding the proper torsional orientation of DNA segments, and hence does not allow describing the experimentally observed oscillations of the cyclization rates due to the approximately 10 bp torsional period of the double helix³¹. Further, the lack of nucleotide sequence and intrinsic curvature information in the basic WLC model limits its predictive capability, since experimental studies have shown that periodic stretches of consecutive adenine-thymine base pairs possess curved equilibrium conformations³². Such periodic DNA sequences

¹Theoretical Division, Los Alamos National Laboratory, Los Alamos, 87545, United States of America. ²University of New Mexico Comprehensive Cancer Center, Albuquerque, NM, 87102, USA. Correspondence and requests for materials should be addressed to B.S.A. (email: boian@lanl.gov)

exhibit a natural sequence-dependent static bending, also known as intrinsic curvature, that can result in higher cyclization rates³³.

Incorporating the 3D structure and sequence-dependent intrinsic curvature of DNA allows a better representation of the nature of the elasticity of the double helix³⁴. As a first approximation, the intrinsic curvature can be considered as an equilibrium property of each DNA segment³⁵. To more realistically calculate cyclization properties of DNA, a coarse-grained Monte Carlo approach incorporating the 3D structure, intrinsic curvature, and DNA sequence was developed by Levenet, Crothers, and Zhang^{36,37}, by Manning, Maddocks, and Kahn³⁸, and by Czapla, Swigon, and Olson³⁹.

Recently, we applied the Czapla-Swigon-Olson (CSO) model³⁹ to perform an extensive characterization of the effect that distinct sets of structural parameters have on estimating J-factors of DNA sequences with different lengths²⁴. Our analysis encompassed seven sets of structural and conformational parameters to calculate the J-factors of 86 DNA segments with experimentally measured J-factors, with lengths between 50 bp and 330 bp, as well as of 20,000 *in silico* generated random sequences, with lengths between 350 bp and 4,000 bp. Our analysis revealed that a set of parameters based on an examination of nucleosome positioning⁴⁰ provides results most consistent with the experimental measurements. Importantly, we showed that all sets of structural parameters perform well for the *in silico* generated sequences longer than ~350 bp and that the choice of structural parameters and the nucleotide sequence of the DNA segment makes little difference in estimating the cyclization of these long sequences. In contrast, all investigated sets of structural parameters performed poorly when used for estimating J-factors of DNA segments shorter than ~100 bp, yielding predictions with up to six orders of magnitude difference compared to experimental findings. The inability to describe the cyclization rates of ultra-short DNA sequences suggests that taking into account the intrinsic curvature and 3D structure of DNA fails to capture the salient physics for short DNA segments.

In this paper, we develop a new computational approach for evaluating DNA cyclization by linking the CSO model³⁹ with our Extended Peyrard-Bishop-Dauxois (EPBD) approach^{41–43}. The CSO model accounts for DNA intrinsic curvature and 3D structure, while the EPBD approach accurately describes DNA opening dynamics that produces local single-stranded regions (i.e., DNA bubbles) in the double helix⁴⁴. This coupled CSO-EPBD model is applied to 86 DNA sequences with experimentally characterized J-factors previously examined in ref. 24 (Supplementary Table S1). Our analysis demonstrates that, without changing any of the original parameters in the CSO model or the EPBD model, the coupled CSO-EPBD model is able to accurately determine the J-factors of ultra-short DNA sequences with most predictions being within an order of magnitude of experimental measurements. Further, the coupled CSO-EPBD model continues to accurately describe the J-factors of longer sequences, and we also demonstrate that this model is applicable to DNA sequences containing a base pair mismatch.

Results

Coherent delocalized phonon-like modes and DNA looping. The structural integrity of biological macromolecules is primarily governed by hydrogen bonds (H-bonds), which have natural vibration frequencies in the terahertz range^{45,46}. H-bonds are much weaker (~few $k_B T$ s) than covalent bonds, causing these macromolecules to experience slow conformational motion resulting from the inherent thermal fluctuations at biological temperatures. The DNA molecule's slow motions induce spontaneous openings and re-closings of the double helix known as “DNA breathing”^{47,48}, or spontaneous “base flipping”⁴⁹, or “DNA bubbles”⁴⁵, or “coherent delocalized phonon-like modes”⁴⁵ (The term “delocalized” refers to the fact that the low-frequency (THz) vibrational modes of the bio-macromolecules are indeed delocalized in the wavenumber space, while they are localized in the real space, forming DNA bubbles that coherently span only a few base pairs). The frequency of occurrence of these transient openings depend entropically both on the local sequence and its structure. The propensity for DNA bubbles is related to the local stability and flexibility of the molecule⁵⁰ and it plays a key role in many biological processes such as: reading of the genetic code (transcription)^{51–53}, DNA production (replication)^{54,55}, protein-DNA binding^{56–60}, DNA methylation⁶¹, and DNA repair mechanisms⁶².

It has been proposed that the propensity for DNA bubbles and the flexibility of a DNA segment are interrelated^{8,63}. As such, in addition to the elastic moduli of the double helix, loop formations are facilitated by “flexible hinges”. In turn, these “flexible hinges” are due to the fluctuation of DNA base-pair openings resulting in more flexible single-strand DNA (ssDNA) segments in the, otherwise, rigid dsDNA. Thus, the experimental results demonstrating unusually high J-factors for short DNA sequences may be due to spontaneous internal strand-separation. Although energetically expensive and therefore rare, such strand-separation can be entropically favorable and may assist DNA local bending when forming loops between 70 and 100 bp. As explained in ref. 8 (consistent with the values obtained in experiments) the formation of 3 bp ssDNA regions (i.e., DNA bubbles) are associated with a free energy cost of ~6 $k_B T$ to 15 $k_B T$. Hence, modified WLC approaches incorporating DNA's ability to spontaneously develop local ssDNA regions, realized via parameters for a phenomenological density of hinges or kinks, have been explored with some success^{8,13,19}. However, these models largely ignored the DNA sequence as well as the intrinsic curvature and 3D structure of the double helix.

Developing a predictive understanding mandates a more specific ability to characterize the sequence-dependent dynamics of the transient opening and re-closing of dsDNA while also accounting for DNA's intrinsic curvature and 3D structure. Here, by combining the CSO and EPBD models, we develop such an approach and evaluate its predictive capabilities on DNA sequences with previously measured J-factors.

Description of the CSO-EPBD model. The CSO model (see Methods) utilizes equal standard variations for the fluctuations of the tilt and roll angles (i.e., isotropic bending). The model uses a 147 bp persistence length for dsDNA and, as such, mandates a root-mean-square fluctuation of 4.84° for the tilt and roll angles. In contrast, the CSO model standardly uses 4.09° as the root-mean-square fluctuations for the twist angle. Our CSO-EPBD model builds upon the original CSO model but it also includes coherent delocalized phonon-like modes in DNA,

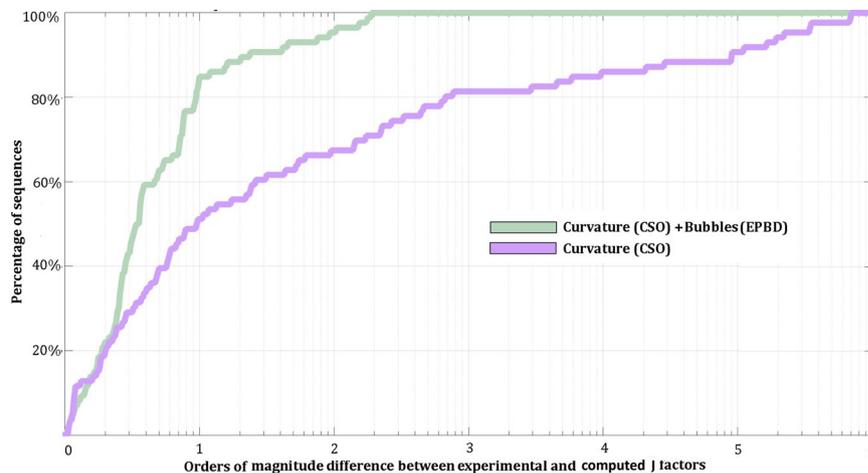


Figure 1. Estimated J-factors for all examined DNA sequences. The two different curves with colors correspond to the CSO and CSO-EPBD model calculations. The y-axis reflects the orders of magnitude difference between experimentally measured and computationally derived J-factors. The x-axis corresponds to the percentage of sequences for a given order of magnitude difference.

which correspond to entropically-driven single stranded regions in the DNA sequence. These single stranded regions are more flexible than dsDNA and act as “flexible hinges” in the otherwise rigid dsDNA molecule since the persistence length of single stranded DNA (ssDNA) is only ~ 3 bp⁶⁴. To incorporate the ssDNA persistent length in our CSO-EPBD framework, the root-mean-square for the roll and tilt angles were set to $\sim 30.08^\circ$, while $\sim 27.96^\circ$ was used for the root-mean-square of the twist angle (both consistent with the assumptions in ref. 39). Note that, in this study, we do not use the naturally straight DNA model that was one of the models proposed in the original CSO publication³⁹. Rather, we use the CSO model with experimentally measured parameters to account for presence of static curvature in the DNA fragments.

Our EPBD model (see Methods) describes the transient base-pair openings of the double helix by quantifying the probability for each base pair to be in a single-stranded state and hence to act as a hinge. As such, for each examined DNA sequence, we generated a profile that captures the probability for each base pair to be in a single-stranded state due to thermal fluctuations of the double helix.

The EPBD probability profiles (Supplementary Table S2) were imbedded in the CSO model by using a Markov Chain Monte Carlo approach. For each base pair in a sequence, a number was generated from a uniform random distribution between 0 and 1 and this number was compared to the EPBD opening probability for this base pair. When this random number was lower than or equal to the EPBD opening probability for this base pair, the DNA base-pair was considered to be in a single-stranded state and the root-mean-square values for ssDNA were used for this base pair when generating a particular chain. Conversely, when the random number was higher than the EPBD opening probability, the DNA base-pair was considered to be in a double-stranded state and the root-mean-square values for dsDNA were used for this base pair. In all cases, we used the same set of structural DNA parameters previously shown to yield optimal results when compared to experimental measurements (Supplementary Table S3). It is important to note that by associating the EPBD-generated opening profiles with the hinge density in the formulation of Yan and Marco^{8,65}, we obtained (by using Eq. (33) in ref. 65) that the free energy of the transient double-strand openings, calculated by EPBD derived probabilities, requires on average $\sim 10 k_B T$, in agreement with arguments made in refs 10 and 65.

Applying the CSO-EPBD model to DNA sequences with different lengths. Our implementation of the new CSO-EPBD model was used to calculate the J-factors of 86 DNA sequences with experimentally determined J-factors and lengths between 50 bp and 325 bp. Comparisons were performed between experimentally determined J-factors, J-factors derived using the original CSO model, and J-factors derived using the CSO-EPBD model (Supplementary Table S4). These comparisons show that the J-factors calculated using the CSO-EPBD model are much closer to experimental measurements compared to the J-factors calculated using the original CSO model (Supplementary Table S4). To quantify the differences between the two *in silico* calculated J-factors and the experimental measurements, we calculated the percentage of sequences for which their computationally estimated J-factors are within a particular absolute distance from the actual experimental measurements (Fig. 1). The results based on the CSO-EPBD model are much closer to experimental measurements: $\sim 85\%$ of J-factors are within an order of magnitude of experimentally measured values. In contrast, only $\sim 51\%$ of the J-factors calculated by the original CSO model are within the same range.

Importantly, both the CSO-EPBD model and the original CSO model perform equally well for sequences with lengths longer than 100 bp (Fig. 2). However, the original CSO model is unable to accurately describe J-factors of sequences shorter than 100 bp since less than $\sim 25\%$ of sequences are within an order of magnitude of experimental measurements (Fig. 3). In contrast, the results from the CSO-EPBD model are within an order of magnitude for $\sim 82\%$ of all examined ultra-short sequences. Further, the J-factors of the remaining $\sim 18\%$ of ultra-short sequences are within approximately two orders of magnitude from the experimental measurements. In contrast,

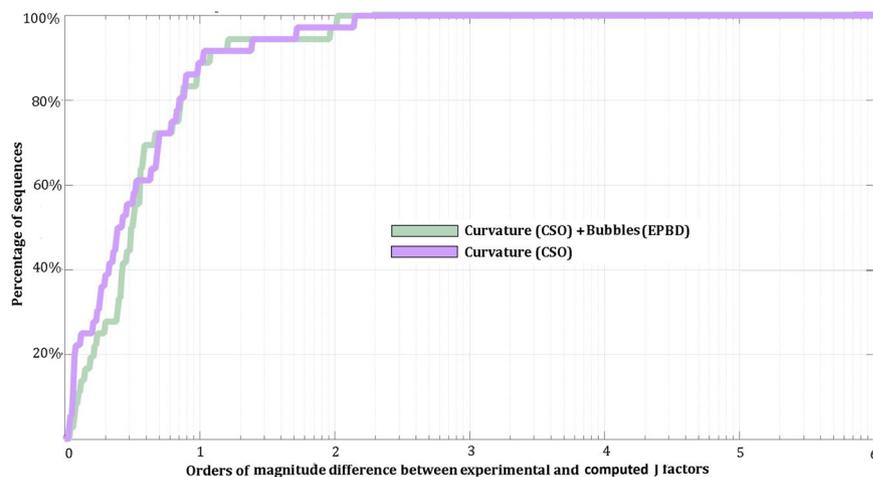


Figure 2. Estimated J-factors for all sequences with lengths longer than 100 bp. The two different curves with colors correspond to the CSO and CSO-EPBD model calculations. The y-axis reflects the orders of magnitude difference between experimentally measured and computationally derived J-factors. The x-axis corresponds to the percentage of sequences for a given order of magnitude difference.

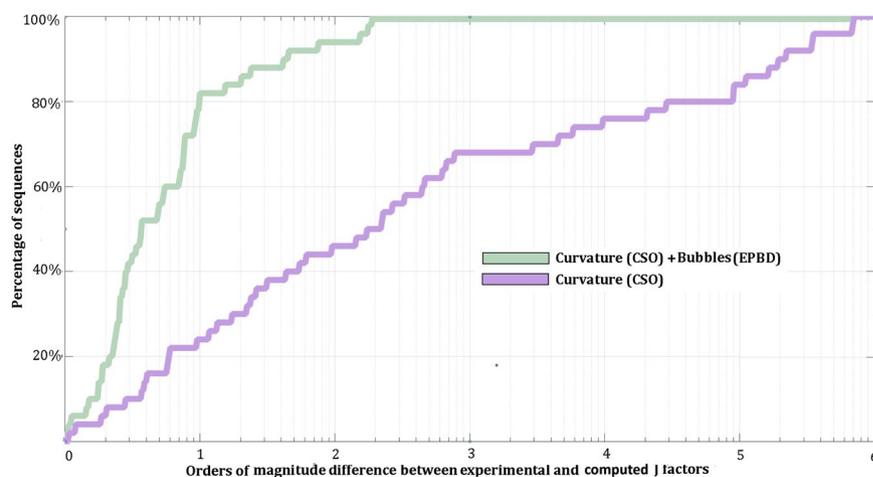


Figure 3. Estimated J-factors for all sequences with lengths shorter than 100 bp. The two different curves with colors correspond to the CSO and CSO-EPBD model calculations. The y-axis reflects the orders of magnitude difference between experimentally measured and computationally derived J-factors. The x-axis corresponds to the percentage of sequences for a given order of magnitude difference.

J-factors calculated by the CSO model for many sequences are four, five, and even six orders of magnitude smaller than the experimental results. Overall, these results support that our CSO-EPBD model provides a significantly better description of DNA cyclization for ultra-short segments compared to the original CSO model (Fig. 3). It is also noteworthy that the CSO-EPBD model is able to describe the J-factors of ultra-short DNA sequences almost as well as the ones of long DNA sequences: ~82% of ultra-short versus ~89% of long sequences are within an order of magnitude of experimental measurements.

Applying the CSO-EPBD model to a DNA sequence with a base pair mismatch. One limitation of most previous models is that they cannot be used to examine the cyclization of a DNA sequence containing a specific nucleotide modification, for example, a mismatched base pair. In contrast, our new CSO-EPBD model can be used to calculate the J-factors of DNA sequences with the presence of nucleotide modifications. Here, our EPBD simulations take into account a sequence mismatch by assigning a probability of one to observe this base pair open in the profile of a DNA sequence (i.e., the mismatch is always in a single-stranded state; see Supplementary Table S5). To evaluate the ability of CSO-EPBD to estimate the J-factors of sequences containing a mismatch base pair, we applied our model to a DNA fragment that was previously experimentally characterized for both having a C:C mismatch in the middle of the sequence as well as not having such a mismatch¹⁵. The J-factor of the sequence with a C:C mismatch, calculated by CSO-EPBD, is $4.99\text{E}-10$, which is more than an one order of magnitude increase compared to the J-factor of the sequence without a mismatch ($J = 3.32\text{E}-11$). Comparison

between the calculated J-factors by CSO-EPBD and previously performed experimental measurements¹⁵ reveals good agreement for both cases.

Discussion

In this study, we developed a new theoretical model for describing DNA cyclization by combining the EPBD model with the CSO model. This CSO-EPBD model was used to perform an extensive examination of the effect that sequence dependent coherent delocalized phonon-like modes of DNA have on estimating J-factors of DNA sequences with different lengths. We applied our implementation of the new CSO-EPBD model to 86 DNA segments with experimentally characterized J-factors, with lengths between 50 bp and 325 bp. Our analysis demonstrates that the CSO-EPBD model yields results in good agreement to experimental measurements. Further, even for sequences with lengths less than 100 bp, the CSO-EPBD model performs very well, yielding more than 82% of the calculated J-factors within an order of magnitude of experimental measurements. The analysis of J-factors for sequences longer than 100 bp shows that the CSO-EPBD model gives results indistinguishable from the results obtained by the CSO model alone.

Even though the calculated J-factors are significantly improved and within an order of magnitude from experimentally measured values, the results derived by leveraging the CSO-EPBD model still differ from the exact experimental measurements for ultra-short sequences. This suggests the need to explore more accurate models for DNA breathing and DNA curvature in the future. One possible explanation for these differences is that the set of structural parameters used in CSO-EPBD is not sufficiently accurate for describing some of the sequences. Another likely explanation is that the EPBD calculated average opening profiles neglect the effect of adjacent base pairs opening in a correlated fashion. Future studies taking into account the specific, rather than the mean, structural parameters of a DNA sequence and incorporating DNA breathing dynamics directly into the CSO model may be needed.

There are other factors that could be affecting our results. Our examination relies on previous measurements of J-factors and structural parameters. However, these experimentally derived values exhibit standard errors that were ignored in our analysis and the mean values were assumed to be representative. Further, parts of the experimental data analyzed in this paper have been considered problematic by some researchers⁶⁶. Nevertheless, assuming the accuracy of the experimentally measured J-factors as well as the mean values of the used structural parameters, the present study provides a systematic evaluation of the role of DNA bubbles in calculating cyclization rates.

Importantly, the present work, for the first time, supports a direct link between sequence-dependent bubbles and more accurate calculation of J-factors for short DNA segments. In addition, considering recent experimental observations⁴⁵ demonstrating that DNA bubbles are indeed in the terahertz range leads to an exciting avenue for future exploration of J-factors in the presence of a strong, pulsed terahertz field. Such future studies will allow elucidation of the connection between terahertz irradiation and DNA functionality⁶⁷.

Methods

The EPBD model. To describe the bubbles in DNA, here, we use the EPBD model of DNA⁴³, which is an extension of the original Peyrard-Bishop-Dauxois (PBD) DNA model^{41,42} that includes sequence-specific stacking potentials. The EPBD model is a quasi-two-dimensional nonlinear model that describes the transverse opening motion of the complementary strands of double-stranded DNA, while distinguishing the two sides (left - v_n and right - u_n) of the DNA double strand, and can describe the presence of nucleotide modifications. The potential of the EPBD model, V_{EPBD} , is (see, e.g., ref. 59)

$$V_{EPBD} = \sum_{n=1}^N U(u_n; v_n) + W(u_{n-1}, u_n; v_{n-1}, v_n), \quad (1)$$

where,

$$U(u_n; v_n) = D_n(e^{-a_n(u_n - v_n)} - 1)^2, \quad (2)$$

and,

$$W(u_{n-1}, u_n; v_{n-1}, v_n) = \frac{K_{n,n-1}^u}{2}(u_n - u_{n-1})^2 + \frac{K_{n,n-1}^v}{2}(v_n - v_{n-1})^2 + \frac{\rho}{4}e^{-\beta[(u_n - v_n) + (u_{n-1} - v_{n-1})]} \left(\sqrt{K_{n,n-1}^u}(u_n - u_{n-1}) - \sqrt{K_{n,n-1}^v}(v_n - v_{n-1}) \right)^2. \quad (3)$$

Here, the sum is over all N base pairs of the DNA sequence. For each base pair, V_{EPBD} includes two degree of freedom; u_n and v_n representing the relative displacement from the equilibrium of the corresponding nucleotide, located, respectively, in the right or left strand of the double helix. These displacements quantify the transverse stretching of the hydrogen bonds between complementary nucleotides. The first term, $U(u_n; v_n)$, is a Morse potential for the n^{th} base pair that represents the combined effects of the hydrogen bonds between the complementary bases and electrostatic repulsion of the backbone phosphates. The parameters D_n and a_n depend on the nature of the base pair (A-T versus G-C, i.e., two hydrogen bonds versus three hydrogen bonds). The second term represents a quasi-harmonic approximation of the stacking interactions between consecutive nucleotides, which influences their transverse stretching motion. The exponential term effectively decreases the stacking interaction when

one of the nucleotides is displaced away from its equilibrium position, e.g., when one of the nucleotides is out of the DNA stack. The stacking force constants, $K_{n,n-1}^u$ and $K_{n,n-1}^v$, depend on: (i) the nature of the nucleotide, (ii) on its closest neighbor, and (iii) on the location of the nucleotide - the right or left DNA strand. The dinucleotide stacking force constants were determined in ref. 43 by comparison with UV-melting curves of DNA oligomers. Here, we exploited EPBD simulations to calculate the opening probability profiles for DNA sequences investigated in ref. 24 using EPBD Markov Chain Monte Carlo (MCMC) simulations (for EPBD parameters and MCMC details see ref. 59). The opening probability derived from MCMC simulations is defined as the fraction of the simulation time a base pair remains open/stretched by more than 2.5 Å, which correspond to a local melting in the EPBD framework. Note that, as we have previously demonstrated^{44, 57, 58, 60}, the DNA local breathing/openings of a base-pair depend not only on the closest neighbors but also on the long-distance flanks of that base-pair.

The CSO model. In the CSO model³⁹, each configuration of a DNA segment with N base pairs depends on its sequence via the equilibrium (minimum energy) values of the DNA's standard conformational and structural parameters⁶⁸; twist angle, roll angle, tilt angle, shift displacement, slide displacement, and rise displacement: $\Theta^{0n} = (\theta_1^{n0}, \theta_2^{n0}, \theta_3^{n0}, \theta_4^{n0}, \theta_5^{n0}, \theta_6^{n0})$. For each, n^{th} bp these parameters are defined in relation to the previous $(n-1)^{\text{th}}$ bp and characterize the static bending. We previously demonstrated²⁴ that the CSO model with trinucleotide equilibrium values, Θ^{0n} , taken from ref. 69 exhibits a superior performance. This study takes the same approach as in ref. 24 and it uses trinucleotide dependencies where, for each XYZ trinucleotide sequence, specific values are assigned to the twist angle, roll angle, tilt angle, shift displacement, slide displacement, and rise displacement (Supplementary Table S3) based on the experimental measurements in ref. 40.

The deviations from these equilibrium values, $\Delta\Theta$, are caused by the thermal fluctuations and are controlled by the DNA elastic moduli, f_{ij}^n , where, f_{ij}^n are the elements of the symmetric 6×6 elastic force matrix F . The energy of each base pair, G , in harmonic approximation depends only on these deviations, $\Delta\Theta$, and is given by,

$$G_n(\Theta^{st}, F, \Theta^{0n}) = \frac{1}{2} \sum_{i=1}^6 \sum_{j=1}^6 f_{ij}^n \Delta\theta_i^n \Delta\theta_j^n, \quad (4)$$

$$\Delta\theta_i^n \equiv \theta_i^{st} - \theta_i^{n0}. \quad (5)$$

The equilibrium parameters θ^{n0} define the static bending and hence the intrinsic curvature of the DNA segment, while θ^{st} are stochastic variables representing the fluctuations. After a diagonalization of the force-constant matrix F , the energy for each consecutive base pairs can be written in terms of a diagonal matrix D and normal variables ω , as: $G_n = \frac{1}{2} \Omega^T D \Omega$. Therefore, the probability for a given configuration becomes,

$$P_n(\Theta) = \prod_i \frac{1}{\sqrt{2\pi\beta D_{ii}}} \exp^{-\frac{\beta}{2} D_{ii} \omega_i^2}. \quad (6)$$

This expression represents the probability as a product of independent terms and normal variables that describe the changes of the parameters on the directions of the principal axes of deformation. Hence, to account for the thermal fluctuations, the conformational parameters can be considered as given by normal distributions with specific expectation values and standard deviations: The expectation values define the static bending and are represented by Θ^{0n} , while the standard deviations define the flexibility and depend on DNA's elastic moduli. Originally the CSO model assumes that the fluctuations of the tilt and roll angles are equal (isotropic bending) and for the root-mean-square fluctuations of the dsDNA the value of 4.84° is used, corresponding to a persistence length of ~147 bp. Furthermore, the CSO model uses 4.09° as the root-mean-square fluctuations in the twist angle value, which corresponds to a global twisting constant compatible with previously measured equilibrium topoisomer distributions of DNA mini-circles⁷⁰.

Within the CSO model, propensity for cyclization of a specific DNA sequence is estimated by Monte Carlo simulations (for details of the algorithm see refs 24 and 39). The simulations sample the configuration space of the chains obtained by generating series of DNA sequences with structural parameters drawn from normal distributions with given expectation values and standard deviations corresponding to dsDNA. The J-factor that characterizes the cyclization can be determined^{36, 37, 39} as a product of probabilities describing the contribution of the spatial configuration:

$$J = \frac{4\pi}{N_A} W(r \approx 0) \Gamma_r(\cos\gamma \approx 1) \Phi_{r, \cos\gamma}(\phi \approx 0). \quad (7)$$

Here, $W(r \approx 0)$ is the probability for a DNA segment to be circular, i.e., to possess the end-to-end distance $r \approx 0$ (in practice, a threshold of 30 Å was used). The factor $\Gamma_r(\cos\gamma \approx 1)$ is the conditional probability that the normal vector of the first and last base pairs are (almost) aligned when the ends of the fragments coincide, i.e., the cosine of the net bending angle is ≈ 1 . In practice, a threshold of $\cos\gamma > 0.86$ was used when $r \approx 0$. The term $\Phi_{r, \cos\gamma}(\phi \approx 0)$ is the conditional probability that the first and last base pairs coincide. In practice, a threshold of $\cos\phi > 0.86$ was applied when both $r \approx 0$ and the first and the last base pairs are coplanar, i.e., $\cos\gamma \approx 1$. The factor $\frac{4\pi}{N_A}$, where N_A is Avogadro's number, is the normalization associated with the uniformly distributed probability density of bimolecular association. As in our previous works²⁴, periodic boundary conditions were used for simulating each sequence. More specifically, the first two base-pairs of each sequence were added to the end of the sequence resulting in two extra base-pairs. Note that supplementary materials (Table S2) also includes the opening profiles for these two additional base-pairs.

References

- Sinden, R. R. *DNA structure and function* (Elsevier, 2012).
- Garcia, H. G. *et al.* Biological consequences of tightly bent DNA: The other life of a macromolecular celebrity. *Biopolymers* **85**, 115–130 (2007).
- Wu, H. M. & Crothers, D. The locus of sequence-directed and protein-induced DNA bending. *Nature* **308**, 509–513 (1984).
- Hogan, M. & Austin, R. Importance of DNA stiffness in protein-DNA binding specificity. *Nature* **329**, 263–266 (1987).
- Travers, A. Protein-induced DNA bending. In *Nucleic Acids and Molecular Biology*, 136–148 (Springer, 1988).
- Kahn, J. D. & Crothers, D. M. Protein-induced bending and DNA cyclization. *Proceedings of the National Academy of Sciences* **89**, 6343–6347 (1992).
- Baumann, C. G., Smith, S. B., Bloomfield, V. A. & Bustamante, C. Ionic effects on the elasticity of single DNA molecules. *Proceedings of the National Academy of Sciences* **94**, 6185–6190 (1997).
- Yan, J. & Marko, J. F. Localized single-stranded bubble mechanism for cyclization of short double helix DNA. *Physical Review Letters* **93**, 108108 (2004).
- Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).
- Cloutier, T. E. & Widom, J. Spontaneous sharp bending of double-stranded DNA. *Molecular Cell* **14**, 355–362 (2004).
- Li, G., Levitus, M., Bustamante, C. & Widom, J. Rapid spontaneous accessibility of nucleosomal DNA. *Nature Structural & Molecular Biology* **12**, 46–53 (2005).
- Cloutier, T. E. & Widom, J. DNA twisting flexibility and the formation of sharply looped protein–DNA complexes. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 3645–3650 (2005).
- Yuan, C., Chen, H., Lou, X. W. & Archer, L. A. DNA bending stiffness on small length scales. *Physical Review Letters* **100**, 018102 (2008).
- Mathew-Fenn, R. S., Das, R. & Harbury, P. A. Remeasuring the double helix. *Science* **322**, 446–449 (2008).
- Vafabakhsh, R. & Ha, T. Extreme bendability of DNA less than 100 base pairs long revealed by single-molecule cyclization. *Science* **337**, 1097–1101 (2012).
- Mazur, A. K. & Maaloum, M. DNA flexibility on short length scales probed by atomic force microscopy. *Physical Review Letters* **112**, 068104 (2014).
- Zheng, G., Czapla, L., Srinivasan, A. & Olson, W. K. How stiff is DNA? *Physical Chemistry Chemical Physics* **12**, 1399–1406 (2010).
- Geggier, S. & Vologodskii, A. Sequence dependence of DNA bending rigidity. *Proceedings of the National Academy of Sciences* **107**, 15421–15426 (2010).
- Vologodskii, A. & Frank-Kamenetskii, M. D. Strong bending of the DNA double helix. *Nucleic Acids Research* **41**, 6785–6792 (2013).
- Shin, J., Lee, O.-C. & Sung, W. How a short double-stranded DNA bends. *The Journal of Chemical Physics* **142**, 155101 (2015).
- Rosanio, G., Widom, J. & Uhlenbeck, O. C. *In vitro* selection of DNAs with an increased propensity to form small circles. *Biopolymers* **103**, 303–320 (2015).
- Irobalieva, R. N., Fogg, J. M., Catanese, D. J. Jr. & Sutthibutpong, T. *et al.* Structural diversity of supercoiled DNA. *Nature Communications* **6**, 303–320 (2015).
- Sutthibutpong, T., Matek, C., Benham, C. & Slade, G. G. *et al.* Long-range correlations in the mechanics of small DNA circles under topological stress revealed by multi-scale simulation. *Nucleic acids research* **44**, 9121–9130 (2016).
- Alexandrov, L. B., Bishop, A. R., Rasmussen, K. Ø. & Alexandrov, B. S. The role of structural parameters in DNA cyclization. *BMC Bioinformatics* **17**, 1 (2016).
- Wang, Q. & Pettitt, B. M. Sequence affects the cyclization of DNA minicircles. *The Journal of Physical Chemistry Letters* **7**, 1042–1046 (2016).
- Jeong, J., Le, T. T. & Kim, H. D. Single-molecule fluorescence studies on DNA looping. *Methods* (2016).
- Schleif, R. DNA looping. *Annual Review of Biochemistry* **61**, 199–223 (1992).
- Shimada, J. & Yamakawa, H. Ring-closure probabilities for twisted wormlike chains. application to DNA. *Macromolecules* **17**, 689–698 (1984).
- Levene, S. D., Giovan, S. M., Hanke, A. & Shoura, M. J. The thermodynamics of DNA loop formation, from j to z. *Biochemical Society Transactions* **41**, 513–518 (2013).
- Jacobson, H. & Stockmayer, W. H. Intramolecular reaction in polycondensations. I. The theory of linear systems. *Journal of Chemical Physics* **18**, 1600–1606 (1950).
- Shore, D. & Baldwin, R. L. Energetics of DNA twisting: I. Relation between twist and cyclization probability. *Journal of Molecular Biology* **170**, 957–981 (1983).
- Drak, J. & Crothers, D. M. Helical repeat and chirality effects on DNA gel electrophoretic mobility. *Proceedings of the National Academy of Sciences* **88**, 3074–3078 (1991).
- Crothers, D. M., Drak, J., Kahn, J. D. & Levene, S. D. DNA bending, flexibility, and helical repeat by cyclization kinetics. *Methods in Enzymology* **212**, 3–29 (1992).
- Hagerman, P. J. Sequence-directed curvature of DNA. *Annual Review of Biochemistry* **59**, 755–781 (1990).
- Zhurkin, V., Ulyanov, N., Gorin, A. & Jernigan, R. Static and statistical bending of DNA evaluated by Monte Carlo simulations. *Proceedings of the National Academy of Sciences* **88**, 7046–7050 (1991).
- Levene, S. D. & Crothers, D. M. Ring closure probabilities for DNA fragments by Monte Carlo simulation. *Journal of Molecular Biology* **189**, 61–72 (1986).
- Zhang, Y. & Crothers, D. M. Statistical mechanics of sequence-dependent circular DNA and its application for DNA cyclization. *Biophysical Journal* **84**, 136–153 (2003).
- Manning, R. S., Maddocks, J. H. & Kahn, J. D. A continuum rod model of sequence-dependent DNA structure. *The Journal of Chemical Physics* **105**, 5626–5646 (1996).
- Czapla, L., Swigon, D. & Olson, W. K. Sequence-dependent effects in the cyclization of short DNA. *Journal of Chemical Theory and Computation* **2**, 685–695 (2006).
- Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology* **191**, 659–675 (1986).
- Peyrard, M. & Bishop, A. R. Statistical mechanics of a nonlinear model for DNA denaturation. *Physical Review Letters* **62**, 2755 (1989).
- Dauxois, T., Peyrard, M. & Bishop, A. R. Entropy-driven DNA denaturation. *Physical Review E* **47**, R44 (1993).
- Alexandrov, B. S. *et al.* A nonlinear dynamic model of DNA with a sequence-dependent stacking term. *Nucleic Acids Research* **gkp016** (2009).
- Bishop, A., Rasmussen, K., Usheva, A. & Alexandrov, B. S. Entropy-driven conformations controlling DNA functions. In *Disorder and Strain-Induced Complexity in Functional Materials*, 273–292 (Springer, 2012).
- González-Jiménez, M. *et al.* Observation of coherent delocalized phonon-like modes in DNA under physiological conditions. *Nature Communications* **7** (2016).
- Turton, D. A. *et al.* Terahertz underdamped vibrational motion governs protein-ligand binding in solution. *Nature Communications* **5** (2014).
- Guéron, M., Kochoyan, M. & Leroy, J.-L. A single mode of DNA base-pair opening drives imino proton exchange. *Nature* **328**, 89–92 (1987).

48. Fei, J. & Ha, T. Watching DNA breath one molecule at a time. *Proceedings of the National Academy of Sciences* **110**, 17173–17174 (2013).
49. Roberts, R. J. On base flipping. *Cell* **82**, 9–12 (1995).
50. von Hippel, P. H., Johnson, N. P. & Marcus, A. H. Fifty years of DNA “breathing”: reflections on old and new approaches. *Biopolymers* **99**, 923–954 (2013).
51. Siebenlist, U. RNA polymerase unwinds an 11-base pair segment of a phage λ promoter. *Nature* **279**, 651–652 (1979).
52. Choi, C. H. *et al.* DNA dynamically directs its own transcription initiation. *Nucleic Acids Research* **32**, 1584–1590 (2004).
53. Alexandrov, B. S. *et al.* DNA dynamics play a role as a basal transcription factor in the positioning and regulation of gene transcription initiation. *Nucleic Acids Research* **38**, 1790–1795 (2010).
54. Potaman, V. N. *et al.* Unpaired structures in SCA10 (ATTCT) n.(AGAAT) n repeats. *Journal of Molecular Biology* **326**, 1095–1111 (2003).
55. Phelps, C., Lee, W., Jose, D., von Hippel, P. H. & Marcus, A. H. Single-molecule fret and linear dichroism studies of DNA breathing and helicase binding at replication fork junctions. *Proceedings of the National Academy of Sciences* **110**, 17320–17325 (2013).
56. Ambjörnsson, T. & Metzler, R. Binding dynamics of single-stranded DNA binding proteins to fluctuating bubbles in breathing DNA. *Journal of Physics: Condensed Matter* **17**, S1841 (2005).
57. Alexandrov, B. S. *et al.* DNA breathing dynamics distinguish binding from nonbinding consensus sites for transcription factor γ 1 in cells. *Nucleic Acids Research* **40**, 10116–10123 (2012).
58. Jablensky, A. *et al.* Promoter polymorphisms in two overlapping 6p25 genes implicate mitochondrial proteins in cognitive deficit in schizophrenia. *Molecular Psychiatry* **17**, 1328–1339 (2012).
59. Nowak-Lovato, K. *et al.* Binding of nucleoid-associated protein Fis to DNA is regulated by DNA breathing dynamics. *PLoS Comput Biol* **9**, e1002881 (2013).
60. Duan, J. *et al.* A rare functional noncoding variant at the gwas-implicated mir137/mir2682 locus might confer risk to schizophrenia and bipolar disorder. *The American Journal of Human Genetics* **95**, 744–753 (2014).
61. Klimašauskas, S. & Roberts, R. J. M. Hhal binds tightly to substrates containing mismatches at the target base. *Nucleic acids research* **23**, 1388–1395 (1995).
62. Ye, Y. *et al.* Enzymatic excision of uracil residues in nucleosomes depends on the local DNA structure and dynamics. *Biochemistry* **51**, 6028–6038 (2012).
63. Manning, G. S. Breathing and bending fluctuations in DNA modeled by an open-base-pair kink coupled to axial compression. *Biopolymers* **22**, 689–729 (1983).
64. Mills, J. B., Vacano, E. & Hagerman, P. J. Flexibility of single-stranded DNA: use of gapped duplex helices to determine the persistence lengths of poly (dT) and poly (dA). *Journal of molecular biology* **285**, 245–257 (1999).
65. Yan, J., Kawamura, R. & Marko, J. F. Statistics of loop formation along double helix DNAs. *Physical Review E* **71**, 061905 (2005).
66. Du, Q., Smith, C., Shiffeldrim, N., Vologodskaja, M. & Vologodskii, A. Cyclization of short DNA fragments and bending fluctuations of the double helix. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 5397–5402 (2005).
67. Alexandrov, B. S. *et al.* Specificity and heterogeneity of terahertz radiation effect on gene expression in mouse mesenchymal stem cells. *Scientific reports* **3**, 1184 (2013).
68. Dickerson, R. E. Definitions and nomenclature of nucleic acid structure parameters. *Journal of Biomolecular Structure and Dynamics* **6**, 627–634 (1989).
69. Goodsell, D. S. & Dickerson, R. E. Bending and curvature calculations in B-DNA. *Nucleic Acids Research* **22**, 5497 (1994).
70. Heath, P. J., Clendenning, J. B., Fujimoto, B. S. & Schurr, M. J. Effect of bending strain on the torsion elastic constant of DNA. *Journal of Molecular Biology* **260**, 718–730 (1996).

Acknowledgements

L.B.A. is supported through a J. Robert Oppenheimer Fellowship at Los Alamos National Laboratory. This research used resources provided by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the US Department of Energy National Nuclear Security Administration under Contract No. DE-AC52-06NA25396. Research performed at Los Alamos National Laboratory was carried out under the auspices of the National Nuclear Security Administration of the United States Department of Energy.

Author Contributions

B.S.A., L.B.A., K.Ø.R. and A.R.B. conceived the research. L.B.A. and B.S.A. performed the simulations, including development of the code. B.S.A., L.B.A., K.Ø.R. and A.R.B. wrote the paper. All authors discussed the results and commented on the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:[10.1038/s41598-017-09537-y](https://doi.org/10.1038/s41598-017-09537-y)

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017